UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
ICT International Doctoral School

CONCEPT CHALLENGE GAME

A GAME USED TO FIND ERRORS FROM A MUTILANGUAL LINGUISTIC RESOURCE

Hanyu Zhang

Advisor

Prof. Fausto Giunchiglia
Università degli Studi di Trento

Abstract

Multilingual semantic linguistic resource is critical for many applications in Natural Language Processing (NLP). While, building large-scale lexico-semantic resources manually from scratch is extremely expensive, which promoted the applications of automatic extraction or merger algorithms. These algorithms did benefit us in creation of large-scale resources, but introduced many kinds of errors as the side effect. For example, Chinese WordNet follows the WordNet structure and is generated via several algorithms. This automatic generation of resources introduces many kinds of errors such as wrong translation, typos and false mapping between multilingual terms. The quality of a linguistic resource influences the performance of the further applications directly, which means the quality of a linguistic resource should be the higher the better. Thus, finding errors is inevitable.

However, till now, there is not any efficient method to find errors from a large-scale and multilingual resource. Validating manually by experts could be a solution, but it is very expensive, where the obstacles come from not only the large-scale dataset, but also multilingual. Even though crowdsourcing is a method for solving large-scale and tedious task, it is still costly. By thinking in this scenario, we plan to find an effective method that can help us finding errors in low cost.

We use games as our solution and adopt Universal Knowledge Core (UKC) with respect to Chinese language as our case study. UKC is a multi-layered multilingual lexico-semantic resource where a common lexical element from a different language is mapped to a formal concept. In this dissertation, we present a non-immersive game named Concept Challenge Game to find the errors that exist in English-Chinese lexico-semantic resource. In this game, people will face challenges in English synsets and have to choose the most appropriate option from the listed Chinese synsets. The players are unaware when finding errors in the lexico-semantic resource. Our evaluation shows that people are spending a significant amount of time playing and able to find different erroneous mappings. Moreover, we further extended our game to Italian version, the result is promising as well, indicating that our game has the ability to figure out errors in multilingual linguistic resources.

Keywords
[game with a purpose, multilingual, lexico-semantic resource, knowledge base, mobile application, translations]

Acknowledgement

It has been more than four years for my Ph.D studying, and at the same time, it is the 7th year that I am staying in Trento. This small quiet city is almost my second hometown and I have to say this is the best town I have seen in my life, which is a beautiful, calm and friendly town in the mountain. I am familiar with every building here, where has delicious food and where is fun to play. I hope I can return back to this lovely city, constantly. Now, at the end of my Ph.D studying, I have to express my appreciation to some people who supported me all the time, sincerely.

The first person I would like to thank is my supervisor Professor Fausto Giunchiglia even though I had thanked him once in my master thesis. But, I still need to say once more that it gives me great pleasure to be one of your students. Thanks to take me to the research field and shaping me the research methodology.

Second, I have to say thanks to my family with supporting for more than 20 year of my studying career and 30 years life. The most fortunate thing of my life is to be your child.

Also, special appreciation to my friends and colleagues. Very glad to meet and work with all of you.

At last, let me say thanks who providing me the kindly help.

Contents

Chapter 1

1 Introduction

1.1 The Context

A linguistic resource plays a crucial role in semantic applications such as semantic web, NLP and word sense disambiguation. For capturing the semantics, and meanwhile enabling the characteristic of human-readability, a linguistic resource should contain corresponding independent natural language representations for every real-world concept and its instances. WordNet, which was developed by Princeton University, is such an English linguistic resource and used to be treated as a standard. But, since its initial target is an English Electronic Dictionary, and with the rapid development of globalization and modern information technologies, a single language linguistic resource becomes hard to meet the requirements of multilingual services. To give an ability of interoperability access languages, several WordNets were developed by following the English WordNet structure. One of them is MultiWordNet [43] consisting of several European languages. While, many language versions of WordNet are not developed yet, and even developed ones are either cannot be used effectively as they could not achieve critical mass (means they cannot achieve the application level) or cannot meet the English WordNet quality. Furthermore, even for the English version, its coverage is often unsatisfiable for performing some domain specific tasks.

Towards overcoming the coverage problem, some large-scale and multilingual linguistic resources like YAGO3 [44] were developed. The methodologies for building these large-scale multilingual linguistic resources are mainly based on the algorithms of extracting from existing resources. That was based upon the booming development of the encyclopaedia Wikipedia[1] and extraction algorithms. While, the automatic extraction methodology brings to us several drawbacks as well, for example, the over-reliance on the quantity and quality of existed multilingual resources. In reality, there are no existing resources to be extracted in some cultures, and even if resources are available, they are still hard to achieve the application level. For instance, in the English version of Wikipedia, it contains 4 million entities, yet its Chinese version has less than 1 million. And since the uneven quality of the linguistic resources and the extraction algorithms, more errors are introduced by comparing to the manually created linguistic resources. Universal WordNet was created following WordNet structure and extracting data from 200 language editions of Wikipedia, where accuracy is ranged from 59%- 83% [45]. As we mentioned, the multilingual versions are suffering from the lack of contents according to multilingual resources. Some experiments are trying to use machine translation to overcome the issue of lacking existing resources, but the quality of translation is questionable. A notable example is BabelNet [46] where is a linguistic resource linking Wikipedia encyclopaedic entries to WordNet lexicon automatically, using statistical machine translation to fill the missing translations for resource-poor languages. Its accuracy is close to 70%.

[1] Wikipedia: https://www.wikipedia.org/

[2] http://www.yeeyan.org

In Table 1, WordNet and its derivatives are created by human-manually, in which supposed to be closed to 100% accuracy. Chinese WordNet [56] was created by a semi-automatic method. They extracted some resources to populate the database and then validate by human efforts. The last two were created fully automatically, which is large-scale but with relatively low accuracy.

Name	Accuracy	Coverage	Gloss	Languages
WordNet	100%	117,659	Yes	1
MultiWordNet	100%	**57,934**	**Poor**	7
IndoWordNet	100%	**35,000**	Yes	17
EuroWordNet	100%	**60,000**	**Poor**	7
Chinese WordNet	80-90%	117,000	**No**	2
Universal WordNet	**59%~83%**	~117,659*200	**No**	200
BabelNet	**67%~76%**	3032406	**Poor**	50

Table 1 The accuracy of some linguistic resources

As showed in Table 1, the accuracy is decreasing significantly when the data size increasing, where because of either automatic or semi-automatic creation methodology. An idea for overcoming these challenges comes from the success of crowdsourcing [47], which is an approach widely used to solve time-consuming and tedious tasks. It collects needed works by soliciting contributions from an online community, rather than from traditional employees or experts. Especially in knowledge collecting and translating aspects [48], crowdsourcing demonstrated its advantages by comparing to other methods. Terminology of ‘crowdsourcing’ was firstly indicated by Jeff Howe in 2006 used to describe a new business model, in which tasks are distributed through Internet. Brabham further defined crowdsourcing [49] and created a typology of crowdsourcing [50] basing on unsolidified theoretical knowledge of crowdsourcing situation. Due to the extraordinary number of Wikipedia contributors, it has been demonstrated that crowds can outperform linguists in terms of coverage. The research shows that the ‘wisdom of crowds’ based resources are not generally superior to ‘wisdom of linguists’ based resources. However, it is worthwhile to note that collaboratively created knowledge sources are strongly competitive in linguistic knowledge sources on the majority of datasets [51]. Furthermore, crowdsourcing has demonstrated its advantages in the field of translation work, which strengthens our confidence that crowdsourcing will be a good solution. For example, Yeeyan[2] is the largest community translation site in China with more than 400,000 registered users and 30,000 community translators. Community translators use their spare time and multilingual skills to translate interesting stuff they read on the web and share them with Chinese readers. Wordreference[3], a free and multilingual online dictionary, uses its forum to discuss and collect words and meanings from its users, and those words and meanings will be used for their online dictionaries

[2] http://www.yeeyan.org

[3] http://www.wordreference.com

after verifying by experts or seniors. Duolingo[4], a free language study tool, collects language translation from its students when they are practicing the language. Other examples like Facebook[5] and Twitter[6], crowdsourced language translations of their websites based on their huge amount of users. One of the significant challenges of crowdsourcing is how to encourage people to contribute, e.g. incentive mechanisms. Generally speaking, the incentive of crowdsourcing includes Payment, Altruism, Enjoyment, and Reputation, etc. [52] While, Scekic classified it into Pay per performance, Quota/Discretionary bonus, Deferred compensation, Relative, etc. [53]

Several works were proposed that adopting crowdsourcing to populate a linguistic resource in order to overcome the issues of automatic methods. For example, Universal Knowledge Core (UKC)[55]. UKC provides a unified platform to accommodate lexico-semantic resources in multiple languages. It utilizes synsets from WordNet to generate language elements. The generated language elements are then mapped to corresponding lexical elements from different languages using concepts. The architecture is language independent and has a multi-layered ontology called Concept Core. It consists of numerous multilingual resources called Local Knowledge Core (LKC). Each LKC is a copy of UKC restricted to one language. To save effort, Most LKCs were bootstrapped with its corresponding existing linguistic resource. For Chinese case, Chinese WordNet was imported as bootstrapping. For Italian, MultiWordNet was imported.

1.2 The Problem

Since the differences of the creation methodologies, usually, linguistic resources built by expert groups manually always have the highest quality, i.e. human-level accuracy. WordNet and many domain ontologies are the best examples for manual creation. But, as we mentioned, large-scale semantic multilingual linguistic resource is inevitable. These large-scale resources were primarily based on automatic or semi-automatic creation method, which introduced a lot of errors as the side effect. Some of them built through extraction algorithm such as DBpedia and YAGO can achieve near-human quality [57, 58]. Even though crowdsourcing could be a solution for building a large-scale linguistic resource, errors still exist either from bootstrapping or crowdsourcing itself. For example, in UKC, the bootstrapping of Chinese WordNet brought several kinds of errors. Apparently, the quality of a linguistic resource directly influences the performances of its related applications, in which case the quality of a linguistic resource is the higher the better. Low quality will cause bad performances of the corresponding applications.

While, because of the dataset of a linguistic resource is extremely huge, keeping the human-level quality becomes a nontrivial task. At the initial stage, the maintenance work for a small-scale linguistic resource is mainly based on the human effort, which is very costly for a large-scale one obviously. Besides, the trend of multilingual also increases the cost of the maintenance work significantly. A human validator needs to understand not only the knowledge of linguistic resource, but also has to master at least two languages. Furthermore, according to the errors in linguistic resources are always related to semantic relations and word senses, most kinds of errors are hard to detect by machine automatically.

[4] https://www.duolingo.com

[5] https://www.facebook.com/?sk=translations

[6] https://about.twitter.com/company/translation

Maintenance work includes several parts. One of its most important parts is to fix errors on which we are focusing. Fixing error has two parts of work, finding and correcting, in which finding is considered the hardest one. Finding errors from millions records and with multilingual simultaneously is very costly even by using crowdsourcing. By considering this scenario, we are going to look for a low cost way to find errors from a large scale, multilingual linguistic resource for a long-term period. The basic contemporary obstacles faced are:

- Large-scale data size
- Multilingual data type
- Multi-type errors
- A long term maintenance method since dataset is increasing
- People needs knowledge background of semantic multilingual linguistic resource
- Low cost

1.3 The Solution

In order to further reduce the cost of finding errors in a large-scale multilingual linguistic resource, we adopt games as our solution, which derived from the idea named Games with a Purpose (GWAP). The idea of GWAP has been wildly used in many domains such as, Foldit [34], in which non-scientists players are salving protein structure prediction problems, ESP game, where players are labelling images with words, Page Hunt, which is used to improve search engines. And MobiMission [35] is used for geospatial tagging systems. Several previous works with respect to GWAP are tried in order to reduce the cost of the maintenance work, such as Infection [36], which is a video game used to validate common sense knowledge in a knowledge base. However, these games are mainly based on monolingual aspect and two players. Generally speaking, games are helpful for long-term and tedious works. Also, it is in low cost. Furthermore, we can design several games to solve multiple kinds of errors.

We selected a representative linguistic resource named UKC, more specifically, Chinese LKC, Italian LKC, as our case study. Because 1), it imported several existing resources as bootstrapping. These resources are either human made or semi human made, which can cover most creation types. 2), crowdsourcing is the primary creation method. If we use it, we can also know whether our game works on crowdsourcing created resources. 3), it is large-scale and multilingual. These features are the contemporary obstacles we faced for maintaining a semantic linguistic resource.

While, since there are multi-type errors in a semantic linguistic resource, one game may not enough to figure out them all. According to this, we developed a UKC game framework to maximize the reusable components. After that, we create a serious game named word challenge as the first example to validate a knowledge base, which was inspired from English Vocabulary Challenge game. This game is mainly focusing on figuring out the wrong mappings of synsets between target and source language in a large-scale multilingual knowledge base.

Our evaluation shows that players spend significant amount of time playing the game at a short time span. The constraint for the players was the type of the phone they were using to play the game. Many participants were unable to play as they owned an iPhone, but the system was developed for Android phones. The implementation of the game indicates that the game players with very limited linguistic background can also be involved in finding different kinds of error in a multilingual lexico-semantic resource which can be a major help while building a high quality lexico-semantic resource. We further extended our game to Italian language to verify the performance of the game for the other languages. A promising result shows that our game has the ability to find errors for the other languages.

1.4 Research issues

Game has been proved effective in solving long-term and tedious tasks, but some issues from game perspective need to be worked out as well. Before designing a game, we need to specify the error types. So, at first, we introduce the general error types in a semantic multilingual linguistic resource. After that, we introduce the challenges in game designing. For example, game incentive model, game generation, etc.

1.4.1 Error types

In general, we consider an error in three aspects, semantic relation, word sense and word form. Since we use Chinese language as our case study, the following examples are mainly presenting in Chinese. From WordNet point of view, a concept refers to ‘traveling across’ is represented by a word ‘Crossing’, which is named as a synset as well. As shown in Figure 1, synset ‘Crossing’ has two children, ‘Ford, Fording’ and ‘Traversal, Traverse’. And it has a hypernym ‘Travel, Traveling, Travelling’. The behind knowledge is very easy to understand. There are two ways of crossing, either ‘the act of crossing a stream or river by wading or in a car or on a horse’ (ford, fording) or ‘taking a zigzag path on skis’ (traversal, traverse). This kind of knowledge is universal in most of cases, but there are exceptions. In some culture, it might have some problem.

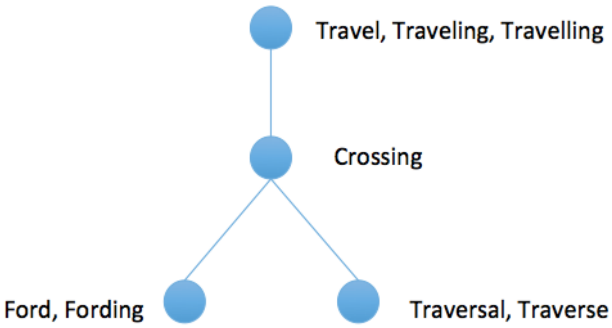


Figure 1 Semantic relations

- **Semantic relation errors**

Semantic relation errors are these relation errors. Figure 2 shows the semantic relation of ‘Bay, Embayment’, ‘Bight’ and ‘Gulf’ and also its corresponding Chinese relations. Without considering the semantic relation, the Chinese part is correct, as ‘Bight’ is ‘海湾’, ‘Gulf’ is ‘海湾’ and ‘Bay, Embayment’ is ‘海湾’ as well. But when we take sematic relation in consideration, it is an error. ‘海湾’ cannot be itself, we should remove the two children and their relations. We call this kind of error as semantic relation error. Semantic relation error has several derivative types.

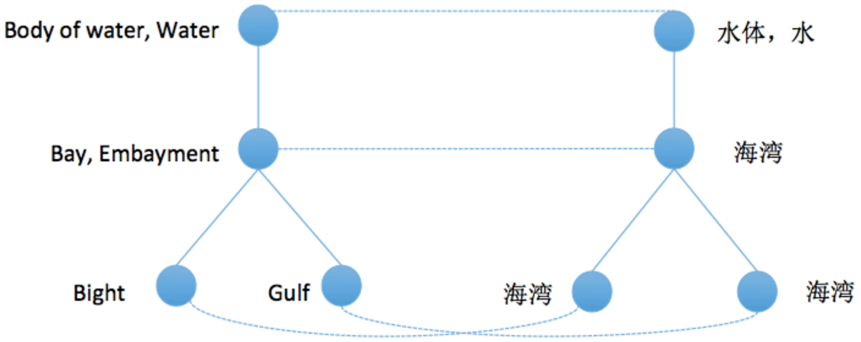


Figure 2 Semantic relations in terms of Chinese and English

- **Word sense errors**

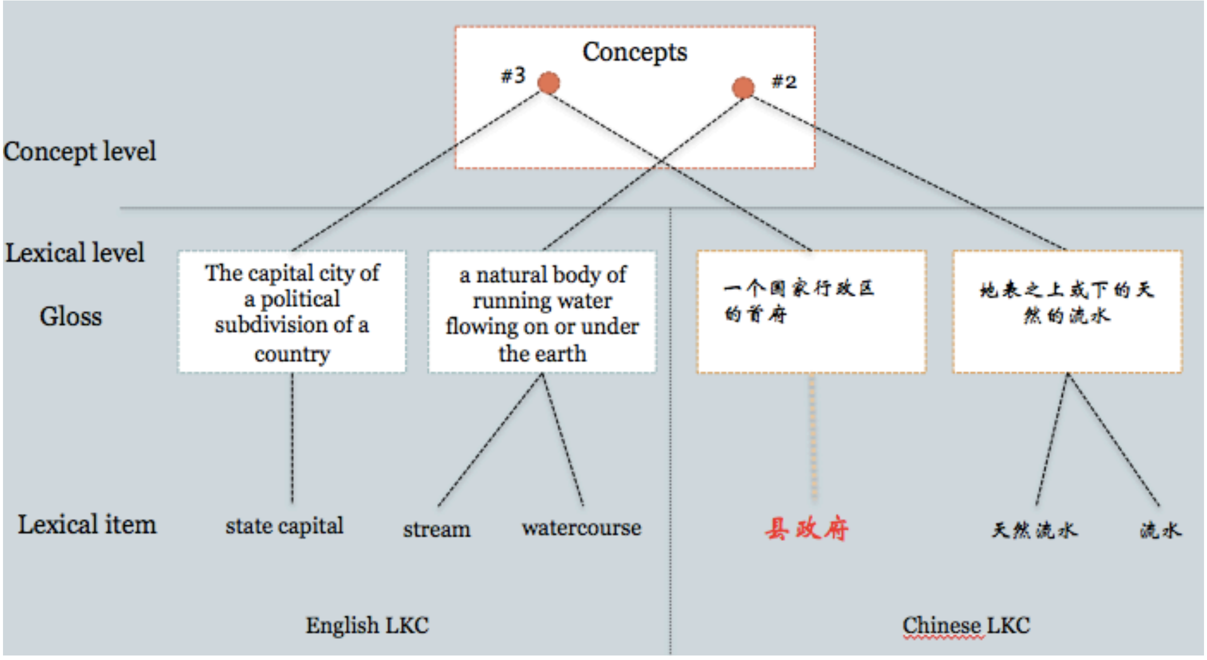


Figure 3 An example of word sense error

Word sense errors are these errors related to word senses. Figure 3 is an example of the word sense error. In this example, it has two concepts, 3 and 2, and each concept has two language representations. The English vocabulary representation (synset) of concept 3 is ‘state capital’, Chinese is ‘县政府’. The English synset for concept 2 is ‘stream, watercourse’, Chinese one is ‘天然水流’. An erroneous case in Chinese LKC inside UKC is concept 3. In this case, the synset ‘state capital’ in English has the correct gloss in Chinese but the synset ‘县政府’ (county government) is an incorrect translation. Furthermore, the sense of the terms in both languages is different, since the English term refers to a location, whereas the Chinese counterpart refers to an organization.

The previous example is an error of mapping between Chinese and English synsets. While, a synset itself can have several problems also. For example, ‘Not rich enough’ and ‘Partial correct’ ‘Not rich enough’, used to indicate the coverage of a single synset. For example, for English synset ***{hostel, hostelry, inn, lodge, auberge}***, its corresponding Chinese synset is ***{旅社，旅店}***. While, in reality, Chinese words **旅店、旅馆、旅社、酒店、旅舍、栈房、客栈、客店** are used to express this sense commonly. 'Partial correct' refers to a part of a synset is not correct. For example, English synset ***{hostel, hostelry, inn, lodge, auberge}***, its corresponding Chinese synset is ***{旅社, 旅店, 茶馆}.*** While, the first two words in this Chinese synset are correct, but the last word means teahouse is not correct.

- **Word errors**

Word errors refer to the errors related to a word, for instance, 'Typo' or 'Not a word'. Typo is easy to understand. 'Not a word' refers to some words of a synset are of the correct meaning, but not a real word in the language. For instance, the Chinese synset **{全体教职员工和学生}**, this Chinese means all staff and students, which is correct for this meaning for the synset **{school}**, but, strictly speaking, it is phrase instead of a word.

1.4.2 Challenges

We decided to use games as the solution to deal with the long-term maintenance work. While, during the game designing and developing, we met three challenges. The first difficulty should be Game design and incentive model. We need to design a fun game, which can attract more players subsequently. Second, we need to think how to transfer linguistic resource data to game-like data. The third problem is feedback quality control. The evaluation of the competence of the player is inevitable and also how we can validate a linguistic resource from the game feedback. In the following of this section, we will briefly introduce how we solve these four issues respectively.

Game design and incentive model

In GWAP theory, there are several existing game modes based on the agreement method including input-agreement, output-agreement, inversion-problem, etc. These methods are mainly used to verify unlabeled data and need two players to get an agreement in general. In our case, the requirement is different. Our data needed to validate is harder than the normal case. For example, in the ESP game, the task is to label what objects are in an image. The results are mainly concentrated in some easy words like human, car, and color, etc. It is very easy to recognize that there is a car or woman in an image. But our task is to validate semantic relations, multilingual mappings and words, which are hard even for experts. A validator needs to master not only language knowledge, but also knowledge base background. Thus, we need to narrow down our agreement method rather than using these existing ones. Moreover, since the data in a linguistic resource is serious, it is not that funny like pictures or common senses in a game point of view. So, we restrict our game into a serious game. To design the first game, we investigated all famous Chinese knowledge game types because that an existing popular game mode is easier to be accepted by players. Furthermore, in additional to the general incentives of a serious game, we also added some competitive incentives like leaderboard, first player claim, etc.

Game-like data generation

After investigation of the existing Chinese knowledge games, we found that the question-answer game mode can satisfy our requirements. So, in this part, we need to consider how to transfer knowledge base data into game-like data. To solve this, we built a games framework, which can generate several types of question-answer pairs. Also, to generate richer data, we integrated 2 additional databases, WordNet Domains and lemmatized word frequency as shown in Figure 4. WordNet domains developed by FBK, and word frequency data was extracted from British National Corpus. We use English LKC as questions since it has the best quality. And target Chinese LKC, which we need to validate, as the answers. After generation, games framework provides 2 types of questions, 7 difficulty levels and 4 kinds of option sets.

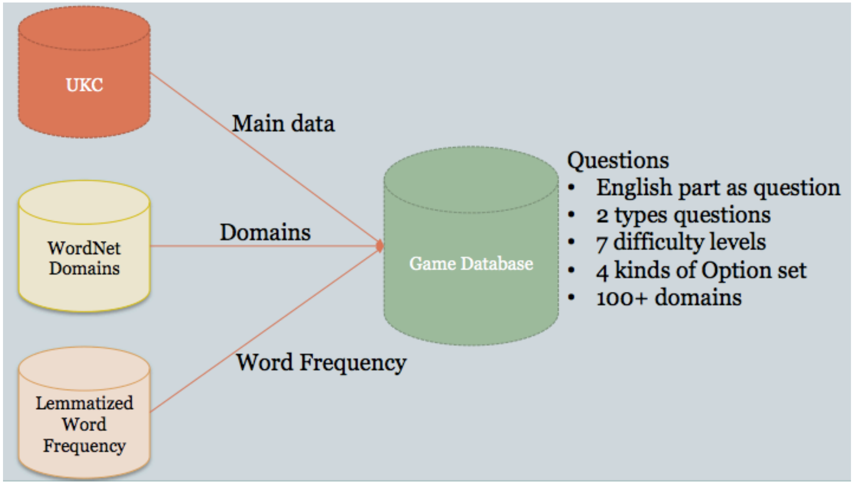


Figure 4 UKC games framework database

Feedback quality control

The collected game feedbacks are used to validate our UKC records as a voting system. The basic assumption is 'the most selected answer is the correct answer for a question'. We adopted several algorithms, for example, X square, super majority, relative majority, DS evidential theory, etc., by consider different aspects, like probability of each option, user accuracy, etc., to select the final result, and the relative majority method shows the potential to generate the best results.

Since it is a voting system, understanding the user actions is inevitable. Questions answered by guessing or randomly input will decrease the system performance obviously. Thus, we need to evaluate a player's performance. To improve the feedback quality, we need to filter out these bad answers, for example, a player's accuracy is only 20%, and his answers might be useless. Instead of creating a small size gold standard, we use UKC data to judge the correctness directly. The Chinese LKC was bootstrapped by Chinese WordNet and its evaluated accuracy is around 80-90%. While, since our task is approximately distinguish good answers and bad answers rather than to specify a player's exact accuracy, like 32.5%, we think the original data from UKC can fulfill our requirement. Furthermore, to further improve the feedback quality, in Concept Challenge Game, all game elements are designed to let a player to answer questions honestly. For example, in the game result, skip honestly is without punishment, but answer wrong by guessing will reduce the game score. We also provided two thresholds to filter answers, which are accuracy level and honestly level.

1.5 Structure of the Thesis

The following thesis is organized as, in Chapter 2, first, we briefly introduced our case study Universal Knowledge Core main structure including natural language core and concept core. Second, we specifically introduced the natural language core bootstrapping procedure in terms of Chinese language. Since we imported Chinese WordNet as bootstrapping, in this chapter, we also introduced all existing Chinese linguistic resources and why imported Chinese WordNet was imported in details.

Since we use Game with a purpose (GWAP) as our solution, after introducing our case study, in Chapter 3, we introduced what is GWAP and some examples of GWAP. Our focus is Chinese Language, thus in this Chapter we introduced all popular Chinese knowledge game modes as well.

As we mentioned before, there are several kinds of errors in a multilingual linguistic resource, one game might be not enough to figure them all. By this scenario and at the same time maximize the reusable components, we developed UKC game framework. So, in Chapter 4, we introduced how we create the UKC game framework and its main components.

As we have investigated all the existing Chinese knowledge game modes, we found that a game named Vocabulary Challenge Game can embed our purpose perfectly. So in Chapter 5, we did a survey of existing Vocabulary Challenge Games. And in Chapter 6, we presented the designing procedure of our game named Concept Challenge game that inspired from Word Challenge Game.

Chapter 7 introduced the evaluation of the Concept Challenge Game and Chapter 8 is the implementation including game data integration, game framework architecture, etc. Chapter 9 is the summary of the thesis.

Chapter 2

2 Universal Knowledge Core

2.1 Universal Knowledge Core

WordNet [3] is an English language dictionary based on synsets, containing gloss and sense. The significant innovation of WordNet comes from its semantic structure, and because of that, it plays an important role in NLP and AI filed. Numerous languages are following the step of English WordNet, developing WordNets in their own languages. However, it introduced several limitations. Especially, it is in British English, where the glosses given for the terms are mainly focusing on the British society and culture. For example, the explanation of "primary school" is "*a school for young children; usually the first 6 or 8 grades*", which is obviously biased towards the British educational system. In this scenario, WordNet cannot be used directly in some multilingual and multicultural environments.

Universal Knowledge Core (UKC) was developed based on solving the above limitations 2. It provides mappings between language-independent concepts connected with semantic relations, synsets composed by synonymous words and lexical gaps in case a certain language cannot express a concept. DERA methodology [5] and its guiding principles [6] are employed in order to avoid bias on any cultural, spatial or temporal. The UKC consists of two fundamental components: Natural language core and Concept core.

2.1.1 Natural Language Core

Nature language core is composed by words, senses, synsets and exceptional forms, where a synset in a given language is connected to a concept, word senses are organized into four part-of-speech (POS) noun, verb, adjective and adverb, one word may have more than one POS and synonym word senses with the same POS are grouped into synset.

Word: A word is the basic lexical unit of the NL core represented as a lemma. It can be multiword, phrasal, collocation, etc.

Sense: A sense is the meaning of a word. A word can have one or more senses, with a same of different POS tag. A sense of a word is owned to only one synset.

Synset: A synset is a set of words sharing the same meaning. In fact, words in a synset have semantically equivalent relations. Each synset might be accompanied by a gloss consisting of a definition and optionally example sentences. Also each synset in a language is related to a concept in the concept core.

Exceptional form: An exceptional form is an inflected representation of a lemma, for instance, wives (plural form of the noun lexeme wife) and best (irregular superlative of the adjective lexeme good).

Lexical relation: A lexical relation is a relation between the words of different synsets. In fact following relations are of this kind: antonym, derivationally related form, pertainym or derived from adjective, participle-of-verb and homograph-of. Even though the WordNet has not only lexical relations but also semantic ones, in the UKC lexical relations are part-of the NL core because they hold between words that are language dependent.

2.1.2 Concept Core

The concept core consists of concepts and semantic relations between concepts. The architecture is language independent and is under a multi-layered ontology.

Concepts: A concept is a language independent representation of a synset. For example, country, city, person. The concept *city* can be represented as *city* in English, *città* (chit'a) in Italian, *xom* (khot) in Mongolian, 城市 in Chinese. Thus, a concept is connected to multiple synsets with the same meaning in different languages.

Semantic relations: A semantic relation is a property connecting concepts to build the hierarchy or semantic network. *Is-a* and *part-of* are the examples of semantic relations.

2.2 Chinese Local Knowledge Core

Universal Knowledge Core provides a unified platform to accommodate lexico-semantic resources in multiple languages. Its Natural Language Core consists of several multilingual resources called Local Knowledge Core (LKC). LKC is a copy of UKC restricted to one language. English is available for all LKCs as the label language for concepts. For example, Chinese LKC is restricted in Chinese and English. Due to the reason that creating a linguistic resource from scratch is costly, we imported an existing Chinese linguistic resource as a bootstrapping. Importing various existing resources will not only be more cost-effective, but also helpful for combining information from different resources, keeping the ultimate knowledge more consistent and reliable.

Before importing, we did an investigation of all existing Chinese linguistic resources. As showed in Table 2, we found six Chinese linguistic resources, but three of them were built by Chinese language only. In addition to these 6 resources, Chinese WordNet in Taiwan is in traditional Chinese. Contemporary Chinese Predicate Verb Dictionary [7] is the result of an initial small-scale experiment, which is operated after learning and adjusting foreign semantic description theory. '905' Semantic Project [9] and Hownet [10] are aiming to create a large-scale linguistic resource. Beida-SemDict [11] is a product of machine translation. Since the success of English WordNet, a lot of cultures are developing WordNet in their own language. CCD [11,13] and Chinese WordNet [15] (CWN) are the attempts to be geared with international linguistic resource standards. It is therefore difficult to indicate which is a better linguistic resource because they all have different focuses. For example, in the hierarchy perspective, WordNet is using the tree structure to depict knowledge, yet HowNet uses a net structure. In some cases, tree structure has the better performance, for instance, concept 'willow, it is a tree (hypernym) and it can be divided into white willow, silver willow, etc. (hyponym). However, for example, concept 'dust', people do not recognise it in the tree structure in fact. Furthermore, WordNet uses synset to indicate word sense but HowNet uses sememe. Both of them have a good performance to describe a word sense. CCD was manually built which is costly but with high quality. In addition, some unique Chinese semantic relations and concepts are added in, which is helpful in understanding Chinese language better. CWN is created by automatic translation and manually validation, which means it is low cost but not in high quality, also, it lacks of Chinese specific elements. Till now, only HowNet and CCD keep updating, as they are commercial products, having a long-term financial support. However, they are really expensive to use. Since words and meanings are increasing and changing, the development of a linguistic resource should keep updating. A maintenance methodology is preordained for the linguistic resources.

Our focusing is to bootstrap the Chinese LKC where English-Chinese connections are necessary. Thus, we only focus on multilingual linguistic resources, that is, *Chinese WordNet* (CWN), *HowNet* [7], *CCD*, and *Chinese WordNet*[8] in Taiwan. Three conditions are primarily considered, including the quality, the dataset structure and the user license. Through studying, we learned that *HowNet* has the highest quality; *Taiwan Chinese WordNet* is built by traditional Chinese, which is different with the simplified Chinese; *CCD* has a higher quality and express in simplified Chinese, but expensive to use. CWN is free to use and has a huge coverage, but the quality is not high. As a result, the first two resources are abandoned, since 1), the data structure of *HowNet* is different with WordNet, lacking the mapping between them. 2), the differences between traditional and simplified Chinese are not only conformed to the characters, but also in the expression. In such case, the transformation of expression from traditional Chinese to simply Chinese could cause some unexpected problems. 3) HowNet and CCD are not open source, and expensive even for educational purpose. For example, for HowNet, educational usage is around 20,000 euros.

Thus, the existing resource that we plan to import is simplified *Chinese WordNet* (CWN), which is free to use and fully following the data structure of English WordNet, and we found the glosses mapping between CWN and the extended version of WordNet, which would guarantee the success of importing. Another method is to find ID mapping between WordNet IDs for different versions. CWN developed by the cooperation of Department of Computer Science and Engineering at Southeast University and Department of Computer Science at Vrije Universiteit Amsterdam. Considering that there are a lot of high quality dictionaries nowadays, in the translation procedure, they adopted some algorithms in order to reduce the human efforts, including minimum distance algorithm and intersection algorithm. And finally, they did manual validation for all results. By using these algorithms and manual correction, the resulting Chinese WordNet contains 118000 Chinese words and 115400 synsets.

Name	Time	Institution	Scale	Construction Method	Language	Status

[7] http://www.keenage.com

[8] http://lope.linguistics.ntu.edu.tw/cwn/

Contemporary Chinese Predicate Verb Dictionary	1990-1993	Tsinghua University, Chinese People University	1000Verbs 3000Senses	Manually	Chinese	Closed
'905' Semantic Project	1990-1995	Beijing Language University	40,000Words 50,000 Senses	Manually	Chinese	Closed
HowNet	1988-	Chinese Academy of Sciences	2199 sememes and 116533 records	Manually	Chinese/ English	Alive/ Commercial Products
CCD	2000-	Beijing University	70,000 Concepts	Manually	Chinese/ English	Alive
Beida SemDict	1996-	Beijing University	65330 Words	Manually	Chinese	Alive
CWN	2008-	Southeast University Vrije Universiteit Amsterdam	118,000 words 115,400 synsets	Automatic/ Manually Validation	Chinese/ English	Closed/ Free

Table 2 Existing Chinese linguistic resources

The rest of this Section is arranged as: first, analysing HowNet, CCD and CWN respectively. But since CWN is our target resource, CWN is with more details. Second, we introduced how we evaluate CWN. Third, we introduce how we imported CWN into UKC as the bootstrapping of Chinese LKC.

2.2.1 Chinese Concept Dictionary

Chinese Concept Dictionary (CCD) is a WordNet-like semantic lexicon of contemporary Chinese, which is well structured mathematically from computational lexicography perspective. Labelled tree structure is expected to be adopted as the basic description method for hypernymy & hyponymy hierarchy. CCD is compatible with English WordNet in the construction of concept, which does not mean that the CCD is the same with WordNet. In fact, the differences between Chinese and English in the description structure and processing method have been noticed. The difficulties in automatic analysis caused by the lack of morphological constraints in Chinese. One aim of the CCD is to offer more knowledge helpful to Chinese syntactic-semantic analysis.

CCD Structure

CCD follows the English WordNet structure essentially, using synonymous set (synset) as Concept, and including relations between concepts. As same as English WordNet, it contains Noun, Verb, Adverb and Adjective. Main relations are synonymy, antonymy, hypernymy & hyponymy, meronymy, etc.

The main structure of the CCD is hypernymy & hyponymy relation, which makes CCD as a forest with distinct categories. The CCD inherits the set of initial trees provided by English WordNet for noun-part and verb-part. That because they found that the initial English concept classification is also effective for Chinese concept. The CCD is compatible with English WordNet in the structure of concept, however, in order to emphasize the features of Chinese language, it refines the concept content and relations between concepts according to simplified Chinese characteristics. For example, *subarea* relation and *part of time* relation. The content of Noun in CCD belongs to notional words, where are the nominal in the grammatical point of view. Similarly, in the grammatical perspective, the content of Verb is verb and predicate pronouns.

Creation Procedure

First, they extracted English WordNet structure as the bootstrap, including synsets and relations. Second, as showed in Figure 5, they developed a visualized and data-sensitive application, named the Visualized Auxiliary Construction of Lexicon (VACOL) (Figure 6), in order to display, modify and enhance the extracted information. Since the diversity of cultures, the concept mapping between WordNet and CCD is not only one-to-one, but also many-to-many. For example, in Figure1, C5 is mapping to E1, E6 and E5, C4 is only mapping to E2.

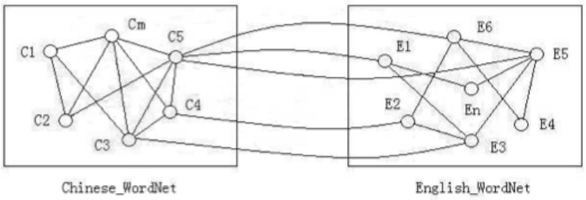
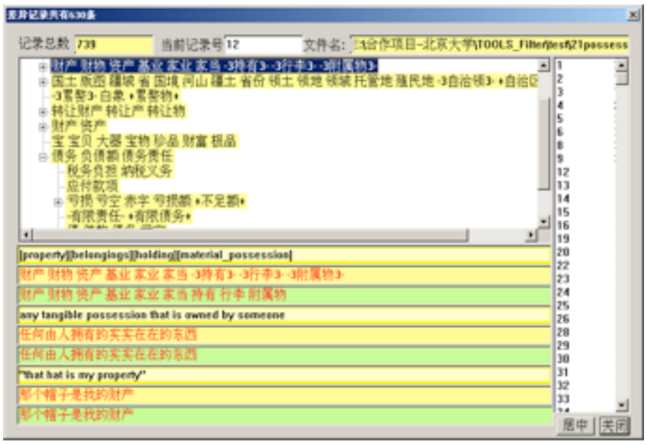


Figure 5 Mappings between CWN and EWN [12]

Lexicographers considered the following conditions in the process of development.

1. By considering current node’s hypernymy & hyponymy, if this node has its corresponding Chinese concept, they just translate the node content to Chinese.
2. If this node does not have a corresponding Chinese concept, the following conditions are taking into consideration.
   - If this English concept is too general, create a hyponym as the son node.
   - If it is too specific, delete this node from CCD.
   - If inappropriately classified, need to move all the succeeding son nodes.

As a result, the CCD has around 50000 concepts, and since it used labelled tree structure, each Chinese concept can find its corresponding English concept.


差异记录共有630条

记录总数 739　当前记录号 12　文件名: :\合作项目-北京大学\TOOLS_Filter\esf\21possess

财产 财物 资产 基业 家业 家当 ‹持有› ‹行李› ‹附属物›
国土 版图 疆域 省 国境 河山 疆土 省份 领土 领地 领域 托管地 殖民地 ‹自治领› ‹自治区
‹累赘› 白象 ‹累赘物›
转让财产 转让产 转让物
财产 资产
宝 宝贝 大器 宝物 珍品 财富 极品
债务 负债额 债务责任
税务负担 纳税义务
应付款项
亏损 亏空 赤字 亏损额 ‹不足额›
‹有限责任› ‹有限债务›

[property][belongings][holding][material_possession]
财产 财物 资产 基业 家业 家当 ‹持有› ‹行李› ‹附属物›
财产 财物 资产 基业 家业 家当 持有 行李 附属物
any tangible possession that is owned by someone
任何由人拥有的实实在在的东西
任何由人拥有的实实在在的东西
"that hat is my property"
那个帽子是我的财产
那个帽子是我的财产

Figure 6 Visualized Auxiliary Construction of Lexicon [12]

2.2.2 HowNet

HowNet is a common-sense linguistic resource revealing conceptual relationships and attribute relationships of concepts. Semantic, that is, relations between concepts are the soul of HowNet, as well as the knowledge. The relationships used to represent knowledge can be divided into Concept Relationship (CR) and Attribute Relationship (AR). Those relations construct Concept Relation Net (CRN) and Attribute Relation Net (ARN). Different individual has different CRN, even for the same concept. This reflects different levels of knowledge among people. As a knowledge base, the knowledge structured by HowNet is a net instead of a WordNet tree as showed in Figure 7. It is dedicated to demonstrating the general and specific properties of concepts.

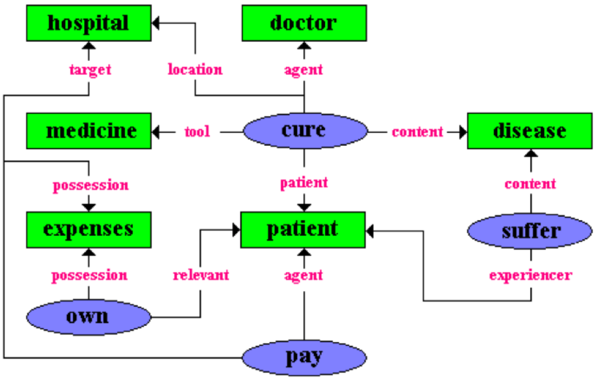


Figure 7 HowNet Relations [10]

Construction and Methodology

The first step of HowNet creation is defining sememes, which are as difficult as defining morpheme. Generally speaking, a sememe is the smallest basic semantic unit that cannot be divided any further. For instance, “human being” can be regarded as a sememe by ignoring a complex concept encompassing a set of attributes. The hypothesis is that all concepts can be reduced to the relevant sememes. So that there exists a close set of sememes, from which, composes an open set of concepts. In this case, if they can manage the close set of sememes to describe concept relations as well as attribute relations, an ideal knowledge base would be conceivable.

The establishment of sememe set is based on the meticulous examination of about 6000 Chinese characters. For instance, in Event class, they extracted 3200 sememes at the beginning. After necessarily merging, 1700 sememes were left for further classification and finally 700 sememes were made. In the following process, they made the necessary adjustment and extension when the set cannot satisfy the requirements. For example, a word with multiple concepts, and if the existing set of sememes failed to classify all the concepts, then they will have to adjust the tagging set. After 10 years of developments, over 2000 sememes had been created.

Data Format in HowNet

Knowledge dictionary is the heart of the entire HowNet system. Each entry in the dictionary consists four items by ignoring the language types.

W_X= word / phrase form
G_X = word / phrase syntactic class
E_X = example of usage
DEF = concept definition

X can be replaced as C standing for Chinese and E where indicates English language. E_C refers to examples of Chinese. For example, the entry ‘打’ is showed in below. W_C= 打 and W_E = buy are entries for Chinese and English respectively.

NO.=000001
W_C=打
G_C=V
E_C=~酱油,~张票,~饭,去~瓶酒,醋~来了
W_E=buy
G_E=V
E_E=
DEF=buy|买

All the examples (in E_C) and definitions (in DEF) are made manually. A complicate rule is made for demonstrating the concept in order to represent the inter-relation between concepts and their attributes. For example, concept delicious is defined as:

DEF=aValue|属性值,taste|味道,good|好,desired|良

The creation of HowNet has been more than a decade. Till now, 2199 sememes, 100168 Chinese words and 29868 concepts are produced. However, since the complexity and size of Chinese language, it still has a long way from the end.

2.2.3 Chinese WordNet (CWN)

Due to the reason that WordNet has more than two hundred thousand words, it is a huge workload work to translate it manually into the other languages. Thus, to reduce the effort of human being in manual translation, in CWN creation, three algorisms are used as the auxiliary tools, which are minimum distance algorithm, intersection algorithm and word co-occurrence algorithm. But since the low accuracy of the third one, only the first two algorithms were adopted. CWN is fully abided by English WordNet structure. They leave an empty mapping when a synset cannot find its equivalent Chinese translation. Since CWN is free to use and open source, we decided to import it into UKC. To understand the quality, an in-depth investigation and study of the creation procedure were made as in the following.

Minimum distance algorithm

Minimum distance algorithm is offered to calculate the minimum distances between English word explanation and its corresponding synset sense. The corresponding explanation with respect to the smallest minimum distance is the one that most similar with the synset sense. The Chinese words of this explanation are the Chinese translation for this synset. The dictionary used to looking up the explanations is ***American Heritage Dictionary***. The main procedure of this method is shown in the following:

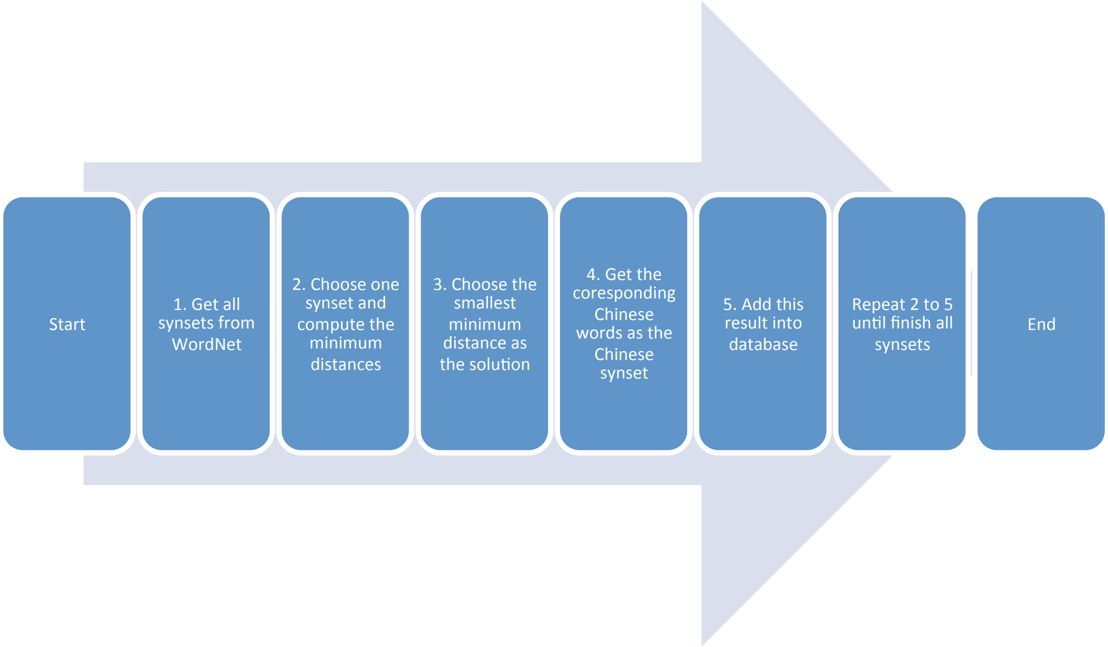

Start
1. Get all synsets from WordNet
2. Choose one synset and compute the minimum distances
3. Choose the smallest minimum distance as the solution
4. Get the coresponding Chinese words as the Chinese synset
5. Add this result into database
Repeat 2 to 5 until finish all synsets
End

Figure 8 Main procedure of minimum distance algorithm

As showed in Figure 8,

1. Extract all synsets from WordNet.
2. Choose a synset by sequence and for every word in this synset, get meanings from American Heritage Dictionary (English-Chinese).
3. Compute the minimum distances between this synset sense and every meaning respectively (In order to improve accuracy, in this algorithm, using 0.6 as add and remove operation cost; using 1 as modification cost).
4. Depending on the smallest minimum distance, choose the related Chinese word/words as the corresponding Chinese synset (only Chinese words, without Chinese gloss).
5. Add the Chinese synset into Database.
6. Repeat 2 to 5 until all the synsets are finished.

To understand better, a short example is in the follwing. A synset "the departure of a vessel from a port" only contains a single word "sailing". In American heritage dictionary, sailing means:

1.The skill required to operate and navigate a vessel; navigation.
航海术：驾驶和航行一条船所需的技巧；航行术
2. The sport of operating or riding in a sailboat.
帆船运动：驾驶或航行帆船的一项体育运动
3. Departure or time of departure from a port.
启航：离开港口；离开港口的时间

It is obvious that the synset sense is similar with the third meaning of sailing in the dictionary. The developers assumed that the minimum distance of two sentences indicates that those two sentences are the most similar ones. In order to verify this, we computed the minimum distance of "the departure of a vessel from a port" to every meaning of sailing in American heritage dictionary by applying minimum distance algorithm (using a single word as the minimum unit). After computing, we got:

[WordNet Gloss]: "the departure of a vessel from a port"

Sentence	Minimum distance
The skill required to operate and navigate a vessel navigation	0.666
The sport of operating or riding in a sailboat	0.622
Departure or time of departure from a port	0.475 {(0.6+0.6+0.6+1+1)/8 =0.475}

After computing, the minimum distance is the third meaning of sailing as we expected. Then, we choose the corresponding Chinese word "启航" as the Chinese translation for the synset "the departure of a vessel from a port". 1000 random samples were selected and 160 of them are not correct. In this case, the accuracy is 84%.

Intersection algorithm

Intersection algorithm is used to calculate the intersection among the translated Chinese sets (by pairs), which are translated from English word of synset. The result intersection set is the Chinese translation of this synset. The dictionary in which adopted this algorithm is Xdict. In this dictionary, it contains 177842 words, which are more than WordNet (155287 words). Also, editors separated the Chinese translation of every English word into different groups depending on the different senses. Those groups are separated by semicolon. For example, the translation of word "thing" is "物,东西;所有物;事,事情,事件;局面;事业;举动,行动;题目,主题;细节,要点".

Figure 9 is the primary procedure of intersection algorithm. In Figure 9, SEsyn={Esyn$_i$ | 1 ≤ i ≤ n}; Esyn$_i$={E$_{ij}$ | 1 ≤ j ≤ n }; CE$_{ij}$= {C$_{ijk}$ | 1 ≤ k ≤ n }; **SEsyn** is a set contains all synsets in the WordNet. **Esyn$_i$** is a synset in WordNet. **E$_{ij}$** is a single word in the synset Esyn$_i$ and it maps to the corresponding translated Chinese set **CE$_{ij}$**. **C$_{ijk}$** is a single Chinese word in CE$_{ij}$. **Intersection(CE$_{ij}$)** means get the Chinese intersection set of CE$_{ij}$. **Count()** is utilized to count the number of elements in this set. **Chinese_group()** is a function detecting ";".

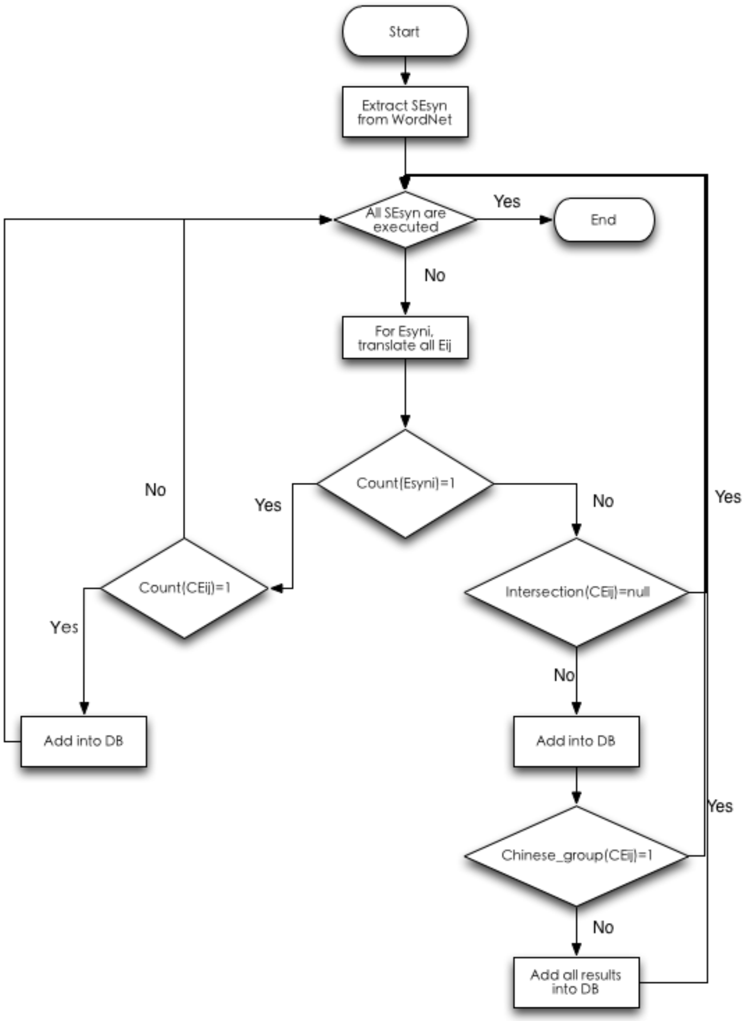


Figure 9 Main procedure of Intersection Algorithm

When calculating the intersection set, two situations used to determine that two words are equal:

1, they are the same word.

2, the minimum distance between those two words are smaller than 0.3. Generally, in Chinese, if two words' minium distance smaller than 0.3, it has a large possibility that those two words have the same meaning. Especially when those two words appear in the related translation group.

In the following example, we illustrate how to calculate the intersection set. In a synset {a vaguely specified concern; several matters to attend to; it is none of your affair; things are going well}, there are three relating words, which are: matter, affair and thing, and their translations in the Xdict dictionary are shown as follows:

□ affair n. 事情,事务;恋爱事件

□ thing n. 物,东西;所有物;事,事情,事件;局面;事业;举动,行动;题目,主题;细节,要点

□ matter n. 物质;麻烦,毛病;事情,问题;内容,素材

The intersection between affair and thing is “事情”. But since there are semicolons, the Chinese words in the same group should be added into this intersection set. The result is “事情，事务，事，事件”.

The intersection between affair and matter is “事情”. It also contains semicolons, so the result is “事情，事务，问题”.

The intersection between thing and matter is “事情”. Since the semicolons, the result is “事情，事，事件，问题”.

After combining those three sets together, we got the finial result “事情，事务，事，事件，问题”, and this is the finial result of the Chinese translation for this synset.

The number of synsets, which at least one word can be looked up in Xdict dictionary, are 80376. E.g., intersection algorithm is based on these 80376 words. The developers selected 1000 random samples and 192 of them are not correct. In this case, the accuracy is 80.8%.

Human translation

The goals of the former algorithms are to decrease the difficulty of human translation. The result of minimum distance algorithm is saved in database table wn_sdcv_chinese; the result of intersection algorithm is saved in database table wn_chinese. Since the accuracy of minimum distance algorithm is higher than intersection algorithm, during human translation, wn_sdcv_chinese has a higher priority for deciding the final translation, which means wn_sdcv_chinese table will be checking first. Synset is the smallest unit in the procedure of human translation. The system provides to linguists the synset content (English word, synset Id, POS) and its alternative translation. This alternative translation is calculated by the following procedure.

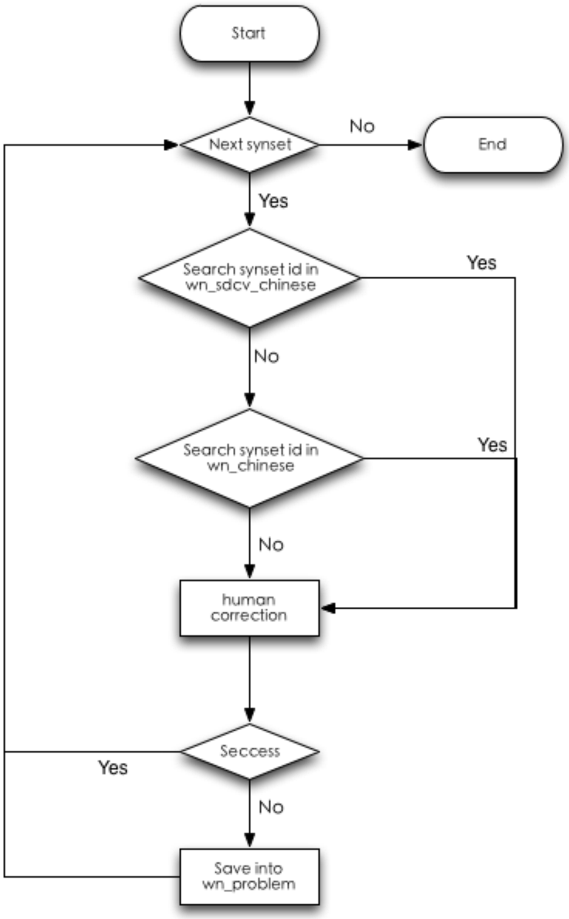


Figure 10 Main procedure of human translation

At first, the system shows the English gloss of this synset to the linguists, after that, shows the corresponding English word and the alternative translations. The linguistic has to decide the correctness of these alternative translations. The first choice of the alternative translation comes from database table wn_sdcv_chinese by search the synset id and if there is no such id exist, means that the minimum distance algorithm failed in translating this synset. In this case, system will search in table wn_chinese. If it still fails, the system will keep empty in alternative translation form, e.g. this synset has to be translated manually. If this synset is hard to translate or hard to decide the correctness, this synset will be added to wn_problem table and discuss by experts later.

2.2.4 Space Domain Translation

To evaluate the overall accuracy of the Chinese WordNet and discover possible problems during the UKC localization, an experiment that English to Chinese space domain translation manually was preformed. English space domain was generated from [16]. In this section, we will introduce the process of Chinese space domain translation, the CWN accuracy evaluation and the problems during the translation.

Before describing the localization process, some significant features of Chinese need to be declared. In addition to Mandarin Chinese, there are many dialects like Cantonese[9] and Taiwanese

[9] Wiki: https://en.wikipedia.org/wiki/Cantonese

Hakka[10]. Chinese language, also known as Mandarin, is the mother tongue for most Chinese speakers and written language for all Chinese people. Due to the different language families, English and Chinese have many significant differences. Our focus is on the semantic differences between vocabularies, rather than the difference on phonological or language families. There are several elements, lexical gap, associative meaning and extent. Lexical gap [17] refers to a kind of specified object or concept that is unique in one culture, but missing in the others. So, sometimes, it is hard to find the corresponding word. In general, lexical gaps between Chinese and English are always with culture aspects, for example:

Ecological culture: Some Chinese words “三伏”，“九伏” refer to the hottest weather in a year, which are hard or cannot find corresponding English equivalents.
Material culture: Some traditional food or clothes like, “粽子” (traditional Chinese rice-pudding), “长袍马褂”(robe and mandarin jacket).

Another situation is associative meaning. Associative meaning refers to the words and phrases that existing in both English and Chinese, but the meaning of these words and phrases cannot be completely overlapped. For example, in Chinese ‘magpie’ forecasts good news, so magpie is welcome by Chinese people. But on the contrary, in some countries, people think ‘magpie’ is too talkative and sometimes even symbolizes the stealing.

Generally speaking, the extent of English vocabulary is wider by comparing to Chinese vocabulary. The exact meaning of English words is decided by the context when translating Chinese and English, while the meaning of Chinese words is comparatively independent and stable. Taking kinship (family ties) as an example, English call those who are of the same generation with their parents “uncle” for male and “aunt” for female; and call the younger generation “nephew” for the boy and “niece” for the girl. This situation is totally different in Chinese where every person has a unique appellation in his family. Another famous case is ‘I’. ‘I’ needs to be translated into ‘朕’ when the target person is a king, and ‘我’ when the target is a normal people, etc.

2.2.4.1 English Space Domain

A domain can be defined as any area of knowledge or field of study that we are interested in or that we are communicating about. Domains may include traditional fields of study (e.g. medicine, physics), applications of pure disciplines (e.g. engineering, agriculture), any aggregate of such fields (e.g. physical sciences, social sciences) or capture knowledge about our everyday lives (e.g. music, sport, recipes, tourism). Our focus is on the domain Space. Notice that Space has always played a central role in all library classification systems.

The domain under examination is decomposed into its basic constituents, each of them denoting a different aspect of meaning. Each of these components is a facet. For instance, in Space daomain, the facets may include bodies of water, geological formations and administrative divisions. More precisely, a facet is a hierarchy of homogeneous terms describing an aspect of the domain, where each term in the hierarchy denotes a different concept. In the original library science ap-

[10] Wiki: https://en.wikipedia.org/wiki/Taiwanese_Hakka

proach, since the purpose is to classify bibliographic material, each concept denotes a set of documents while links between concepts in the facet hierarchies denote subset relations. In our approach, since the purpose is to describe Space in terms of real world objects, each concept may denote a class, an entity, a relation or an attribute, while links denote a much richer set of relations. For instance, in the former case the term river denotes the set of all documents about rivers, while in the latter case it denotes the set of all real world rivers. Concepts inside a facet are arranged by characteristics, i.e. according to their distinctive properties. For instance, since both river and brook are flowing bodies of water (their characteristic) they are arranged in the same facet, i.e. body of water, and at the same level of the facet hierarchy. When arranged together, siblings sharing the same characteristic form what in jargon is called an array of homogeneous terms.

The space domain is a large-scale geospatial ontology built using the faceted approach from the complete integration of *GeoNames* and *WordNet*, which is also known as space ontology. It currently consists of 17 facets, around 1000 concepts and 8.5 million entities. Facets include *land formation* (e.g., mountain, hill), *body of water* (e.g., sea, lake), *administration division* (e.g., state, province) and *facility* (e.g., university, industry).

2.2.4.2 Space Domain Translation

An experiment, which is a manual translation work of space domain, has been performed in order to evaluate the accuracy of CWN and discover the problems during the UKC localization process. The following dictionaries are adopted as references to judge the correctness of the Chinese translation.

- English WordNet 3.0
- Oxford Advanced Learners Dictionary 7th edition
- American Heritage Dictionary 3rd edition
- Web dictionary[11]
- Youdao dictionary[12]
- Xinhua dictionary[13] (The most authoritative Chinese dictionary)
- 朗道英汉字典 5.0 （Landau）
- 牛津英汉双解美化版 (Oxford)
- 英汉汉英专业词典 (English-Chinese Chinese-English Professional Dictionary)
-

The Translation of lemmas

Every lemma (a word in the synset of the concept) in the Space domain, the first step is looking up the lemma from CWN in order to find the corresponding Chinese synset. Generally, the glosses of Space domain are the same with CWN. Thus, we can find the corresponding Chinese synset by searching the gloss in the CWN database. For example, the gloss of lemma 'Railroad' is 'a line of track providing a runway for wheels', after searching this gloss in the CWN database,

[11] http://www.ichacha.net

[12] http://dict.youdao.com

[13] http://xh.5156edu.com

the ID is indicated, which is “103895665”. This unique ID plus “50000000” is the mapping to the Chinese synset. The corresponding Chinese synset for lemma ‘Railroad’ is “1，铁轨，2 轨道”. If there is no Chinese synset mapping to this ID, we call it Chinese-synset missing problem.

The second step is checking the correctness of this Chinese synset. Several dictionaries are adopted to validate the correctness of the Chinese synset. In CWN, Chinese synset is not luxuriant since every Chinese only contains one or two Chinese words in general. And, those Chinese words are always the first rank sense word for this concept. Then, for enriching our UCK Chinese synset, when some Chinese translations are presented in most dictionaries yet in CWN， we will add new translations beyond on the CWN translation after checking the correctness.

If the lemma is out of CWN, or the corresponding Chinese synset is not correct, we go to the third step. At this step, the Chinese translations mainly come from the dictionaries. For example, the entry ‘Opera house’ is not in the WN and CWN, but in most dictionaries, we can find this lemma and it has a similar sense. In this case, we chose the most frequent words of this sense as the Chinese translation. For this entry is “歌剧院”. There are some exceptions (professional vocabularies), in which the lemma can be found in neither CWN (WN) nor the dictionaries. In this situation, we use online dictionary which contains a lot of specialized dictionaries, for example, geography dictionary.

In some extreme cases, the word cannot be found in anywhere, we go to the fourth step, which is using direct translation. For instance, ‘Continental arch’ cannot be found in specialized dictionaries. ‘Continental’ in Chinese in this sense means “大陆的”， and ‘arch’ in this context means “拱门，拱形的”. Thus, we translated this word like “大陆拱”. But for the reason of ensuring the accuracy, we use the following mechanism. Searching this Chinese word on Google and Google scholar, if there are some matching keywords means that this translation is fine. If not, we try another translation. But, if this lemma is hard to find a matching, we go the last step -- put it into the gap.

The Translation of glosses

Similarly with lemma translation, in order to get the translation more authoritative, the gloss from dictionaries has the highest priority to be adopted. Our translation rule is that:

1) The keywords of the gloss from dictionaries matches more than 80% with the one from the space domain, then this gloss will be adopted. (This situation only occurred 5%-10% for 1000 words).
2) If it is not matching, we use a Chinese dictionary. Searching the Chinese translation lemma and checking whether the corresponding gloss is matching. (Less than 5%, only 10-20 lemmas were matched)
3) Otherwise, we translate the gloss manually.

2.2.4.3 Results

For keeping every aspect clearly during the translation process, we used some marks as assistance. The meanings of every mark are list in the following:

Color Yellow: means current concept, the gloss is different with the corresponding WordNet one.
Color Green: means the concept is duplicated.
Color Orange: means the concept has a problem.
Color blue: means that this lemma cannot be found in both WN and CWN.
Color red: means the corresponding Chinese cased word lemma in the CWN is not correct.

File name	Entry Number	Red	Blue	Yellow	Green	Orange
Abandoned facility	16	0	16	0	0	0
Administrative division	18	0	1	0	0	0
Agricultural land	20	1	5	1	0	1
Attributes	102	7	11	7	19	2
Barren land	7	0	5	1	0	0
Body of water	116	8	60	0	0	0
Facility part 1	99	7	26	3	0	0
Facility part 2	99	4	21	3	0	0
Facility part 3	81	1	24	7	1	1
Facility part 4	60	2	19	3	0	0
Facility part 5	19	0	12	3	0	0
Forest	6	0	1	0	0	0
Geological formation	200	9	87	12	0	1
Land	15	1	5	2	0	0
Plain	13	0	2	1	0	1
Populated place	13	1	9	1	0	0
Rangeland	9	0	0	0	0	0
Region	47	2	27	0	0	2
Relation	66	5	8	1	0	1
Seat of government of a political entity	6	1	0	0	0	0
Wetland	8	1	3	0	0	1
	1020	50	339	45	20	10

Table 3 Final result

As in Table 3, 1020 entities have been translated, 50 of which have problems. In these 50 entries, 21 of them belong to Chinese-synset missing problem; the others are semantic errors or typos. In those 1020 entries, 404 (Blue +Yellow + Green) were not from CWN. In this case, we got the final accuracy of CWN is around 85%.

1) UCK space domain errors

During the translation procedure, some problems are suspected to be errors. Basically, the suspected errors can be separated into three categories.

- **Examples non-related**

Example errors are the problem which those concepts' gloss containing the improper or irrelevant examples.

Country: The territory occupied by a nation*; "he returned to the land of his birth"*; "he visited several European countries"

High salinity: greater than normal in degree or intensity or amount; *"a high temperature"; "a high price"; "the high point of his career"; "high risks"; "has high hopes"; "the river is high"; "he has a high opinion of himself"*

Low salinity: less than normal in degree or intensity or amount; *"low prices"; "the reservoir is low"*

Submarine pass: the location in a range of seamounts of a geological formation that is lower than the surrounding peaks, *"we got through the pass before it started to snow"*

Sisal plantation: an estate that specializes in growing *banana.*

Inclination: the property possessed by a line or surface that departs from the vertical; "*the tower had a pronounced tilt"; "the ship developed a list to starboard"; "he walked with a heavy inclination to the right"*

- **Typo**

Typo is a kind of error that can be detected by the system automatically.

Petrolium basin: an area underlain by an oil-rich structural basin.

Mudflat: No gloss provided.

- **Duplication**

During the translation procedure, we found some duplication. Those duplications have the same concept ID, the same word ranking, e.g., they are not the words in different concepts. Those duplications will cause the inaccuracy problem when counting the finial result. We list them in the following:

- Phytoplankton
- Flora
- High
- Low
- Linear dimension
- Length
- Breadth
- Height
- High
- Low
- Depth
- Volume
- Elevation
- Phytoplankton
- Wilding
- Fungus
- Pest
- Acidity
- Classical architecture

- Clinic

2) Errors of CWN

CWN was created following the largely automatic methodology. All the translations were generated from some dictionaries automatically and then validated by human being. The validators of CWN are a group of bachelors, which means they have a high education level. During the translation process, we found some errors in CWN. By analysing and categorizing those errors, we believe that we can figure out some normal mistakes and problems that we will meet in the following work (crowdsourcing). An error indicates those Chinese translation (s), which is different with the meaning from the dictionaries. Generally, those errors can be split into the following categories.

- **Too general**

Higher-level meaning means the concept is assigned with the higher-level Chinese meaning. Normally speaking, this is correct without the specific requirements. However, in UKC, we need to distinguish the nuances between concepts, e.g., the specific meanings are required. For example, the concepts ,which contain word 'Bay' and 'Gulf' respectively, have the same meaning in Chinese "海湾" in all dictionaries. But if we focus on the glosses of those two concepts *'an indentation of a shoreline larger than a cove but smaller than a gulf'* and '*an arm of a sea or ocean partly enclosed by land; larger than a bay*', we can find the difference between them. In this case, adding an adj. can help us to distinguish them.

- **Typo**

Typo is a normal mistake.

- **Ambiguity**

Ambiguity means a kind of error that assigned wrong meaning (this meaning is one of the translations in the dictionaries) to the synset. For example, the word 'town' has the Chinese translation '市区,镇，城镇'. In the concept *'an urban area with a fixed boundary that is smaller than a city*' we choose '镇，城镇' as our translation instead of '市区'.

- **Wrong POS**

This is an error similar with ambiguity. The word is translated with the wrong meaning belongs to different PoS. For example, the concept '*the persons (or committees or departments etc.) who make up a body for the purpose of administering something*' was translated in '行政', '管理' which are verbs. However, our concept is a noun, e.g. the Chinese translation in CWN is not correct.

3) Issues and Discussions

Some issues were found during the translation, we discussed each of them in the following.

- **Gaps or phrase**

For example, cased word concept 'Loch', *'a long narrow inlet of the sea in Scotland (especially when it is nearly landlocked)'*. Obviously in Chinese we don't have a word to describe the concept of Scotland inlet. However, as we seen in UKC there are a lot of 'adj.+ n.' phrases such as abandoned + n., lost + n. and section of n., e.g. we can use format 'adj.+ n.' or 'adj. + adj.+ n' to express the concept of English word ‘Loch’. In some dictionaries, they treated this kind of case as ‘(ɑdj.) n’, which are the same with CWN. For ‘Loch’, it is ‘（苏格兰的）湖泊’ [meaning is (Scotland) inlet].

- **How to define a gap**

Chinese language has a strong ability to express meanings, thus, how to define a gap becomes a problem. Obviously, a lot of concepts are missing in the dictionaries. How to decide it is a gap rather than the contributors’ knowledge less coverage becomes a problem.

- **Assigning word sense rank**

Assigning word sense rank appears as a difficult task to accomplish since the *Language Translators* contribute the results separately. In the translation work, they were aware of the fact that concepts translated by others might have the same word. But it remained obscure until the whole translation task was finished.

- **Correctness**

What is a correct translation? For instance, if we translate the concept word ‘Chain of ponds’ directly, it is “池塘链”. Unfortunately, it is hard to decide if it is correct or not. In our experiment, we adopt some famous resources as our baseline, such as, Google image, Google scholar and Wikipedia. In this example, the translation ‘池塘链’ are missing matching from all those famous resources, thus, we decide this translation is wrong.

In this work, we introduced an experiment that evaluating Chinese WordNet. The data set we used to evaluate CWN is English space domain and some Chinese-English dictionaries are adopted to judge the correctness of the translation. After experimentation, we calculated the accuracy of CWN is around 83% and some issues and discussions are proposed. Furthermore, we believe that we can reduce the effort of human being by integrating some existed resources.

2.2.5 Chinese LKC Creation

In order to save our efforts, we will attempt to import an existing Chinese linguistic resource at first. After investigation, we found that there are some linguistic resources based on Chinese language. The famous are Chinese WordNet (CWN), HowNet and Chinese WordNet (traditional version). Three conditions are primarily considered, including the quality, the dataset structure and the user license. Through studying, we learned that HowNet has the highest quality; Taiwan Chinese WordNet is produced in traditional Chinese, which is different with the simplified Chinese; Chinese WordNet has a higher quality and express in simplified Chinese. As a result, the first two resources are abandoned, since 1), the data structure of HowNet is different with WordNet, lacking the mapping between them. 2), the differences between traditional and simplified Chinese are not only in the characters, but also in the expression. In such case, the transformation of expression from traditional Chinese to simply Chinese could cause some unexpected problems. Thus, the existing resource that we plan to import is simplified Chinese WordNet (CWN), which is free to use and fully following English WordNet’s data structure, and we found the glosses mapping between CWN and the UKC, which would guarantee the success of importing.

According to the factors mentioned above, we believe that importing CWN will be a good starting point for the entire translation work. We did not find the direct mapping of ID. However, Chinese WordNet imported English glosses from English WordNet, and most of the glosses in UKC are also imported from English WordNet. That is, the majority of UKC concepts should find its corresponding Chinese Synset in CWN. We abide by the following rules when importing,

- Following the data structure of UKC. In order to keep the consistency, we do not import any semantic relations from Chinese WordNet (CWN).
- Only import these concepts that both existed in UKC and CWN
- Use English glosses plus examples as the mapping medium
- Do not import when this concept in UKC has a Chinese synset

We use glosses plus examples as the mapping medium. That because we found that some glosses are too similar, which means using glosses only will cause multiple mapping issues. For example, the following three glosses are too similar. When we use gloss ‘open to question’ as the mapping medium, it maps to the other two:

- **incontestable, indisputable, undisputable**: not open to question; obviously true
- **unquestionable**: not open to question
- **equivocal**: open to question

2.2.5.1 Importing Steps

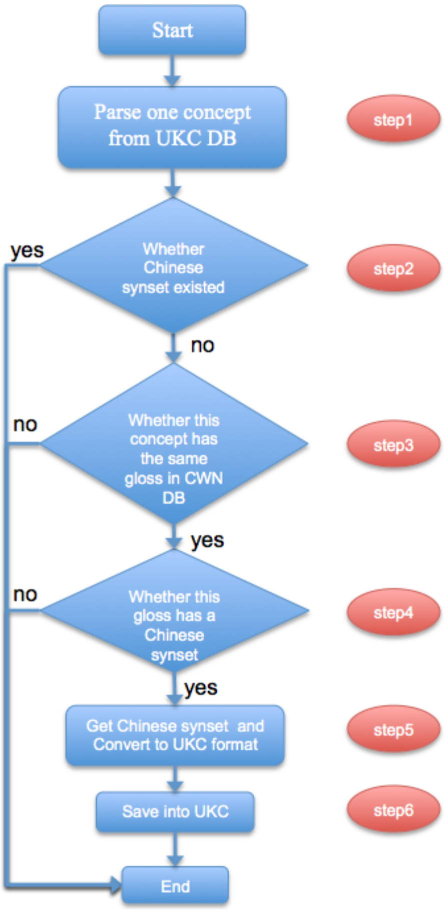

Start
Parse one concept from UKC DB
step1
yes
Whether Chinese synset existed
step2
no
no
Whether this concept has the same gloss in CWN DB
step3
yes
no
Whether this gloss has a Chinese synset
step4
yes
Get Chinese synset and Convert to UKC format
step5
Save into UKC
step6
End

Figure 11 Parsing algorithm of a single concept

Figure 11 shows the steps for importing a Chinese concept. It has 6 steps, which are,
Step 1: Select one concept from UKC database according to the ID order (from 1 to 111244)
Step 2: Check whether the current concept existed a Chinese synsets
Step 3: Check whether we can map this concept to CWN, because some glosses in UKC were not coming from English WordNet.
Step 4: Sometimes, in CWN, some Chinese synsets are empty.
Step 5, Get word ranks and Chinese words from CWN, and then convert to UKC format (Some Chinese words in CWN database are null, need to filter these empty words before converting).
Step 6, save this new Chinese synset into UKC database.

2.2.5.2 Results

In this import work, we import new Chinese words, new senses, POS, new Chinese synsets and word ranks. Since CWN is out of Chinese glosses, we did not import Chinese glosses. Before importing, UKC database (Version 2.5.0) contains 110968 concepts. 814 of them have Chinese synset. These 814 synsets contain 983 Chinese words and 1048 senses. After importing, 96636 (87% of UKC concepts) are covered with Chinese synsets. We imported 95822 new Chinese synsets, 88854 Chinese words and 119096 word senses. 9981 concepts are not mapped to CWN. In these mapped concepts, there are 4885 empty Chinese synsets.

We randomly select 100 concepts twice from UKC database in order to obtain the average accuracy.

At the first time, 14 of them have issues, in which,

- **8 concepts**: gloss is not mapped. (7 missing mapping since CWN is out of glosses. 1 because this gloss in UKC is not from English WordNet.
- **5 concepts**: CWN has these English glosses but without Chinese synset.
- **1 concept**: Wrong format in CWN, '愉快地（的）、迅速地（的）' should be saved in two different rows in the database.
- **Other concepts**: are all mapped and imported correctly.

Second time, 14 of them have issues,

- **10 concepts**: are not mapping.
- **concepts:** in CWN have glosses but without Chinese synsets.
- **Other concepts:** are all mapped and imported correctly.

After evaluation, we found that all the issues are coming from CWN database, besides that, the mapping accuracy is around 100%. During the importing procedure, we got the following issues in CWN.

- **Empty words**

  Empty words are these words that saved as " " in CWN database. There are 784 empty words for these mapped concepts.

- **Empty synsets (not lexical gap)**

  Empty synset is a synset that does not offer any information. There are 4885 empty Chinese synsets in mapped concepts.

- **Irregular format of words**

  Some Chinese words in CWN database are saved in inappropriate formats, for example, '愉快地（的）、迅速地（的）', '被阉割的男歌手(为了保持女高音或男高音那样的高音而在青春期前被阉割的男歌手) ', ' 至上' (words with blanks).

We found two reasons caused not match problem:

- **Some glosses of CWN and UKC come from different resources**

  Some glosses in UKC do not belong to English WordNet. For example: the gloss of *maltese cat* in UKC is 'a short-haired bluish-grey cat breed'; In English WordNet 3.1 is 'a term applied indiscriminately in the United States to any short-haired bluish-grey cat')

- **Typo**

  UKC: the process of becoming rigidly fixed in a conventional pattern of 'thought' or 'behaviour'

> CWN: the process of becoming rigidly fixed in a conventional pattern of ‘thought’ or ‘behavior’

Our goal is translating UKC to Chinese. Before that, in order to save our efforts, we intend to import an existed WordNet structure based Chinese linguistic resource. We found Chinese WordNet satisfied our requirements. But, there is not an explicit ID mapping between UKC and CWN. Thus, in our work, we utilize gloss plus example as the mapping medium. In this work, we imported 98415 Chinese synsets and 88943 Chinese words with 119722 senses. Also, we presented the selection of the Chinese linguistic resource, the basic idea of importing work. Furthermore, we did an experiment to evaluate the importing accuracy, which is close to 100%.

Chapter 3

3 Game with a Purpose

3.1 Game with a Purpose

Game with a Purpose (GWAP), also known as social Game-based Human Computation or Crowdsourcing via games, is a kind of theory in which we can get benefits when user is playing a human computation game. Those benefits are as the side effect during games. Many human computation games have been developed already such as ESP game [21], Peekaboom [23] and Verbosity [24]. Although they are solving different problems, all of them do get side effects during the games.

Computer technique has advanced dramatically over the last five decades, but it is still hard to solve some problems, e.g., artificial intelligence (AI) problems, that most humans take for granted. By this scenario, human computation [18] is proposed, which is a technique that makes use of human abilities for computation tasks where hard for computers but trivial for humans. Currently, the human computation systems can be classified as [19]: 1) Initiatory Human Computation: 2) Distributed Human computation: 3) Social Game-based Human Computation. A case for initiatory human computation is CAPTCHA [20], which is a well-known and wildly used security pass. Some cases of distributed human computation, which is also known as Crowdsourcing, are Wikipedia[14], Yahoo! Answers[15] and Amazon Mechanical Turk[16].

Terminology of 'crowdsourcing' was firstly indicated by Jeff Howe in 2006 used to describe a new business model, in which tasks are distributed through Internet [21]. Brabham further defined crowdsourcing [22] and created a typology of crowdsourcing [23] basing on unsolidified theoretical knowledge of crowdsourcing situation. Due to the extraordinary number of Wikipedia contributors, it has been demonstrated that crowds can outperform linguists in terms of coverage. The research shows that the 'wisdom of crowds' based resources are not generally superior to 'wisdom of linguists' based resources. However, it is worthwhile to note that collaboratively created knowledge sources are strongly competitive to linguistic knowledge sources on the majority of datasets [24]. Furthermore, crowdsourcing has demonstrated its advantages in the field of translation work, which strengthens our confidence that Crowdsourcing will be a good solution. For example, Yeeyan[17] is the largest community translation site in China with more than 400,000 registered users and 30,000 community translators. Community translators use their spare time and multilingual skills to translate interesting stuff they read on the web and share them with Chinese readers. Wordreference [18] a free and multilingual online dictionary, uses its forum to discuss and collect words and meanings from its users, and those words and meanings

[14] Wikipedia website: https://www.wikipedia.org/

[15] Yahoo answers! Site: https://answers.yahoo.com/

[16] Amazon Turk Site: https://www.mturk.com/mturk/welcome

[17] Yeeyan website: http://g.yeeyan.org/

[18] Wordreference website: http://www.wordreference.com/

will be used for their online dictionaries after verifying by experts or seniors. Duolingo[19], a free language study tool, collects language translation from its students when they are practicing the language. Other examples like Facebook[20] and Twitter[21], crowdsourced language translations of their websites based on their huge amount of users. One of the significant challenges of crowdsourcing is how to encourage people to contribute, e.g. incentive mechanisms. Generally speaking, the incentive of crowdsourcing includes Payment, Altruism, Enjoyment, and Reputation, etc. [25] While, Scekic classified it into Pay per performance, Quota/ Discretionary bonus, Deferred compensation, Relative, etc. [26] Game with a Purpose is a such kind of Crowdsourcing exploiting enjoyment as the incentive.

3.2 GWAP Classification

GWAP is a kind of theory in which we can get benefits when a user is playing a human computation game. Those benefits are as the side effect during games. Many human computation games have been developed already such as ESP game [27], Peekaboom [28] and Verbosity [29]. Although they are solving different problems, all of them do get side effects during the games. Depending on different game structures, in [27] Luis Ann categorizes Human computation games into three categories: output-agreement game, input-agreement game and inversion-problem game. We introduce each of them in the following respectively.

Output-agreement

Output-agreement is a game type where the same input is provided to all players and outputs should be based on the common input. They win when they provide the same outputs. An example for output-agreement game is ESP game. In ESP game (Figure 12), for each round, two players are given the identical picture (in this example is an image with a dog), they do not know each other and they cannot communicate. They are requested to type some words to describe this picture. They win once they typed the same word for a picture. The mapped input can be asymmetric. As in this example, player 1 typed dog at time 0:03 and player2 typed dog at 0:11. In this case, dog will be used as a tag of this picture.

[19] Duolingo website: https://www.duolingo.com/

[20] Facebook translation: https://www.facebook.com/

[21] Twitter translation: https://twitter.com/?lang=en

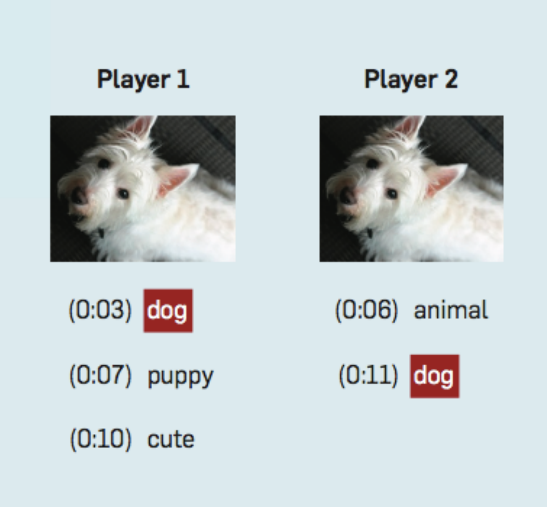


Figure 12 An example of ESP game [27]

Input-agreement

All players are given inputs that are known by the game (but not by the players). The players have to guess whether they got the same input via describing it (output) for each other. They win when they guess correct whether they are or are not in the equivalent input. An example for this is TagAtune [30]. Two players are provided with a same or different song. They have to describe it for each other (Figure 13). At the end, they are asked to determine whether they listened the same some or not based on the written descriptions. They win when they guess correctly. As shown in Figure 14, after getting the question, play1 and player 2 are making descriptions and they have to make a decision at the end.

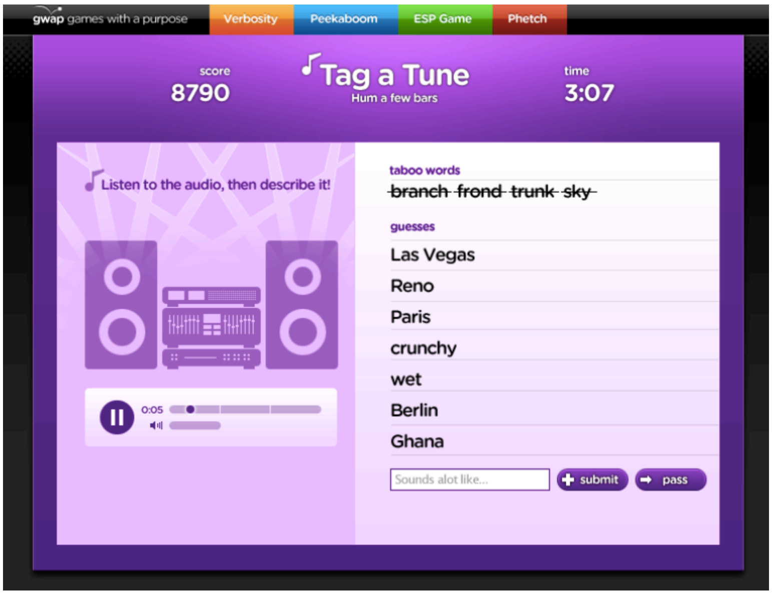


Figure 13 screenshot of TagATune

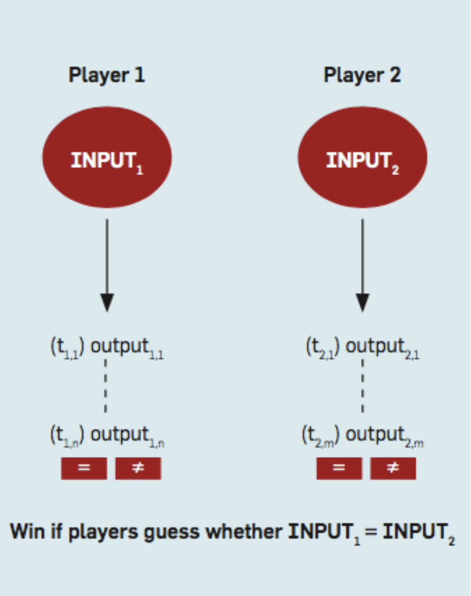

Player 1
Player 2
INPUT_1
INPUT_2
$(t_{1,1})$ $\text{output}_{1,1}$
$(t_{1,n})$ $\text{output}_{1,n}$
$(t_{2,1})$ $\text{output}_{2,1}$
$(t_{2,m})$ $\text{output}_{2,m}$
=
≠
=
≠
Win if players guess whether $\text{INPUT}_1 = \text{INPUT}_2$

Figure 14 Mechanism of input agreement method [27]

Inversion-problem

Two players compose a team. An input is only provided to one player. This player has to produce output for the other player. The second player has to make a guess that what is the input. If the second player can guess the input correctly, then they win the game. An example for this is Verbosity [29]. As shown in Figure 15, Player 1 produced some output with the given input. Player 2 produced outputs based on player 1's output. They win then player 2's output is consistent with the player 1's input.

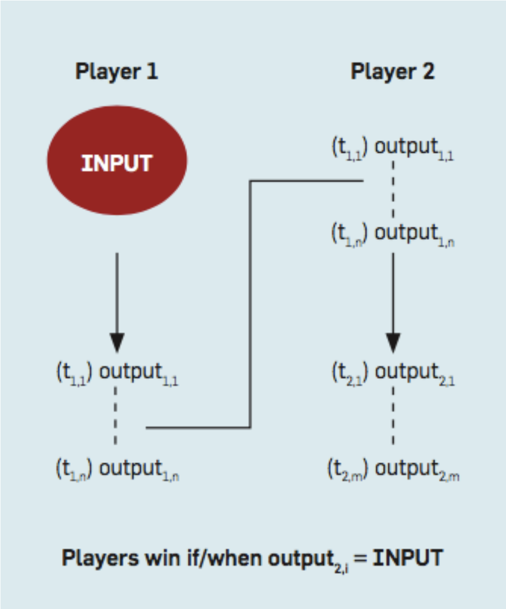
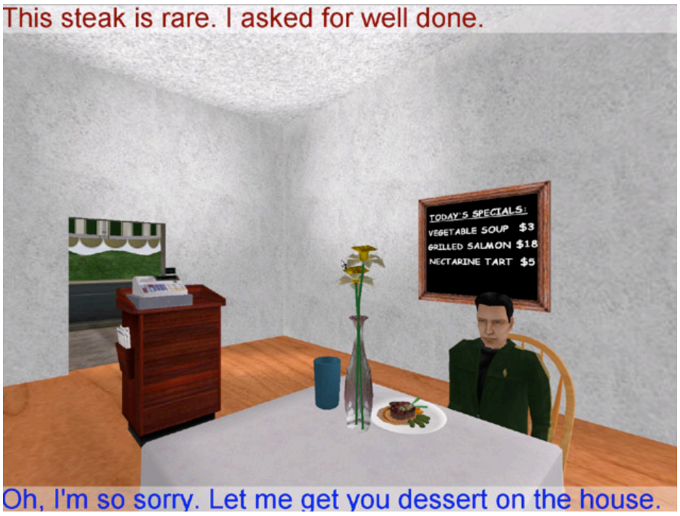

Player 1
Player 2
INPUT
$(t_{1,1})$ $\text{output}_{1,1}$
$(t_{1,n})$ $\text{output}_{1,n}$
$(t_{1,1})$ $\text{output}_{1,1}$
$(t_{1,n})$ $\text{output}_{1,n}$
$(t_{2,1})$ $\text{output}_{2,1}$
$(t_{2,m})$ $\text{output}_{2,m}$
Players win if/when $\text{output}_{2,i} = \text{INPUT}$

Figure 15 Inversion-problem [27]

Output-optimization

Man-Ching Yuen and his collages defined a discrete category, output-optimization [32]. By considering the verification method (otherwise known as quality control), it can be separated into symmetric verification game and asymmetric verification game. All players are given the same input and their outputs are the hints of other players' outputs. One example is the Restaurant game [31], it is a video game that aims to collect social behaviour and language from players in the scenario of a restaurant. Each round of the game takes about 10 minutes. The game automatically finds a partner when a player logins in. They are asked to act as a customer and waiter respectively, and they need to simulate actions like in a real restaurant. The collected data will be used in a new game, e.g. train a conversational virtual agent via machine learning. Everyone who plays the Restaurant Game will be credited as a Game Designer.


This steak is rare. I asked for well done.

Oh, I'm so sorry. Let me get you dessert on the house.

Figure 16 Screenshot of The Restaurant game

Futhermore, in [33], Chien-Ju Ho and Kuan-Ta Chen proposed two fundamental verification mechanisms, simultaneous verification and sequential verification, in human computation game systems. Using game mechanism can separate human computation game into Collaborative Game, Competitive Game and Hybrid Game. And also, if we focus on the number of players, every game can be either a single player game or multiple-player game.

3.3 Related Human Computation Games

Games with a purpose have been wildly used in many domains such as, Foldit [34], in which non-scientists players are salving protein structure prediction problems, ESP game, where players are labelling images with words, Page Hunt, which is used to improve search engines. And MobiMission [35] is used for geospatial tagging systems.

While, similar to our work are games with annotation-based linguistic such as word sense disambiguation and create or validate common sense knowledge. Infection [36] is a video game to validate semantic knowledge bases as shown in Figure 17. In this game, some humans are infected, but not transfer into zombies yet. The problem is that they look at the same. So, in order to protect the city, players are required to recognize these humans by asking some common senses. They use BabelNet [37] as the case study, which is a large-scale multilingual semantic linguistic resource. But, validating concepts between multilingual, which is our task, is trickier than validating common sciences. We think a video game is not suitable for our case. First, in a human perspective, common senses are more understandable than concepts. Put these multilingual concepts into a video game is not fun. Second, the target players for solving our problems should speak at least two languages.

A similar game working with common sense is Concept game [38], which is a fast-paced slot machine game aiming to validate the common sense that collected by automatic algorithms. When a player pulls the lever, the game randomly generates a common sense to the player. The player has to judge whether this is meaningful or meaningless within a specified time.

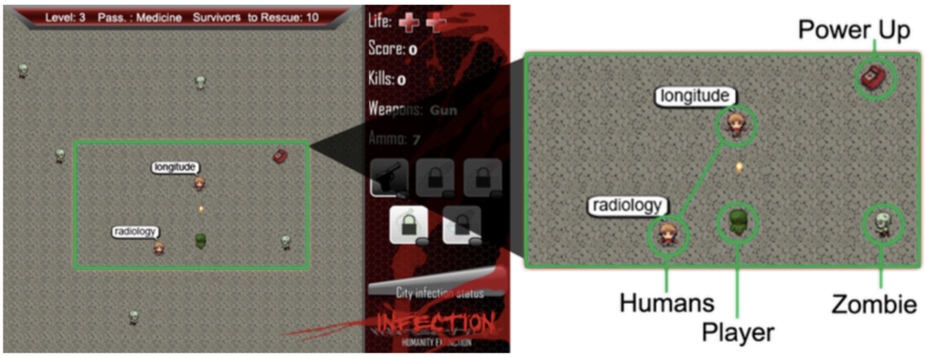
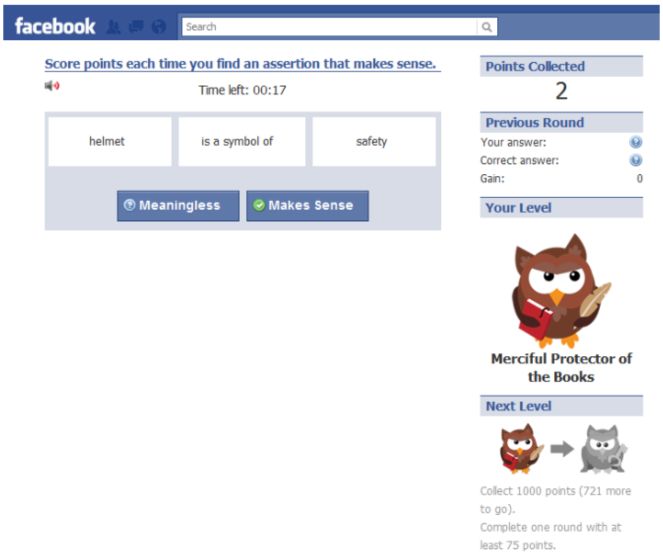


Figure 17 Screenshot of Infection [36]


facebook

Search

Score points each time you find an assertion that makes sense.

Time left: 00:17

helmet	is a symbol of	safety

Meaningless | Makes Sense

Points Collected

2

Previous Round

Your answer:

Correct answer:

Gain: 0

Your Level

Merciful Protector of the Books

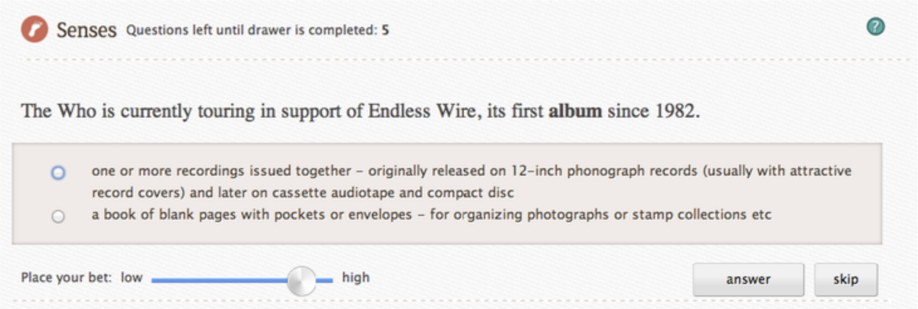
Next Level

Collect 1000 points (721 more to go).

Complete one round with at least 75 points.

Figure 18 Screenshot of Concept Game [38]

Obtaining gold standard data for word sense disambiguation is costly. So, people are attempting to use GWAP as a solution. Wordrobe [39] is such a game aiming to collect word sense disambiguation corpus. In this game, question is a sentence with a highlight word(s), and points are the possible senses of this highlight word. A player is required to read the question and understand the meaning of the highlighted word at first, and then the player has to select one option and place a bet. Game result is given by calculated on the basis of the answers from other players. Thus, the score of a player keeps updating even when they are not playing. Finally, they use absolute vote method to find the best sense for the word in each sentence.


Senses Questions left until drawer is completed: 5

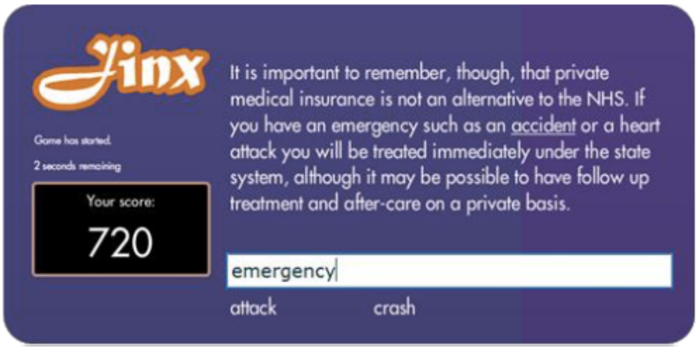
The Who is currently touring in support of Endless Wire, its first **album** since 1982.

○ one or more recordings issued together – originally released on 12-inch phonograph records (usually with attractive record covers) and later on cassette audiotape and compact disc

○ a book of blank pages with pockets or envelopes – for organizing photographs or stamp collections etc

Figure 19 Screenshot of Wordrobe [39]

Jinx[40] as shown in Figure 20 is another game that aiming to collect word sense disambiguation data. Different with Wordrobe, Jinx is a two-player game and following ESP game pattern, that is, an output-agreement game. In this game, two players are assigned as a pair randomly. Players are asked to produce some words/phrases for the highlighted term based on the given sentence. Two players are given an identical sentence, and they win when they produce the same output. Same with ESP game, the mapped outputs do not need to be typed at the same time.


It is important to remember, though, that private medical insurance is not an alternative to the NHS. If you have an emergency such as an accident or a heart attack you will be treated immediately under the state system, although it may be possible to have follow up treatment and after-care on a private basis.

emergency

Figure 20 Screenshot of Jinx [40]

3.4 Existing Chinese Word Games

We want to use existing game instead of creating a new one in order to save effort. We found that errors are from Chinese language part, and then the target player should be Chinese player. Besides, since the content of a linguistic resource is based on words, the focus should be related to word/konwledge games.

At the beginning, we need to understand what Chinese knowlegde games are playing with. Since Chinese language is totally different with alphabetic ones, here, we introduce Chinese language briefly. Chinese as the most spoken language all over the world (over 1 billion Mandarin Chinese speakers), is the only official language in China. Unlike languages with alphabetic such as English, the reading and writing of Chinese present unique features. We call the single characters like “中”, “国” as Chinese characters, and characters like “中国”, “山水”, “衣锦还乡” as Chinese words. Chinese characters are ideographic rather than phonetic. The characters originated with the Oracle Bone Script, and are continuously evolving in both shape and writing style. Figure 21 shows the evolution steps of Chinese character. Chinese characters are the basic semantic components in the Chinese language, which is completely different from English. In the character point of view, a character in Chinese could be equivalent to a word in English, such as “水” (water), and “山” (mountain). A Chinese word is usually equivalent to a phrase in English either, such as “热水” (hot water), “电话”(telephone) and “高山” (high mountain). In the pronunciation point of view, many characters or words have the same pronunciation, that is, homonyms, such as “一”, “衣”, and “医”. In the meaning point view, a Chinese character or a word could have several meanings, that is, polysemy, for instance, “水分”.

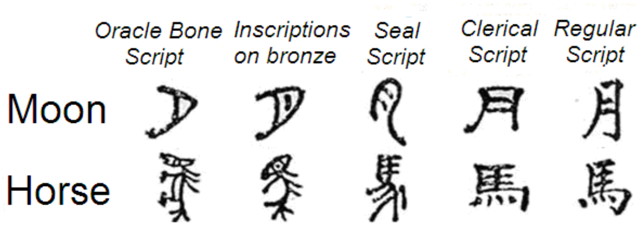


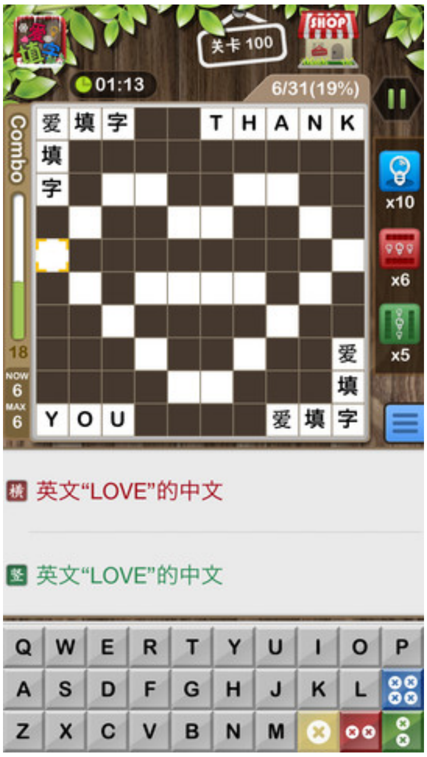
Figure 21 Evolution Steps in Chinese Characters

According to the facts we discussed above, Chinese word/knowledge games can be classified into 3 categories, which are playing with ‘Chinese characters structure writing’, ‘pronunciation’, or ‘meaning/knowledge’ (e.g. idioms, couplet, proverb, etc.) respectively. As our objective, our main focus should be on the games of playing with ‘meaning’. Word games are often designed to test language ability or to explore its properties, and generally engaged as entertainment, but have been found to serve an education purpose as well. Chinese word games are the word games in Chinese language. It has been a long history, which can be traced to thousands of years ago. In this Section, we will introduce several well-known Chinese word games playing with meaning/ knowledge, which are Chinese crosswords game and Chinese puzzle game, couplet game, solitaire game, idioms game, drinking game and vocabulary challenge game.

3.4.1 Chinese Crosswords Game

Since the features of the Chinese language, which are different with English, Chinese crossword games are using Chinese characters as the smallest unit. Normally speaking, a Chinese word is shorter than an English word. For example, English word ‘love’ contains 4 letters, but in Chinese it is ‘爱’ or ‘爱情’, which contains only one or two Chinese characters. At first glance, Chinese crossword game is similar to the English one in the game format. But since a Chinese word is more expressive, its involving content is greater than an English word. A Chinese crosswords game contains not only words, but also common senses, idioms, proverbs, poetries aphorisms, ancient and modern, etc. For example, as shown in Figure 22, at the first row, Chinese ‘爱填字’ means ‘ we like to play crosswords games’ in English. Therefore, the contents of Chinese crossword game are particularly rich, more interesting and challenging.

If we focus on game forms, the Chinese crossword game can be subdivided into graphical and non-graphical game. Graphical game arranges all entries into a certain meaningful graph. This graph can be symmetrical patterns, animals, plants, houses, text pictures, etc. As showed in Figure 22, it is a heart. Non-graphical game is a crossword game without graphical restrictions. Certainly, the primary focus of a crossword game should be the game content.



横 英文“LOVE”的中文

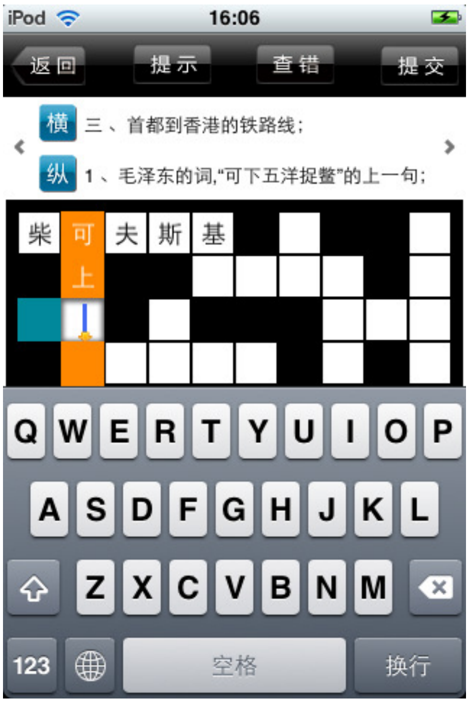
竖 英文“LOVE”的中文

Figure 22 Graphical Chinese Crossword[22]

By considering game content, Chinese crossword game can be separated into non-thematic game and thematic game. Non-thematic game does not have restrictions on the game topic, and generally, the content is common sense. Thematic game constrains all the game entries associated with one topic. For instance, full of poetries, movies, human names, novels, locations, etc.

‘小强填字’ is a famous non-thematic Chinese crossword game. ‘小强填字’ was published by the newspaper Southern Weekly at the beginning of nineties of last century and published the web app in 1999. Its publisher said that it is the first Chinese crossword game in China. We are not sure whether it is the real first one, but it is indeed the most famous one. As showed in Figure 23, in October 2009, they published iPhone version and got the high rating on App Store. Content of Southern weekly crossword game does not have restrictions, which means its content is various. Such as news, poetries and human names.

[22] ITunes Store: https://itunes.apple.com/cn/app/ai-tian-zi-zui-jing-mei-zhong/id565958696?mt=8



横 三、首都到香港的铁路线；

纵 1、毛泽东的词,"可下五洋捉鳖"的上一句；

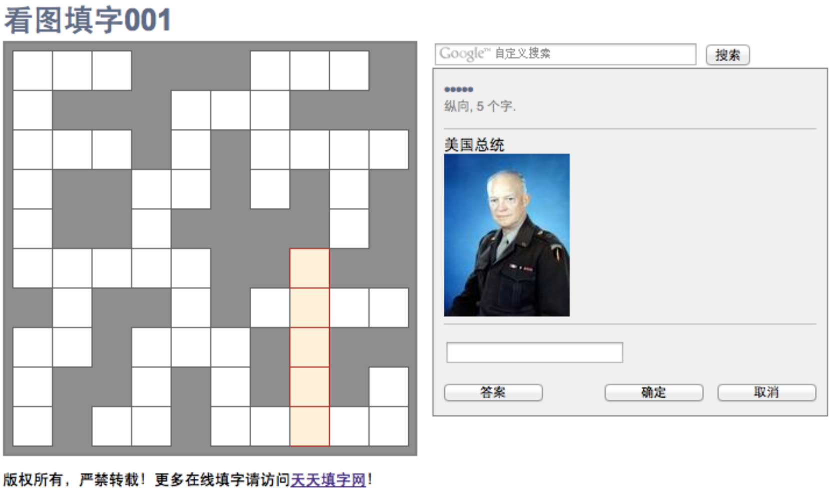
柴 可 夫 斯 基

上

Figure 23 Southern weekly mobile version[23]

'南方体育' is engaging thematic Chinese crossword game. The meaning of '南方体育' is southern sport, a famous sport newspaper in China. There is a sport crossword game on this newspaper. It will be apparent that whether you are a comprehensive sports enthusiasts after filling in one puzzle. Just those footballers' long names are the challenges.

Depending on the clue form, the Chinese crossword game can be separated into two types image-based game and language-based game. Image-based clue crossword game (Figure 24, the clue is an American president), as its name, is a crossword game using the image as the clue. This image can be a person, a building or a meaningful picture, etc. And language-based clue game (Figure 25) is using the Chinese language as the clue.



看图填字001

●●●●●

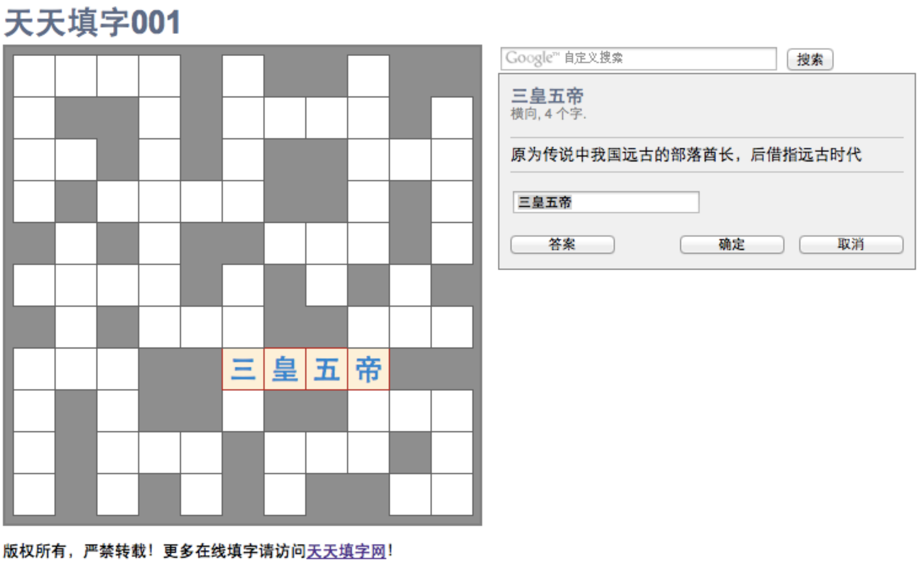
纵向, 5 个字.

美国总统

版权所有，严禁转载！更多在线填字请访问天天填字网！

Figure 24 Image-based clue game and the image title "美国总统" means US president. [24]

[23] ITunes store: https://itunes.apple.com/cn/app/xiao-qiang-tian-zi-mian-fei-ban/id347542956?mt=8



天天填字001

三皇五帝
横向, 4 个字.

原为传说中我国远古的部落首长，后借指远古时代

三皇五帝

版权所有，严禁转载！更多在线填字请访问天天填字网！

Figure 25 Language-based clue game[25]

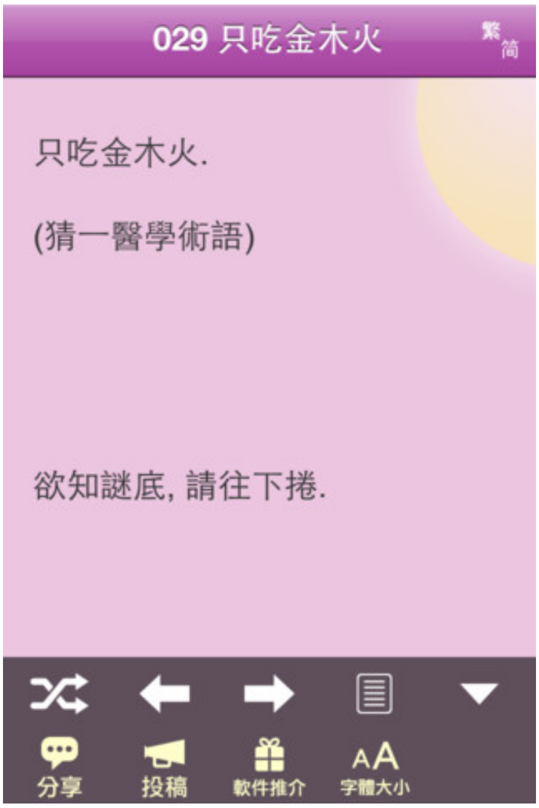
3.4.2 Puzzle Game

In China, puzzle game is derived from a traditional activity ‘Guessing lantern riddles’. It is an essential part of the Lantern Festival (A traditional Chinese festival). Lantern owners write riddles on a piece of paper and post them on the lanterns. If visitors have solutions to the riddles, they can pull the paper out and go to the lantern owners to check their answer. If they are right, they will get a small gift. The activity emerged during people’s enjoyment of lanterns in the Song Dynasty (960-1279 A.D.). As riddle guessing is interesting and full of wisdom, it has become popular among all social strata.

Now, Guessing lantern riddles is not only for the Lantern Festival, but also on newspapers, magazine, TV and even web and mobile application. In the following are the Guessing lantern riddles on iPhone and Android respectively (Figure 26 and Figure 27).

[24] Crossword games :http://www.tzgame.net/archives/191.html

[25] Crossword games : http://www.tzgame.net/archives/470.html



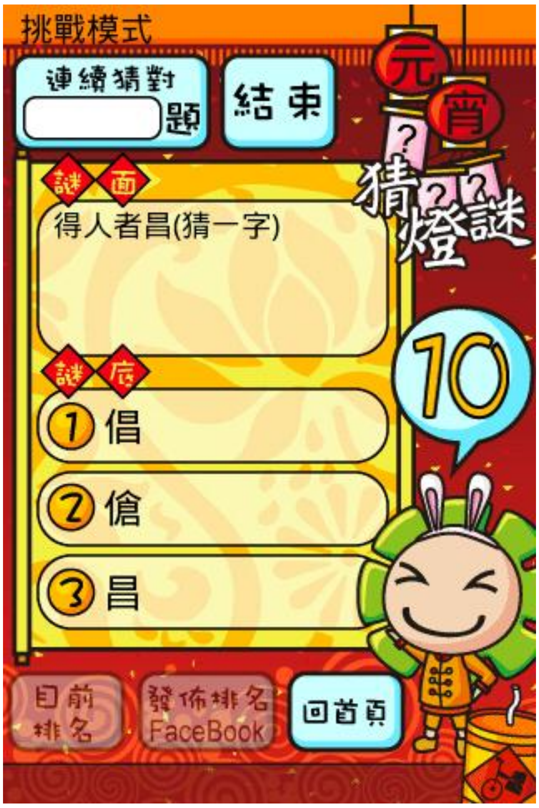
029 只吃金木火

只吃金木火.

(猜一醫學術語)

欲知謎底, 請往下捲.

Figure 26 iPhone Application[26]



挑戰模式

謎面

得人者昌(猜一字)

謎底

① 倡

② 傖

③ 昌

Figure 27 Android Application[27]

People must guess the answer from a word, a poem or a phrase, guessing riddles are as hard as fighting with a tiger, so that lantern riddles have another name 'Lantern tigers'. Some lantern riddles as example are in the following:

English Example:
Riddles: What month do soldiers hate?
Answer: March
Hint: Since in English, "March" has the meaning "a procession of people walking together" (This definition from English WordNet).
Chinese Example:
Riddles: 日复一日（猜一字） (English translation: day after day(a word))
Answer: 昌
Hint: the structure of '昌'is two "日" overlaid. '日' means 'one day' in English

Now, the clue of the puzzle is not only committed in literal, but also could be a picture. And the topic of the puzzle should therefore not be constrained. It could be a brand, person, movie, location, etc. There is an image puzzle game, which is named '看图猜成语'（Guessing Idiom from Picture）,having more then 8000 comments and be evaluated 5 stars in iPhone application store.

[26] ITunes store : https://itunes.apple.com/cn/app/cai-deng-mi/id355658646?mt=8

[27] Android play : https://play.google.com/store/apps/details?id=com.kadahome.lanternriddlelite&feature=search_result#?t=W251bGwsMSwyLDEsImNvbS5rYWRhaG9tZS5sYW50ZXJucmlkZGxlbGl0ZSJd

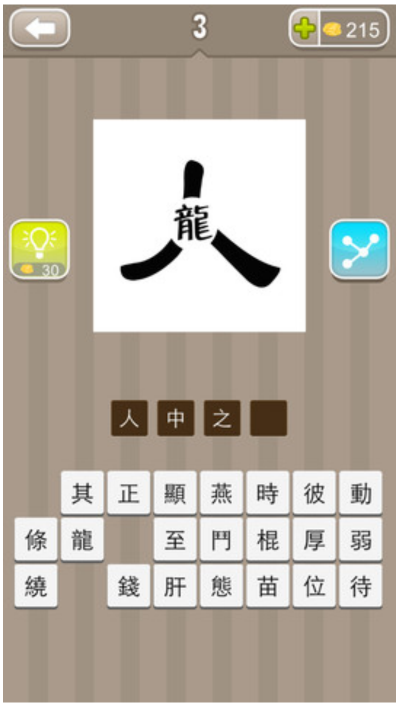


Figure 28 Guessing idiom from Picture (1)

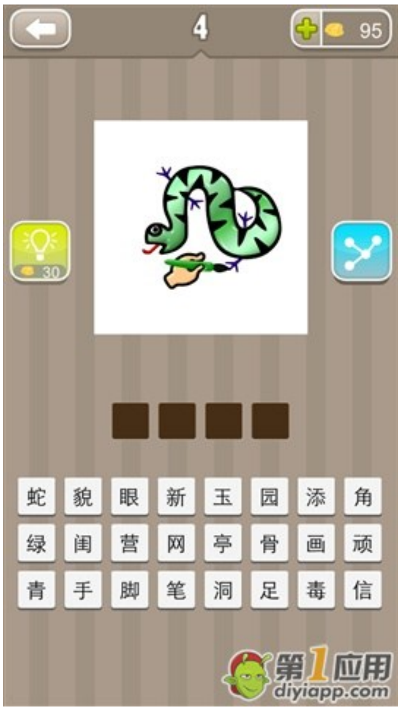


Figure 29 Guessing idiom from Picture (2)[28]

Figure 28 and Figure 29 are the screenshots of the game ‘看图猜成语’. A player has to guess the idiom from the picture showing on the screen. Instead of a keyboard, it uses a set of Chinese characters, in which contains the standard answer, as an input method in order to reduce the game difficulty. The player can seek help from his friend by publishing game status on the social network platform or buy the answer using game gold when they are stuck. In Figure 7, there are two Chinese characters, ‘人’ and ‘龍’， and the character ‘龍’ is in the middle of character ‘人’, so the answer is ‘人中之龙’, which means dragons in human in literal（means this person is outstanding）. In Figure 8, a snake is being drawn feet. In China, we call this ‘画蛇添足’, which means superfluous in English.

Another popular picture guessing game is “疯狂猜图” Crazy Guessing, which has millions fans.

[28] “看图猜成语”: https://itunes.apple.com/cn/app/kan-tu-cai-cheng-yu/id621966617?mt=8

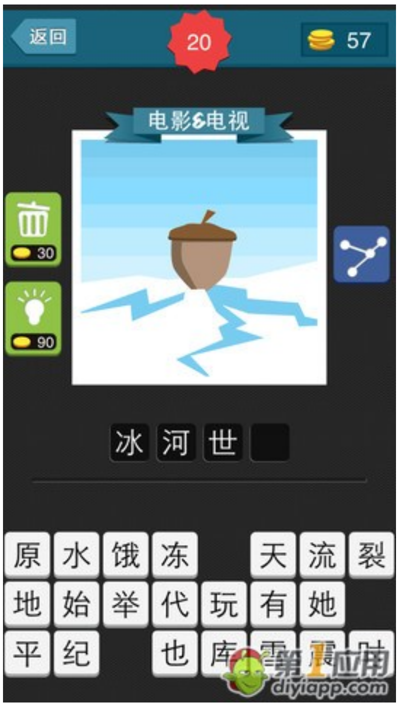


Figure 30 Guessing movies or TV from picture

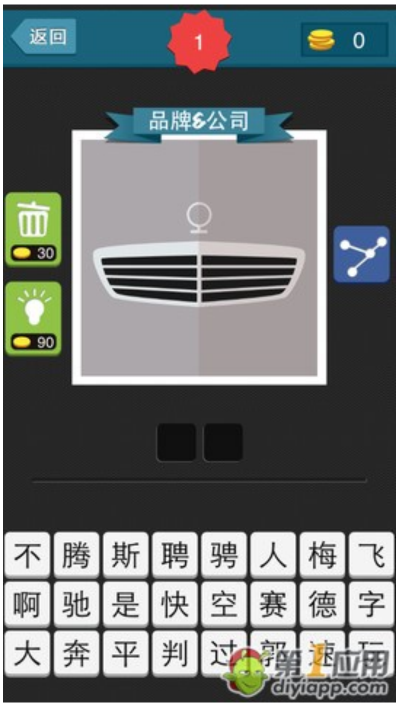


Figure 31 Guessing brand or company from picture

Figure 30 and Figure 31 shows this game is not limited in idioms. In this game, the words can be movies, brands, persons, etc. But it provides a hint of category showing to the player which kind of category it falls in. Figure 30 is the movie guessing and the answer is 'ICE AGE'. Figure 31 is brand guessing, the answer is 'Benz'. In the following are more examples.

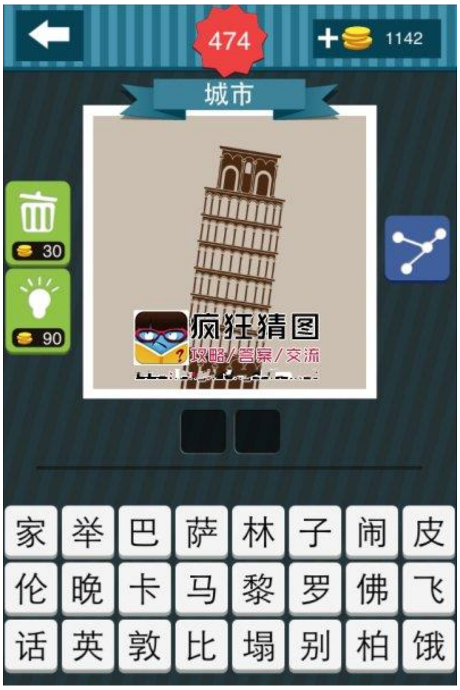


Figure 32 Guessing cities (Pisa)

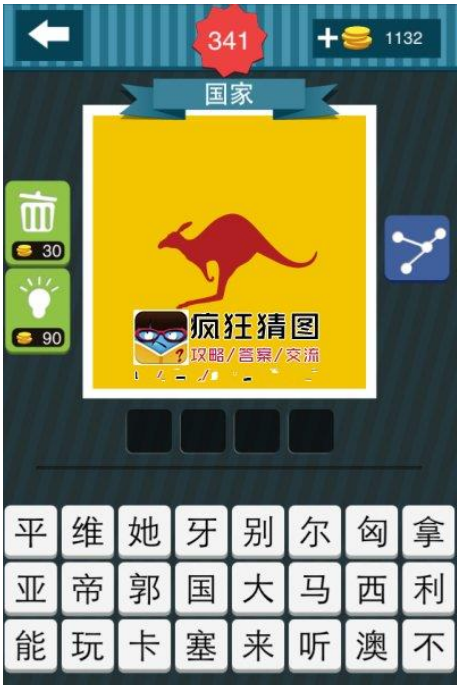


Figure 33 Guessing countries (Australia)

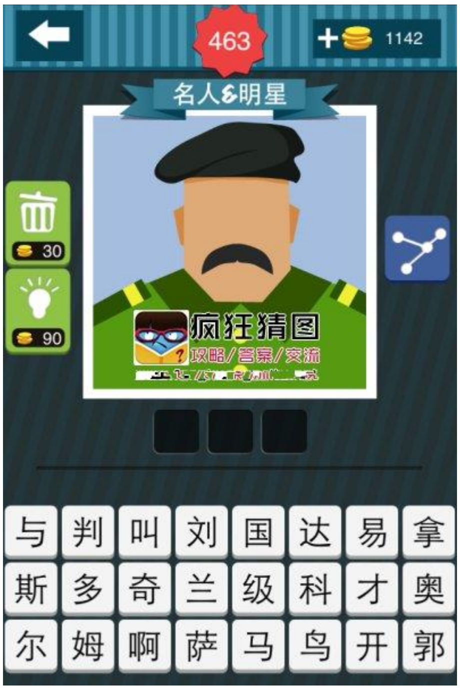


Figure 34 Guessing person (Saddam)

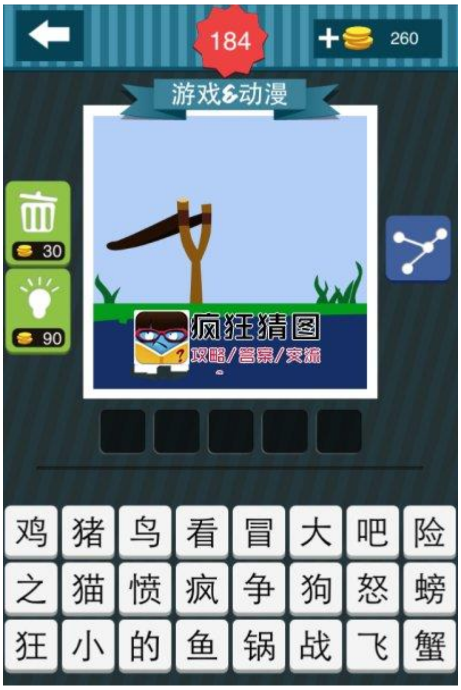


Figure 35 Guessing games (Angry bird)

Guessing picture game is similar to the lantern riddles game to a certain extent. The distinction between them is the riddles carrier, in which picture guessing game is using pictures as the riddles. The rule of picture guessing game is not complicated, which let the player describe the picture depending on specific requirements, such as, idioms, words and brands. It is also similar to the Crossword game with the image clue, but since every picture in picture guessing game is a single puzzle without character count restrictions, it has more developing space. Picture guessing game is very widespread today, especially from 2012, published many kinds of picture guessing games on the web, mobile and social networking platforms.

3.4.3 Couplet Game

A couplet or antithetical couplet[29], the Chinese name is "对联"， is a pair of poetry lines which adhere to certain rules. Unlike poems, they are usually observed on the sides of doors, used as a Chinese New Year's decoration expressing happiness and hopefulness for the coming year, thus, it is named spring couplet as well. It can be better described as a written form of counterpoint. Two poetry lines have a one-to-one correspondence in their metrical length, and each pair of characters must have certain corresponding properties. A couplet is ideally profound yet concise, using one character per word in the style of classical Chinese. A couplet must adhere to the following rules:

1. Both lines must have the exact same number of Chinese characters.
2. The lexical category of each character must be the same to its corresponding character.

[29] Wikipeida : http://en.wikipedia.org/wiki/Couplet_(Chinese_poetry)

3. The tone pattern of one line must be the inverse of the other. This generally means if one character is of the level tone, its corresponding character in the other line must be of an oblique tone, and vice versa. For more information about Chinese tone pattern, see [30].
4. The last character of the first line should be of an oblique tone, which forces the last character of the second line to be of one level tone.
5. The meaning of the two lines needs to be related, in which each pair of corresponding characters have related meanings too.

A short example of a Chinese couplet, the correspondence between individual words of the first and second poetries, is shown as follows:

海 (sea)	天 (sky)
阔(wide)	高(high)
凭(allows)	任 (enable)
鱼 (fish)	鸟 (bird)
跃 (jump)	飞 (fly)

Normally, a Chinese couplet has to be read vertically, the previous example are “海阔凭鱼跃，天高任鸟飞”.

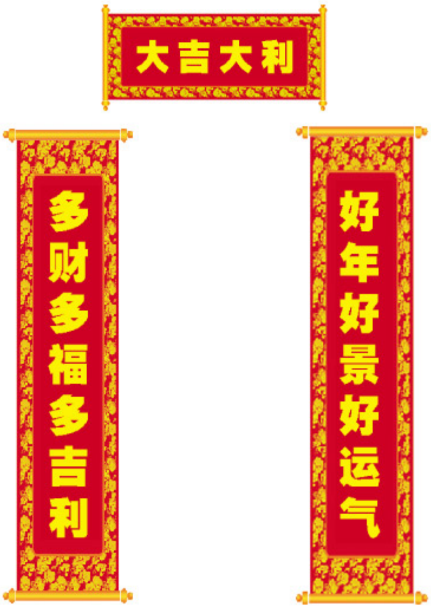


Figure 36 A spring couplet[31]

Figure 36 is an example of the spring couplet. Normally speaking, every couplet has a short title as the finishing touch, in this spring couplet, the horizontal sentence is the title.

Couplet game is legendary and it is always played on some TV shows, online forums and parties. The game is one person provides the first line poetry and let others guess the subsequent poetry. Although there is the rule of guessing the second poetry, it is still hard to judge the correctness of

[30] Chinese Tone Pattern Introduction: http://en.wikipedia.org/wiki/Tone_pattern

[31] Chinese couplet : http://fuckyeahchinesemyths.tumblr.com/post/15278728343/why-we-write-spring-couplets

the answer by a computer. And also because the couplet game is very complicated and the answer is diverse, thus, there is no web or mobile couplet game application. Actually, instead of gaming, couplets are more used to appreciation. There are lots of couplets appreciation applications on book, web and mobile platform.

3.4.4 Solitaire Game

Solitaire game indicates a kind of games that one stuff following another related stuff under a certain rule. This stuff can be a poker, word and photo, etc. And even the word solitaire game is supporting multiple languages such as English, Japanese and Chinese. In this report, we only focus on Chinese word solitaire game. It is a multi-player word game taking the following rules:

First of all, one of the participants provides a word and the others have to provide another word which the beginning is the previous word's last character. Detailed rules are different depending on interests. But in general, it should be of:

1，Without terminology.
2，Without repeating the previous words.
3，Time restriction.

Since there are no restrictions on the using words, the game becomes very easy. Thus, a derivative solitaire game named idioms solitaire game (成语接龙), which restricts words to idioms emerged. Players lose the game if they cannot provide the related idiom. The Chinese idiom, normally contains four characters, is a special language form for Chinese language, which is using a fixed phrase to express a fixed meaning or history (story). At the moment, there are more than 48000 idioms counted by People's Republic of China Ministry of Education. For example: a Chinese idiom "夸父追日"'s literal English translation is Kua Fu Chasing Sun. But, it has a background story, which is:

'Long, long ago there lived a giant man named Kua Fu[32]. He was a person with extraordinary physical power. And he could walk as fast as he flew .One day, he wanted to overcome the scorching sun and started to chase it with flying strides. When he was near the burning sun, he felt extremely thirsty. He couldn't stand it any more so he rushed to the Yellow River and drank up the river. Feeling still very thirsty, he went to Weihe River and drank up the river there too. But he was not satisfied. He decided to go to the north where there was a big lake. Unfortunately, he died on the way because of thirst. '

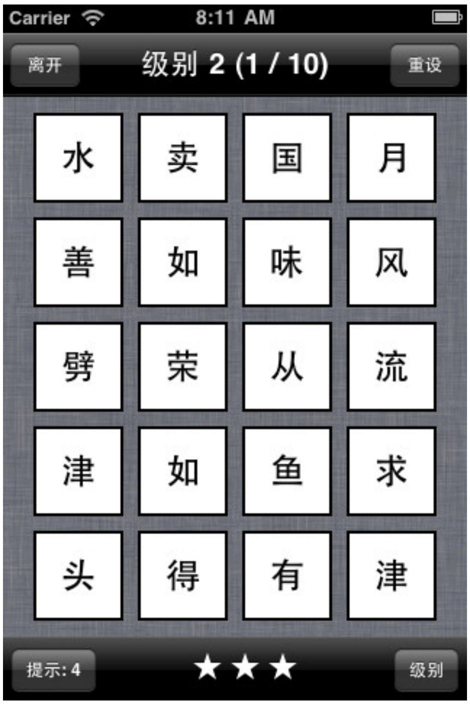
In the following is a Chinese idiom sotarire game example:
胸有成竹 → 竹柏异心 → 心安理得 → 得薄能鲜 → 鲜为人知 → 知不诈愚 → 愚不可及 → 及宾有鱼 → 鱼帛狐篝 →…

[32] iCIBA: http://sl.iciba.com/viewthread-95-525375-1.shtml

Chinese idiom solitaire game is more suitable for the advanced Chinese speakers, at least a person who has a rich accumulation on Chinese idioms. At present, Chinese idiom solitaire game is most playing on the parties, Internet forums and TV shows. But it has a probability developing on the computer and mobile platform, since a lot of Chinese idioms dictionaries exit and it is easier to take the decision of the correctness of the players' answer.

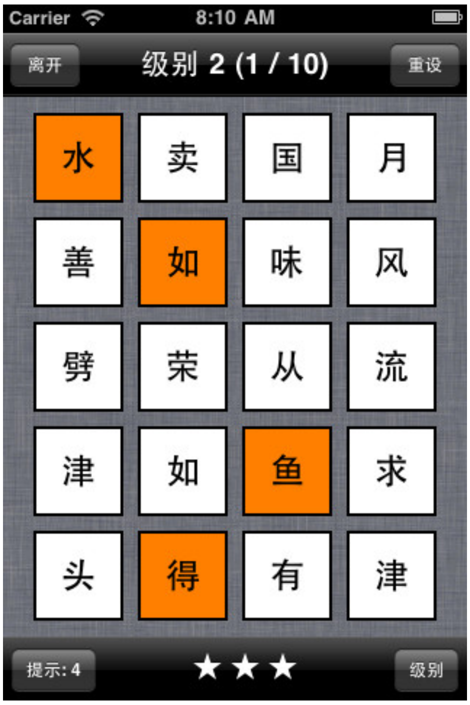
3.4.5 Idioms Game

Idioms game is a Jigsaw game that let the player find some idioms from a set of Chinese characters which are filling in the grids. Its Chinese name is "砌图".



水	卖	国	月
善	如	味	风
劈	荣	从	流
津	如	鱼	求
头	得	有	津

Figure 37 Finding idioms



水	卖	国	月
善	如	味	风
劈	荣	从	流
津	如	鱼	求
头	得	有	津

Figure 38 Finding idioms

Figure 37 and Figure 38 are a game named "砌图游戏之中国成语", English name is Puzzle Game for Chinese Idiom. In this game, a player has to figure out all Chinese idioms from a set of Chinese characters.

3.4.6 Drinking Literal Game

Chinese drinking game, like Chinese drinking, is a unique part of Chinese culture and was under a very long history. Long time ago, alcohol was mainly a beverage in the ceremonial rites. The drinking games, named "酒令"[33] in Chinese, were just aids for drinking. Certainly, there are a lot of aids for drinking, such as archery, chess playing and arrow pitching, aiming to restrict overdrinking to keep drinkers be gentlemen. There were even special designated officials to manage these aids for drinking. Later, drinking games which added entertainment to rites, gradually became artifice to persuade, wager and force overdrinking.

[33] Drinking game of Chinese Alcohol: http://www.warriortours.com/intro/alcohol_game.htm

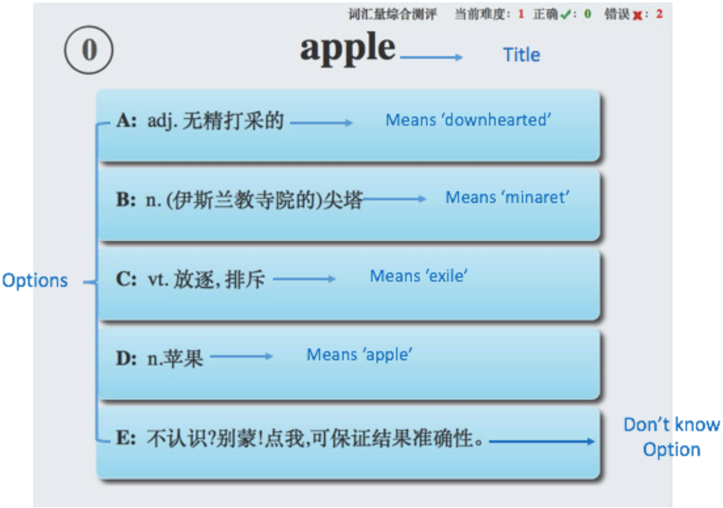
Now "酒令" has many forms depending on the drinker's social status, literacy status and interests, which can be classified into three categories - general game, contest game and literal game.

1. General game includes those games every body can play, such as joke telling, riddling and “传花” (passing flowers one by one). This category usually appears on banquet for ladies.
2. Contest game consists of archery, arrow pitching, chess playing, dicing, finger guessing and animal betting. Among these, the latter two are common practice.
3. Literal game is mainly popular in bookworms since they receive a good education and have refined knowledge and know the essence of Chinese traditional culture. Intellectuals sometimes play the other two category drinking games too, however they consider those games vulgar. Cultured ladies prefer the elegant game, literal game. Usually literal game is unique and artful literal contest, which requires superior wisdom, broad knowledge sphere and fast response. In order to animate atmosphere, players will do their best to produce original, novel, unpredicted and extremely fine literal pieces improviser, with quotations from scriptures, history, poems, proverbs, and fairy tales embedded. Many Chinese drinking games of this category, very artistic, are pleasingly worthy of literary appreciation. Bai Juyi (“白居易”), one of Chinese greatest poets, even though elegant Chinese drinking literal game was much more interesting than a musical accompaniment.

Chinese drinking literal game is very famous, but only a few people are playing. That because this game requires a very background knowledge of the Chinese literal culture. Thus, at the moment, there is not any online or mobile games existed and no one is playing this literal game during drinking. Instead of gaming, now, it plays more on appreciating.

3.4.7 Vocabulary Challenge Game

Vocabulary challenge game is a serious game to evaluate player’s vocabulary size, and always appears in the form of English-Chinese. It allows Chinese speakers to evaluate the size of their English vocabulary. The basic idea is to let a player choose the correct sense of an English word from its corresponding options. The options are 4 or 5 senses that presented in Chinese. The correct one is the Chinese translation of this word. We named one question as one round. We call a question word is a title and its candidate answers as options. We named a group of rounds that a player has to finish once as one section. Normally, one section of a vocabulary challenge game contains 10-30 rounds. A player has to finish all these rounds one by one. And a player is suggested to choose ‘don't know’ option honestly when he is no idea of the current title. After finished a section, the system will evaluate the player’s vocabulary size based on the correctness of the testing rounds.



词汇量综合测评 当前难度：1 正确✓：0 错误✗：2

0

apple

A: adj. 无精打采的

B: n. (伊斯兰教寺院的)尖塔

C: vt. 放逐, 排斥

D: n.苹果

E: 不认识?别蒙!点我,可保证结果准确性。

Figure 39 An Example of Word Challenge Game

Figure 39 is a round of a word challenge game and a section contains 10-30 such round. Word 'apple' is the title of this round. It contains 5 options, A,B,C,D and E, in which A to D are normal options and E is 'Don't know' option. The full translation of option E is 'Don't know? Don't guess! Click me, it guarantees the correctness of the result'. The correct answer of this round is D, which is a Chinese word means 'apple' in English. English translations of options A to C are 'downhearted, minaret and exile respectively. These meanings are far from the correct answer 'apple', even Part of Speech (POS) of them is different.

The platform of English vocabulary challenge game is based primarily on some English study websites through Internet browsers. Students, who are learning English, would like to evaluate their English vocabulary size. Mobile applications for such purpose are not widely developed, but it is popular on the mobile social network. People, who have strong English abilities, would like to test and share the testing result to the social network. Generally speaking, people only play it once during a period, that because of in a short term a substantial growth of the English vocabulary is difficult. But the function of errors list improves the participant times of players.

Mode	Representative Games	Brief explanation	Derived type
Question-answer	Lantern riddles guessing Picture guessing	Guess the given question, the question could be text or picture	Crossword, Vocabulary challenge, Guessing Idiom from Pictures, Crazy Guessing, etc.
No fixed answer	Couplet game Chinese Drinking literal game	Provide the continuing sentence from the given sentence by a common or specific rule. There is no correct answer but has a good answer	
Word chain	Word Solitaire game	Provide a new word, idiom or proverb by the given title's last character	Chinese ***(e.g. idiom, couplet etc.) Solitaire, etc.
Reaction	Idioms game	Find the idioms out from a given characters set as soon	

as possible

Table 4 Chinese language games

Table 4 is a summary of the Chinese language games playing with meaning. The last one ‘reaction’ is more or less playing on the human reaction time, but the familiar with idioms is more helpful for the game. The game type ‘no fixed answer’ and ‘word chain’ is lack of applications on computer platform since the hardness of deciding the answer quality by computer. The famous game types playing with meaning on App Store (IOS), Android shop (Android) and Computer platform are the ‘Question-answer’ and ‘Reaction’. After studying the derived types of these two types, we found that there is a derived type named *Vocabulary challenge* can fulfil our objective, which we can hide our goal protectively. For understanding this game well, we give an introduction in details in Chapter 5.

Chapter 4

4 UKC Games Framework

In the last Chapter, we introduced the concept of Game with a Purpose and existing Chinese game in terms of Chinese word/knowledge. After studying, we got a concrete idea of the how to create a Chinese GWAP game. Due to the data of the target resource UKC and the task of finding errors, the best game mode should be question-answer pair mode. While, to make sure the maximizing reusability of our work and also maximize the cooperation with Entitypedia Games Framework[34], a UKC games framework is created rather than a single GWAP game. UKC games framework concentrates on providing question-answer pair to all games and also some other fundamental game functions.

In this chapter, we will introduce how a question is generated and how to generate answers (options) of each question, how to measure the difficulty level of questions, domains generation, and the possible ways to cooperate with Entitypeida Games Framework, which we were working.

4.1 Questions and Assumed Answers

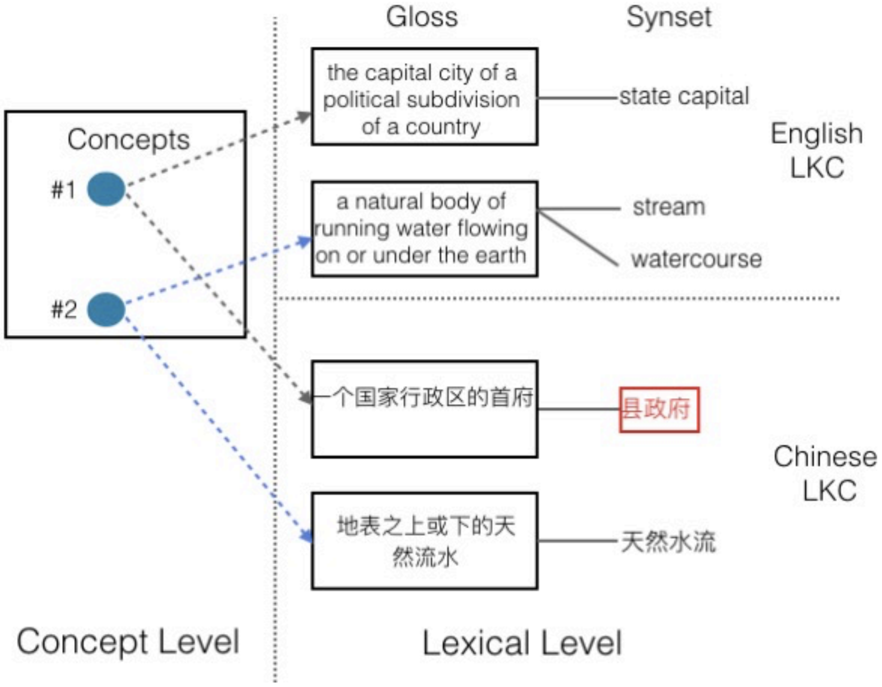


Figure 40 A Small Portion of UKC Structure

Figure 40 is an example of the basic structure of UKC. In this example, it has two concepts, 1 and 2, and each concept has two language representations. The English vocabulary representation (synset) of concept 1 is ‘state capital’, Chinese is ‘县政府’. The English synset for concept 2 is ‘stream, watercourse’, Chinese one is ‘天然水流’. An erroneous case in Chinese LKC inside UKC is concept 1. In this case, the synset ‘state capital’ in English has the correct gloss in Chinese but the synset ‘县政府’ (county government) is an incorrect translation. Furthermore,

[34] Entitypeida Games Framework is a project we were working for. The goal of it is to fix errors in Entitypedia. For more information please read ‘Entitypedia Games Framework’ tech report in Knowdive group.

the sense of the terms in both languages is different, since the English term refers to a location, whereas the Chinese counterpart refers to an organization.

In the UKC games framework, each question-answer pair is a UKC concept connected with two languages where one has to be English and the other is the target language. In this pair, English synset is the question since the English data have the highest accuracy, and target language, e.g. Chinese, is the assumed answer. Our work is to verify whether this pair is mapping correctly. In this example, concept 1 is assigned to Chinese synset '县政府'. After playing, if the most players did not select this as the correct answer, we can guess its assumed answer is wrong. The quantity of questions is limited by a question has connected with at least two languages. In UKC, there are 90,000 concept are connected with Chinese synsets and 30,000 concepts are connected with Italian. Thus, questions for English-Chinese are closed to 90,000, and for English-Italian are around 30,000.

4.2 Measuring the Difficulty Level

The ideal situation of our game is the case that a player can always get a question close to his knowledge boundary, that is, he can always know something in a provided synset. The excessively providing of inappropriate questions will cause the game boring and making the collected data doubtful. So we choose to provide the game models with both level specified/unspecified. One of the goals of the level unspecified model is to evaluate the current English vocabulary size level of a player and after that a player can select his corresponding difficulty level in the continuing play. While, the premise of providing these two game models is to find out the hardness of a question. Here, we used difficulty level as an extent that to express how hard of a question in our game. For each question, it contains two parts, title (a synset) and options. Both of them should influence the difficulty of a round. But, since 1) our target users are mainly Chinese, we assume that the options are known perfectly by the players; 2) The stable options generator, we assume that, in semantic aspect, the difficulty levels are the same for all options we generated. Thus, we only take the difficulty level of the title part (a synset) into consideration.

Yet, till now only few methods can roughly measure the difficulty of a sense (synset). Because the difficulty of sense is subjective, it is difficult to formally compare which is harder between two senses. While, in word sense disambiguation (WSD) field, as to fairly evaluate the system performance, some formal methods were proposed, which is using the Most Frequent Sense (MFS), Entropy or hybrid, to indicate the difficulty of disambiguating a particular term. That was based on two intuitive facts that, for entropy, the more information this sentence contains, the harder of this sentence to be disambiguated; for MFS, a sense might be easier if it is more commonly used, for example 'School' as 'education institution' seems easier than 'grab' as 'a mechanical device for gripping an object'. However, both of them have a limitation that requiring a considerable size labelled training data. We need to measure the difficulty of more than one hundred of thousands synsets, that is, the creation of a huge labelled training data is outweigh the benefits. Besides, the required accuracy of difficulty is lower. In our case, we just want to rank our synsets into some levels in order to provide the appropriate questions. To solve that, we need to, first, approximately rank all the synsets in UKC; second, classify these ranked synsets into levels by adopting an appropriate granularity. But, we provide an evolutionary system in order to increase the overall accuracy of the difficulty level. So, in this section, at first, we discuss how we boost our difficulty level, and after that, we introduce the evolution system.

4.2.1.1 Boosting

In current vocabulary tests, the word frequency list is used as the consensus, which was extracted from a large size corpus, as the difficulty rank. For example, 'captious' (ranks more than 6300) is a harder than 'school' (ranks around 200). Thus, we can probably use the word frequency to rank the synsets list. So as to verify our guess, we propose two options to adopt the word frequency to represent the synset value:

1) The lowest word frequency
2) The average word frequency

According to this method, we can obtain a list of all the senses ranked by their frequencies. Now, we need to consider the granularity of the difficulty level. We did some investigation and found the following information:

1. In testyourvocab.com, which is the most famous vocabulary test website, it shows that the average vocabulary size of Chinese is around 6600.
2. In Chinese examination system, the vocabulary requirement is:

Level	Junior high school	High school	CET-4[35]	CET-6	TOEFL/IELTS	GRE
Vocabulary size	2500	3500	4000	6000	7500	13000

The smallest difference in these levels is 1000 words (from junior high school to high school). So we adopt the synset value/ 1000 as one level. In this case, we use the following method to formalize the consequence.

$$Synset\ \_difficulty = \boldsymbol{Round}(\frac{F}{1000})$$

Where ***Round*** is a rounding function. ***F*** is either lowest frequency or average frequency of a synset.

Notably, in a frequency list, words like "schooling, school house, schooltime, school day" are always with high frequency rank (with low frequency), however, these words are not as hard as the corresponding frequency rank constantly. We call the words we mentioned above derived forms, and the word 'school' headwords. For the reason of balance, we use the word frequency rank of headword to represent its derived forms. Moreover, if a word can neither be found in the word frequency list, nor have a headword, then, we assume that they are hard words, and use the maximal rank as its frequency rank.

In order to compare these two methods, we randomly select some words from WordNet. We adopted the lemmatised frequency list from this link[36] in the following simulation, which con-

[35] CET is the abbreviation of College English Test band

tains 6318 lemmatised words with more than 800 occurrences. It was extracted from British National Corpus (BNC)[37], which is a large size and large spoken component.

No.	Synset	Meaning	Lowest Frequency	Average Frequency	Difficulty Level 1	Difficulty Level 2
1	school	an educational institution	181	181	0	0
2	school, schoolhouse	a building where young people receive education	181	181	0	0
3	school, schooling	the process of being formally educated at a school	181	181	0	0
4	school	a body of creative artists or writers or thinkers linked by a similar style or by similar teachers	181	181	0	0
5	school, schooltime, school day	the period of instruction in a school; the time period when school is in session	181	181	0	0
6	school	an educational institution's faculty and students	181	181	0	0
7	school, shoal	a large group of fish	181	3249	0	3
8	grab	a mechanical device for gripping an object	6318	6318	6	6
9	catch, grab, snatch, snap	the act of catching an object with the hands	2818	4885.5	3	5
10	pasture, pastureland, grazing land, lea, ley	a field covered with grass or herbage and suitable for grazing by livestock	6186	6265.2	6	6
11	eatage, forage, pasture, pasturage, grass	bulky food like grass or hay for browsing or grazing horses or cattle	2145	5241.75	2	5
12	measure, evaluate, valuate, assess, appraise, value	evaluate or estimate the nature, quality, ability, extent, or significance of	403	2964	0	3
13	assess	charge (a person or a property)	1570	1570	2	2
14	tax, assess	set or determine the amount of (a payment such as a fine)	1570	2874.5	2	3
15	assess	estimate the value of (property) for taxation	1570	1570	2	2

Table 5 Simulation table, difficulty level 1 is corresponding to lowest frequency, difficulty level 2 is corresponding to average frequency

After that, we think both of these two kinds of synset values can be utilized to distinguish the synset difficulty level. For example, No. 1, 2, 3, 4, 5, 6 are quite easy, they belong to level 0; No.10, 11 are not frequently used, belongs to level 6; While No. 7, 9,11,12, 14 are showing different. After analyzing the differences, in table [3], we found that a player can guess the meaning from the easiest word sometimes, but only works well for the top meanings of this word, that is, the lowest frequency method underestimates the synset difficulty level on the low word sense rank. While, average frequency method has a better preference on this aspect. In this case, we adopt the average frequency as our solution.

[36] http://www.kilgarriff.co.uk/bnc-readme.html

[37] http://www.natcorp.ox.ac.uk

No.	Synsets	Analyses
7	***school, shoal***	school is an easy word, but this is its last meaning (7/7).
9	***catch, grab, snatch, snap***	Catch is an easy word, but that sense is its eighth meaning (8/10), which is not as easy as word catch's ranking.
11	***eatage, forage, pasture, pasturage, grass***	word grass is easy, but this is its fourth meaning (4/5).
12	***measure, evaluate, valuate, assess, appraise, value***	value is an easy word, but it is the fourth meaning (4/5).
14	***tax, assess***	tax as a verb is not frequently used. The lowest frequency is the word assess. This is its third meaning (3/4). While, if a player knows the word tax, it is not hard to get this meaning.

Table 6

As we mentioned above, the average known words are around 6500, and for GRE, which is generally considered as the highest-level English exam in China, the requirement of words is around 13000 words. But, in UKC, we have more than 150000 words, which means the major part of synsets should be marked as the hardest level. Besides, we use 1000 as the granularity to classify our synset rank, but 1000 words do not map to 1000 meanings. In this scenario, we calculate the following difficulty level distribution as showed in Figure 41. The sum of difficulty level 0 to difficulty level 5 is 19180, and the major part difficulty level 6 is 91550. In table 4, we demonstrate some random words for each level.

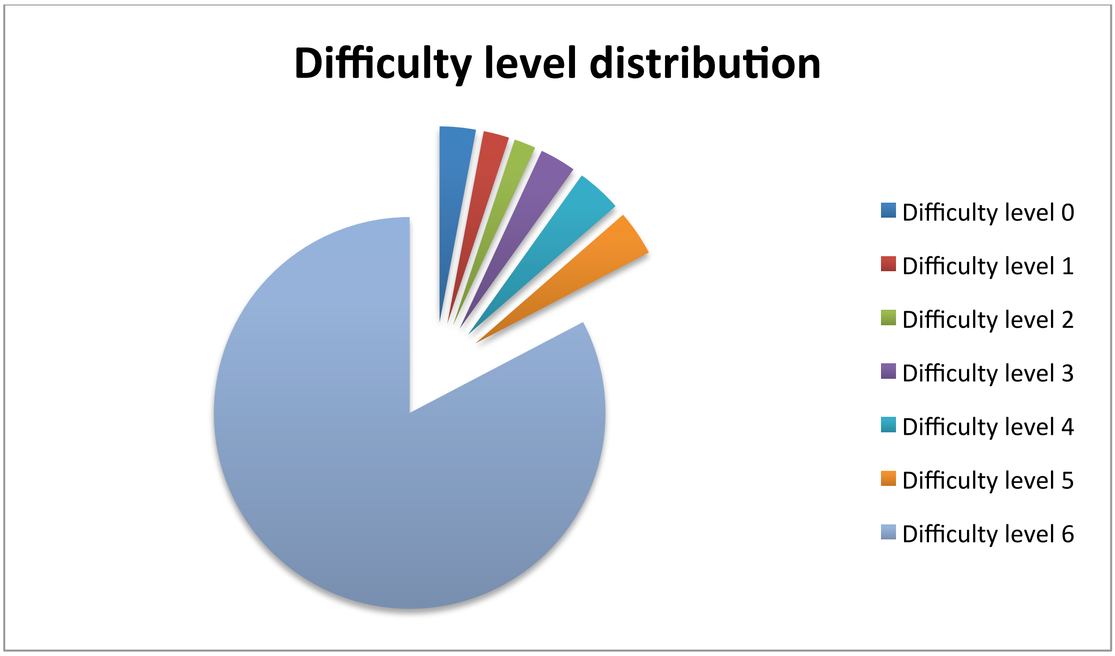


Figure 41 Distribution of difficulty level

Difficulty level	Some Synsets	Difficultly level size
Difficulty level 0	*{name}, {cause}, {similar}, {anything}, {something}*	*3301*

Difficulty level 1	*{article}, {credit}, {contact}, {address}, {device}*	*2351*
Difficulty level 2	*{queen}, {barrier}, {architecture}, {base, base of operations}, { part, separate, divide, disunite }*	*1979*
Difficulty level 3	*{description, verbal description }, {food, nutrient }, {credit, course credit }, {primary, primary election}*	*3307*
Difficulty level 4	*{example, illustration, instance, representative}, { object, physical object, shape bearing object}*	*4116*
Difficulty level 5	*{part, portion, component, component part}, { transmission, transmittal, transmitting }, { alteration, modification, adjustment}*	*4126*
Difficulty level 6	*{boondoggle}, {bootlegging}, {absolution, remission, remittal remission of sin}, {expiation, atonement, propitiation}*	*91550*

Table 7 some random examples for each level

4.2.1.2 The Evolution System

In the pervious section, we introduced how we boost the difficulty level. We separated the whole UKC into 7 levels, and each level has a different quantity of synsets, in which the difficulty level 7 contains the most synsets, that is 91550. In this section, we will introduce how we increase the overall accuracy of the boosted difficulty level system.

The difficulty level of a synset is briefly based on the people thinking that how hard is it. Thus, we can get an assumption that 'If the most players think a synset is harder/easier than the other synsets in the current difficulty level, then this synset should be located in a higher/lower difficulty level instead of the current one'. So, if we can get opinions in terms of difficulty of synsets, we can relocate these synsets and improve the quality of our difficulty level. The best way to get opinions is to ask players directly, but 1) asking an opinion for each question is tedious; 2) against our original goal that hide our objective. So, instead of asking an opinion from player, we propose to use an indirect approach, by evaluating the mistake rate and skip rate, to get opinions.

The accuracy of a question is referring to a percentage of how many players answered correctly. There are two possibilities that if a question is taking a higher mistake rate by comparing with the other questions in the same difficulty level, which is the question is harder or the mapping Chinese synset is wrong. And since we suppose that the game option set is in the same difficulty level, a question is harder can be regarded as the synset is harder. And when the case is in the wrong mapping, move it to next level can give it an improved validation. In this case, we can move this synset into the next difficulty level in both cases. On the other hand, if the mistake rate of a synset is lower than the current mistake rate, we think it should be moved to the previous difficulty level.

However, there are two situations we need to consider. First, a player has the capacity to evaluate the difficulty level. For example, when a player answers a game with unreasonable actions, the accuracy is around 0%-20%. So a player has only 0-20% accuracy for the current difficulty level, his answers are not able to use in the evolution system. Also when a player A of level 3 is playing with level 6 or a level 5 player B is playing with level 2, their answers are meaningless for shaping the difficulty level system. The ideal situation is that, the concept size of a player, who provided the mistake rate data of synset S, has to be around the difficulty level of S, not too high or too low. Suppose that a mistake rate of a synset used to calculate the difficulty level shifting. By analysing the data collected till now, we found that the accuracy range between 65% to 80% can indicate a player is in this difficulty level. In this case, a player's answers upper than 80% or lower than 65% will not be counted in the mistake rate R.

Second, we need to consider when to shift the difficulty level. Suppose that the accuracy for moving up the difficulty level is A+, for moving down is A-. In the case of R-, the counted players should be in or lower than current difficulty level. In this scenario, his answer is meaningless to indicate a synset is easier. Second, we need to figure out the threshold K for deciding R of a synset. That is, how many answers is enough to decide a mistake rate.

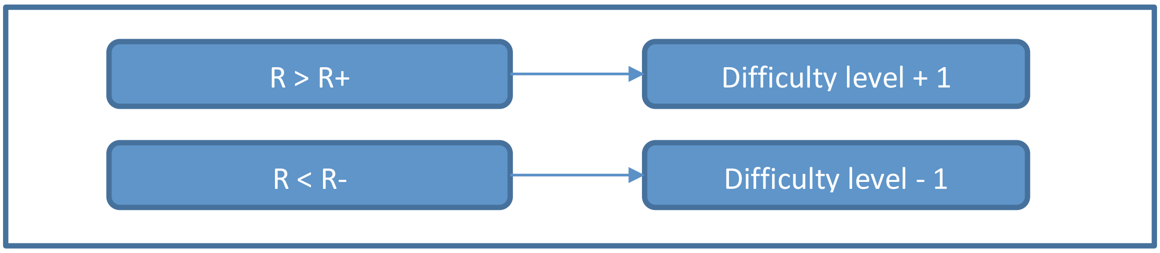


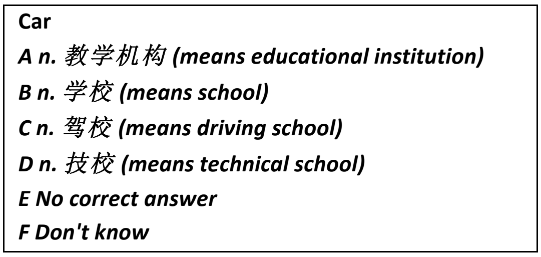
Figure 42 Evolution system

As showed in Figure 42, when mistake rate R of a synset is higher than R+, we move this synset to the next difficulty level. And when R of a synset is lower than R-, we move this synset to its pervious difficulty level. It has a probability that a synset has to be moved into several levels away, but for each time, we only move 1 level. And for each moving, we put it in the middle of that level.

4.3 Option Set Generating

A set of options consists of 5 options in game framework database, and they are Chinese synsets respectively. In this document, we use ***assumed answer*** to indicate the right answer we designed (from UKC), ***real answer*** to indicate the answer from the gold standard and ***player answer*** means the answer from players. So, assumed answer might be able to correct or wrong by comparing with the gold standard. When a player answer is different with an assumed answer, it might be a player's mistake or assumed answer is wrong. The assumed answer is utilized to judge the correctness of player answers.

In a Vocabulary Challenge Game, an option set is always designed as 3 or 4 Chinese options plus an additional 'Don't know' option. While, we prepare 5 options for each option set. That because, in our case, except the above options we discussed, we designed a new option named '***No correct answer'*** in order to solve the problem that the assumed answer is not correct in reality. For example, if we have a wrong record like 'Car'-'学校'(means school in English), a round will be generated as following. The assumed answer is 'B'.



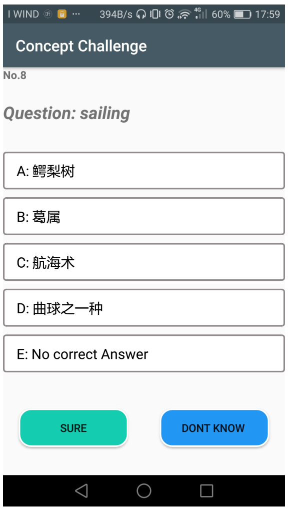
Car
A n. 教学机构 (means educational institution)
B n. 学校 (means school)
C n. 驾校 (means driving school)
D n. 技校 (means technical school)
E No correct answer
F Don't know

Figure 43 an example

In this situation, without the ‘no correct answer’ option, a player is impossible to select a correct answer even he knows the real answer and this question will be a bug. The option *‘No correct answer’* not only solved this problem, but also makes some advantages. In the actual situation, players answer some questions by guessing rather than choosing *‘don't know*’ honestly, because of the correct answer exists in the options definitely. While, as a game trap, if some rounds are designed as with no correct answer intentionally. We decrease the possibility of guessing from the game potentially. In order to make this additional option more useful and not obtrusive, in a real game, we randomly select ‘No correct answer’ as the assumed answer for very few instances as the game trap. While, when we use ‘No correct answer’ as the correct option, we need an additional option to replace the correct answer. That is reason we generated one more option in the option set.

For the normal case, the option set for a certain English synset will be sustained for a long period. For example, synset ‘school’, all players are playing it with the same option sets. Because, the errors are figured out based on the fact that the maximal answers are different with knowledge base record, in the other words, it is a voting system. For a voting system, a vote has to be based on the same target. Options are influencing the game difficulty level, a synset with two different types of options are two different questions actually. To let that voting make sense, the option set for a question will keep stable until we got enough votes.

Besides, the current options generation methods of vocabulary test are primarily based on random option set and related option set. The random option set indicates the options are selected randomly. So, it is obtrusive sometimes and a player can guess the correct answer even for a non-familiar title. In this case, the random option set is decreasing the difficulty level of a round and at the same time narrowing down the reliability of the answers. As showed in Figure 44, Option A ‘鳄梨树’ in English is ‘Avocado’, which means ‘tropical American tree bearing large pulpy green fruits’; Option B ‘葛属’ in English is ‘Pueraria’, means ‘genus of woody Asiatic vines: kudzu’; and Option D in English is ‘Curve, Curve ball’ , means ‘a pitch of a baseball that is thrown with spin so that its path curves as it approaches the batter’; Supposed that a player has no idea of the concept ‘sailing’, but he knew the meaning of ‘sail’. In this case he can guess the correct answer as C, which is the only option related to the concept ‘sail’.



No.8

Question: sailing

A: 鳄梨树

B: 葛属

C: 航海术

D: 曲球之一种

E: No correct Answer

Figure 44 An example of Options

Related option set is on the opposite side, which are providing the related options, which means all the options are similar. It has the advantage that increased the accuracy of testing a player whether knows the concept, but in the other side, the similar options increase the difficulty of selecting the correct answers, the chance of selecting the correct answer is reducing. So we are under the risk that decreases the significance level of the majority answers. Furthermore, how to manage the relation between the question difficulty level and option difficulty level is not evident. For example, a hard question with the related option set might cause a game too hard to play. Thus, in order to identify which is the best option association for our purpose, we choose four kinds of option sets by the degree of the relatedness, which are related option set, semi-related option set, domain related option set and random option set.

- **Related Option set**

To generate a related option set, for a source English synset, at first we find its corresponding Chinese synset C and then, we randomly select 4 concepts from C's direct children, siblings and parent as the related options. As showed in Figure 45 ,the area marked with green. It is notable that we first select from C's direct parent and children, which are the most related concepts of C. If the sum of its parent and children is less than 4, we randomly select its siblings until we get 4 options. In some extreme cases, the sum of C's direct parent, children and siblings are less than 4, we do not generate this English synset's related option set. In real case, some of English Synset are translated into the same Chinese or contained the same Chinese word, in this case, we filter them or treat two similar Chinese synsets as the same answer. For example, a small part of direct Hyponym & Hypernym of Synset ***'{school}; an educational institution;'*** is showed in the following.

Hypernym: {Educational institution}—{教学机构}
Siblings: {college}—{学院}
{university}—{大学}
Hyponym: {Academy}—{学院；研究院}
{Driving school}—{驾校}
{technical school, tech} –{技校}

School

A n. 教学机构 (means educational institution)
B n. 学校 (means school)
C n. 驾校 (means driving school)
D n. 技校 (means technical school)
E n. no correct answer.

- **Semi-related Option set**

The generation process of Semi-related option set is similar with related option set, but since the relatedness degree is lower, we take one more step away comparing to the related option set. So the semi-related options are selected from 2 steps from C' ancestors and children. We do not generate the semi-related when we cannot get adequate options for an option set.

- **Domain-related Options**

Instead of searching 3 steps away from C, we do Domain- related option set, which is selected randomly from the domain of concept C. We did some experiments, the relatedness degree of options selected from 3 steps away is the same with random options. When the mapped concept C has a domain, we randomly select 4 options from its domain as its domain related option set.

- **Random Options**

Randomly selected from the concepts from the whole database that has the same POS (part of speech).

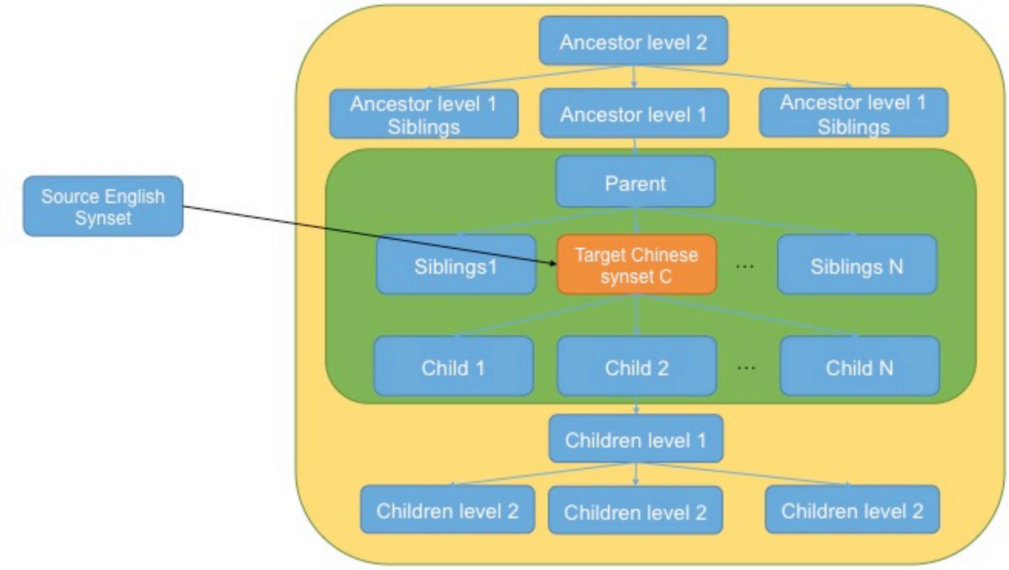


Figure 45 Option Sets Selection

We randomly select 604 English synsets that having the corresponding mapped Chinese synsets, for each difficulty level, we got the following result for each kind of option sets as in Table 8. Table 9 shows the distribution of Italian option sets.

	Related option	Semi-related option	Domain-related option	Random option
Difficulty level 0	136	143	127	147

Difficulty level 1	74	80	65	83
Difficulty level 2	56	59	52	60
Difficulty level 3	50	51	47	52
Difficulty level 4	34	34	33	35
Difficulty level 5	34	39	40	40
Difficulty level 6	158	185	175	188
Sum	542	591	539	604

Table 8 Quantity of Option Sets for 604 Chinese Questions

	Related option	Semi-related option	Domain-related option	Random option
Difficulty level 0	46	50	70	72
Difficulty level 1	37	47	54	55
Difficulty level 2	34	44	53	54
Difficulty level 3	59	71	80	84
Difficulty level 4	64	78	104	106
Difficulty level 5	60	76	101	104
Difficulty level 6	431	511	737	758
Sum	731	877	1199	1233

Table 9 Quantity of Option Sets for 1233 Italian Questions

4.4 Domains

In UKC games framework, we also provide questions classified by domains. Domain information was extracted from WordNet Domains developed by FBK. WordNet Domains (WND) [41] is a lexical resource built by augmenting WordNet with domain labels in a semi-automatic way. Each synset in WordNet has been annotated with at least one semantic domain label. There are 164 labels structured according to the WordNet Domain Hierarchy. In this hierarchy, the first level is doctrines, free_time, applied_science, pure_science, social_science, factotum and each of them are separated into several sub domains where the maximum depth is four. For example, as shown in Figure 46, doctrines domain is composed by Psychology, Art, Philosophy and Religion. The entire domain can be found in link[38] and additional information of how to create WND can be found in paper [42].

[38] http://wndomains.fbk.eu/hierarchy.html

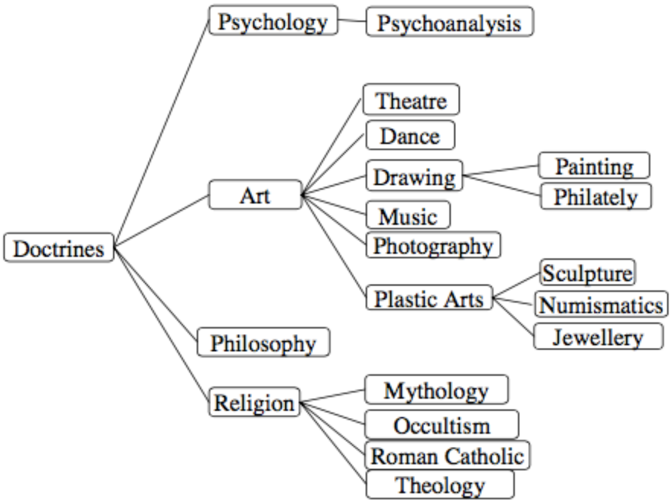


Figure 46 Fragment of WDH

After imported WND to UKC, we count the mapping concepts for each domain as in Appendix. Some domains are very small. For example, domain rugby only has 6 concepts and Cricket has only 24 concepts, etc. Moreover, in addition to random option set are not available always, for example, in Table 8 and Table 9, related option set for 604 Chinese questions are 542, and for 1233 Italian questions are 731, we removed some domains when its amount is too little. As the generated 604 Chinese questions, the distribution of domain is showed in Table 10. If any kind of option set in a domain is less than 20, the games framework deletes the domain automatically. In addition, domain 'factotum' is removed as well since it is not a specific domain.

	Related option	Semi-related option	Domain-related option	Random option
Sport	46	47	50	50
Sociology	22	28	28	28
Play	37	43	44	44
Dance	31	34	34	34
Sum	136	152	156	156

Table 10 Quantity of domains in 604 Chinese Questions

	Related option	Semi-related option	Domain-related option	Random option
Gastronomy	225	195	376	379
Food	58	82	82	92
Factotum	120	141	183	183
Psychological	106	206	213	214
Sum	509	624	854	868

Table 11 Quantity of domains in 1233 Italian Questions

4.5 Login Design

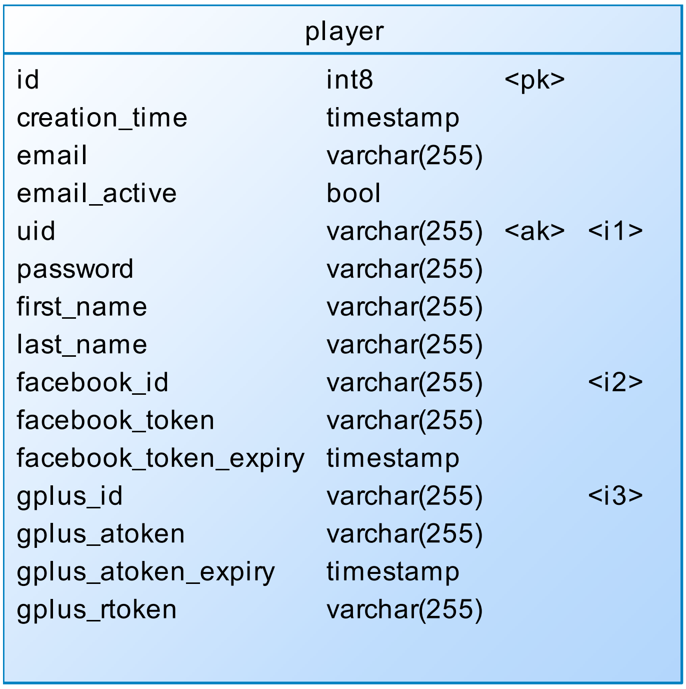
In UKC games framework we provide two kinds of login activity, simplified login and complete login. For some casual word games, we do not require a complicate login system, but need a simple identity to recognize the game player. By this scenario, simple login system is offered. We also provide a complete login system via calling APIs from Entitypedia games framework.

Simplified login requires a unique ID and a player name to login. This unique ID can be Android Device ID, or IPhone Device ID. For example, Android device ID can be call by:

getContext().getContentResolver(),Secure.ANDROID_ID

A player needs to input his player name when the first time login our system.

Complete login is developed via calling APIs from Entitypedia Games framework. Entitypedia Games Framework keeps a single user account database which needs to be used by all games to enable smooth experience across all games of the framework. Figure 47 shows this minimal set of attributes of a player.



player			
id	int8	<pk>	
creation_time	timestamp		
email	varchar(255)		
email_active	bool		
uid	varchar(255)	<ak>	<i1>
password	varchar(255)		
first_name	varchar(255)		
last_name	varchar(255)		
facebook_id	varchar(255)		<i2>
facebook_token	varchar(255)		
facebook_token_expiry	timestamp		
gplus_id	varchar(255)		<i3>
gplus_atoken	varchar(255)		
gplus_atoken_expiry	timestamp		
gplus_rtoken	varchar(255)		

Figure 47 A player table

Chapter 5

5 State of the Art of Vocabulary Challenge Games

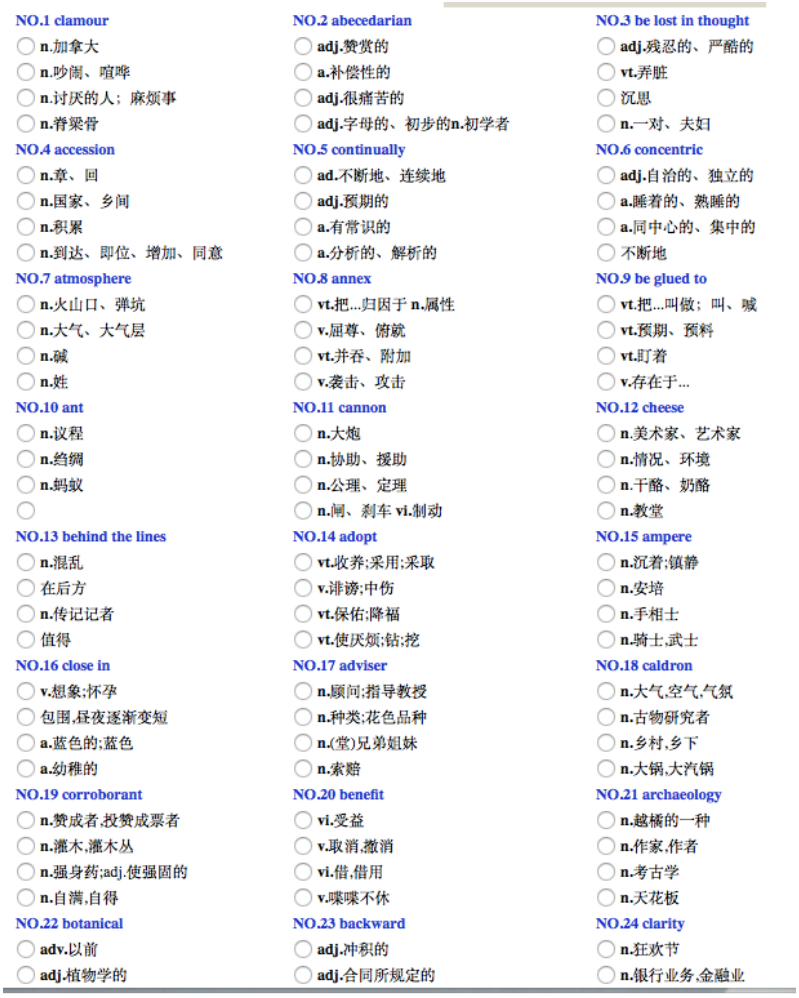
As we discussed in Chapter 3, a vocabulary challenge game will be our solution. In this situation, a comprehensive understanding of the vocabulary challenge game is necessary. Thus, in this chapter, we will introduce the game features of existing vocabulary challenge games. To do that, we investigated most of existing applications of English-Chinese vocabulary challenge game. In order to get a representative survey, we use both Google, which is focusing on web applications, and smart phone stores (apple store and android store), which are focusing on smartphone applications, to search the existing games. The structure of this chapter is: in section 5.1, we provide a classification based on game format; in section 5.2, we use an example named '易记单词' to illustrate all game features of a vocabulary challenge game.

5.1 Vocabulary Challenge Game Classifications

After investigation, we found that there are several kinds of word challenge games. According to game format, that is, how the game looks like, three kinds of game forms are classified as follows.

- **Questionnaire** All the rounds are provided in one questionnaire. That is, all rounds are visible at the same time. After filling all the rounds, system will provide the finial consequence. As showed in Figure 48.
- **Single round** The quantity of a section is the same, but different with questionnaire, only one round is visible each time, as showed in Figure 49. The example of Figure 39 is single round type game either.
- **Selection based** Before challenging, the player has to pick the words he/she knows from the given word list. After that, all the selected words will be test one by one. As showed Figure 50 and Figure 51 respectively. A player has to select what he knows as in Figure 50 and the selected words will be tested in Figure 51. While, there is an exceptional case, a type that only has the selecting part, but without the testing part. That is based on the assumption that a user honestly selects what he knows without cheating and guessing. But for this type, it lacks the element of game and challenge.

However, in the scenario of Web application, cell phone application, etc., the questionnaire format and selection based format are unusual. Most applications are of single round format.



NO.1 clamour
○ n.加拿大
○ n.吵闹、喧哗
○ n.讨厌的人；麻烦事
○ n.脊梁骨

NO.2 abecedarian
○ adj.赞赏的
○ a.补偿性的
○ adj.很痛苦的
○ adj.字母的、初步的n.初学者

NO.3 be lost in thought
○ adj.残忍的、严酷的
○ vt.弄脏
○ 沉思
○ n.一对、夫妇

NO.4 accession
○ n.章、回
○ n.国家、乡间
○ n.积累
○ n.到达、即位、增加、同意

NO.5 continually
○ ad.不断地、连续地
○ adj.预期的
○ a.有常识的
○ a.分析的、解析的

NO.6 concentric
○ adj.自治的、独立的
○ a.睡着的、熟睡的
○ a.同中心的、集中的
○ 不断地

NO.7 atmosphere
○ n.火山口、弹坑
○ n.大气、大气层
○ n.碱
○ n.姓

NO.8 annex
○ vt.把...归因于 n.属性
○ v.屈尊、俯就
○ vt.并吞、附加
○ v.袭击、攻击

NO.9 be glued to
○ vt.把...叫做；叫、喊
○ vt.预期、预料
○ vt.盯着
○ v.存在于...

NO.10 ant
○ n.议程
○ n.绉绸
○ n.蚂蚁
○

NO.11 cannon
○ n.大炮
○ n.协助、援助
○ n.公理、定理
○ n.闸、刹车 vi.制动

NO.12 cheese
○ n.美术家、艺术家
○ n.情况、环境
○ n.干酪、奶酪
○ n.教堂

NO.13 behind the lines
○ n.混乱
○ 在后方
○ n.传记记者
○ 值得

NO.14 adopt
○ vt.收养;采用;采取
○ v.诽谤;中伤
○ vt.保佑;降福
○ vt.使厌烦;钻;挖

NO.15 ampere
○ n.沉着;镇静
○ n.安培
○ n.手相士
○ n.骑士,武士

NO.16 close in
○ v.想象;怀孕
○ 包围,昼夜逐渐变短
○ a.蓝色的;蓝色
○ a.幼稚的

NO.17 adviser
○ n.顾问;指导教授
○ n.种类;花色品种
○ n.(堂)兄弟姐妹
○ n.索赔

NO.18 caldron
○ n.大气,空气,气氛
○ n.古物研究者
○ n.乡村,乡下
○ n.大锅,大汽锅

NO.19 corroborant
○ n.赞成者,投赞成票者
○ n.灌木,灌木丛
○ n.强身药;adj.使强固的
○ n.自满,自得

NO.20 benefit
○ vi.受益
○ v.取消,撤消
○ vi.借,借用
○ v.喋喋不休

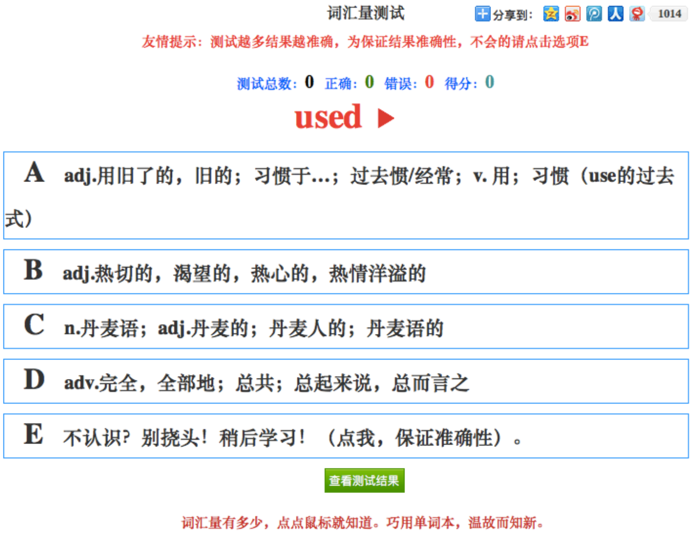
NO.21 archaeology
○ n.越橘的一种
○ n.作家,作者
○ n.考古学
○ n.天花板

NO.22 botanical
○ adv.以前
○ adj.植物学的

NO.23 backward
○ adj.冲积的
○ adj.合同所规定的

NO.24 clarity
○ n.狂欢节
○ n.银行业务,金融业

Figure 48 Questionnaire type



词汇量测试

分享到： 1014

友情提示：测试越多结果越准确，为保证结果准确性，不会的请点击选项E

测试总数：0 正确：0 错误：0 得分：0

used ▶

A adj.用旧了的，旧的；习惯于...；过去惯/经常；v. 用；习惯（use的过去式）

B adj.热切的，渴望的，热心的，热情洋溢的

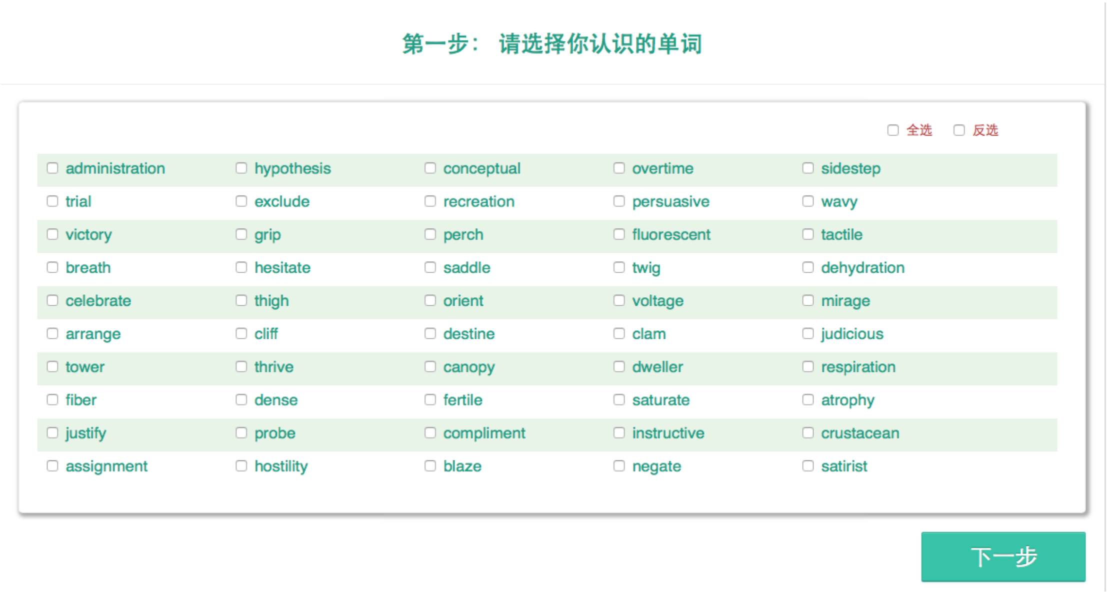
C n.丹麦语；adj.丹麦的；丹麦人的；丹麦语的

D adv.完全，全部地；总共；总起来说，总而言之

E 不认识？别挠头！稍后学习！（点我，保证准确性）。

词汇量有多少，点点鼠标就知道。巧用单词本，温故而知新。

Figure 49 Single round type

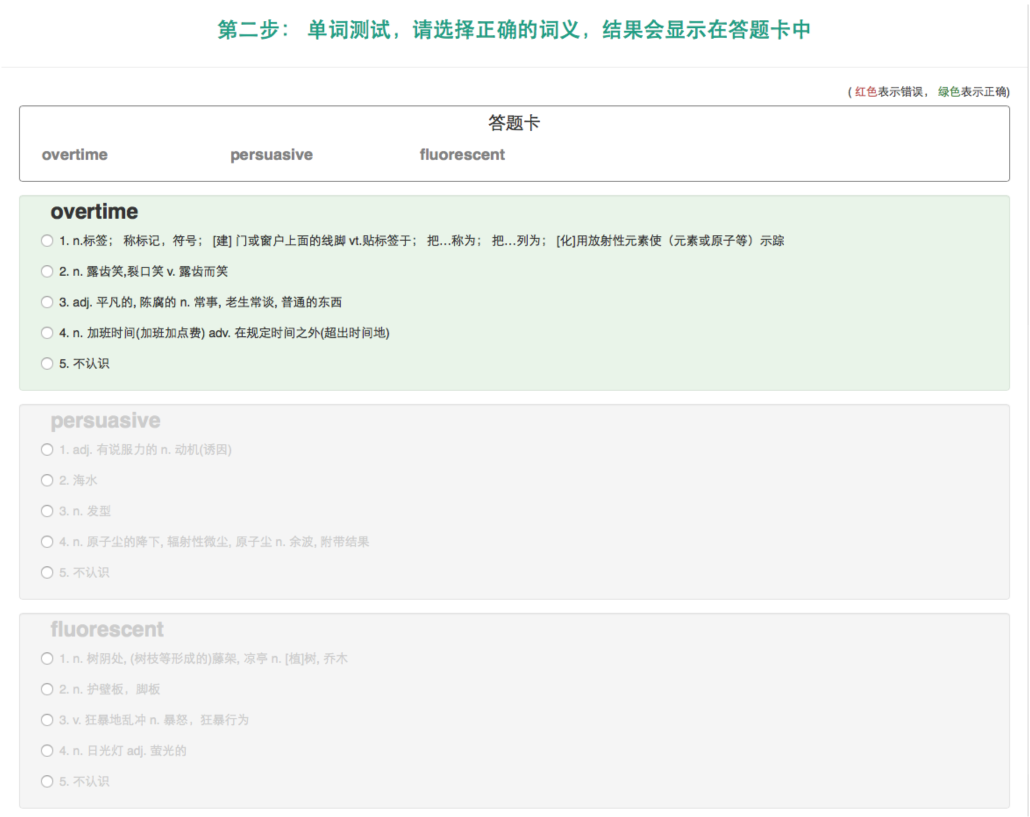


第一步： 请选择你认识的单词

□ 全选 □ 反选

□ administration	□ hypothesis	□ conceptual	□ overtime	□ sidestep
□ trial	□ exclude	□ recreation	□ persuasive	□ wavy
□ victory	□ grip	□ perch	□ fluorescent	□ tactile
□ breath	□ hesitate	□ saddle	□ twig	□ dehydration
□ celebrate	□ thigh	□ orient	□ voltage	□ mirage
□ arrange	□ cliff	□ destine	□ clam	□ judicious
□ tower	□ thrive	□ canopy	□ dweller	□ respiration
□ fiber	□ dense	□ fertile	□ saturate	□ atrophy
□ justify	□ probe	□ compliment	□ instructive	□ crustacean
□ assignment	□ hostility	□ blaze	□ negate	□ satirist

Figure 50 Selection based type 1



第二步： 单词测试，请选择正确的词义，结果会显示在答题卡中

（红色表示错误， 绿色表示正确）

答题卡

overtime persuasive fluorescent

overtime

○ 1. n.标签； 称标记，符号； [建] 门或窗户上面的线脚 vt.贴标签于； 把…称为； 把…列为； [化]用放射性元素使（元素或原子等）示踪

○ 2. n. 露齿笑,裂口笑 v. 露齿而笑

○ 3. adj. 平凡的, 陈腐的 n. 常事, 老生常谈, 普通的东西

○ 4. n. 加班时间(加班加点费) adv. 在规定时间之外(超出时间地)

○ 5. 不认识

persuasive

○ 1. adj. 有说服力的 n. 动机(诱因)

○ 2. 海水

○ 3. n. 发型

○ 4. n. 原子尘的降下, 辐射性微尘, 原子尘 n. 余波, 附带结果

○ 5. 不认识

fluorescent

○ 1. n. 树阴处, (树枝等形成的)藤架, 凉亭 n. [植]树, 乔木

○ 2. n. 护壁板，脚板

○ 3. v. 狂暴地乱冲 n. 暴怒，狂暴行为

○ 4. n. 日光灯 adj. 萤光的

○ 5. 不认识

Figure 51 Selection based type 2

5.2 Game Features

In this section, we will discuss all the game features of a vocabulary challenge game. But because there are a plenty of game features, and a game feature might be different for different games. For example, game A has time limitation for each round, but game B has no time limitation. So, for a better understanding, we chose a game named ‘易记单词’, which is a well known web application and close to our requirements, as our mainline. And at the same time provide the different manifestations from the other applications. So this section is structured as: before talking about each game feature in details, we analysed all the possible game features of ‘易记单词’, and after that we discuss each of them in details.

Figure 52 is the game process of ‘易记单词’ . It has 3 steps, which are welcome step, main game step and result step. The first step is the welcome step, showing how to play the game and some game options. After that is the main game step is similar as what we discussed in the first section of this chapter. And after a player played dozens of rounds, the result step is coming, containing game result and some explains to the result.

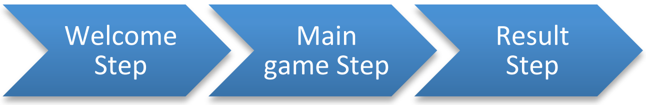
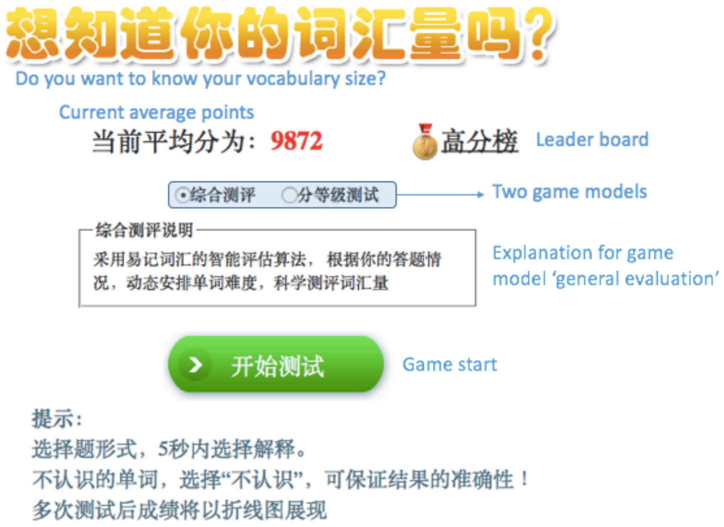


Figure 52 Game process of ‘易记单词’

Figure 53 is the welcome step of ‘易记单词’, including several game features. At the top is the game title, it means ‘Do you know your vocabulary size?’. After that, is average point and leader board, in which leader board is clickable. While, this average point does not point out what average points it is. For example, it might be an average points for a player recently played, or an average score for all players. In our opinion, it is an average point of all players. Since after we played several times, this number did not change. Below that is the game modes. In this game, it has two game modes, ‘General evaluation’ and ‘evaluation based on levels’. The default game model is ‘General evaluation’. While, if a player chooses ‘evaluation based on levels’, the game view is changed to Figure 54. Comparing to Figure 53, the average point becomes a percentage and the content of the explanation box turn into two selectable lists. Below the explanation box is a game start button.


想知道你的词汇量吗?
Do you want to know your vocabulary size?
Current average points
Leader board
Two game models
Explanation for game model ‘general evaluation’
Game start
开始测试

当前平均分为：9872 高分榜

● 综合测评 ○ 分等级测试

综合测评说明

采用易记词汇的智能评估算法，根据你的答题情况，动态安排单词难度，科学测评词汇量

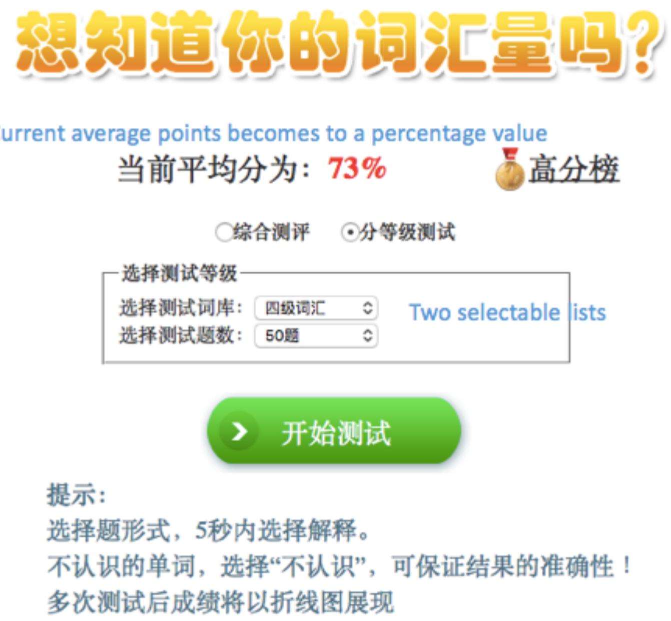
提示：

选择题形式，5秒内选择解释。

不认识的单词，选择“不认识”，可保证结果的准确性！

多次测试后成绩将以折线图展现

Figure 53 Interface for ‘General evaluation’



想知道你的词汇量吗?

Current average points becomes to a percentage value

当前平均分为：73%　高分榜

○综合测评　●分等级测试

选择测试等级

选择测试词库：四级词汇

选择测试题数：50题

Two selectable lists

提示：

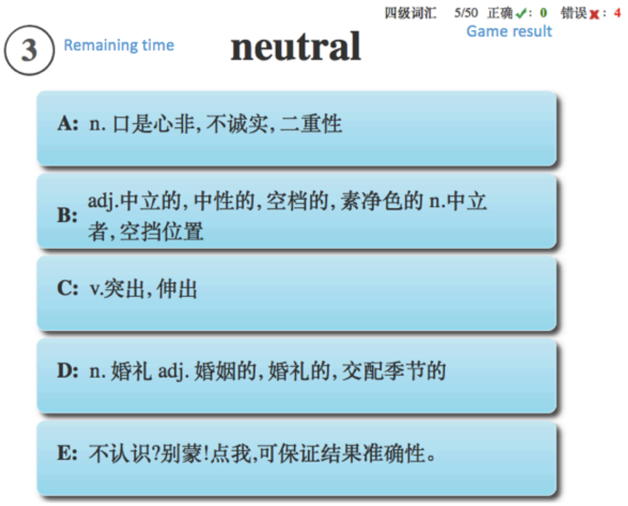
选择题形式，5秒内选择解释。

不认识的单词，选择“不认识”，可保证结果的准确性！

多次测试后成绩将以折线图展现

Figure 54 Interface for 'evaluation based on level'

After clicking the game start button, the game view turns into the main game step as showed in Figure 39. It is from selecting 'general evaluation' option. While, if a player selected 'evaluation based on level', the view will be as Figure 55. In this step, in addition to game title, options, it still has the remaining time and the game result board. For each round, the time limitation is 5 second. And this game result is a dynamic game result, showing the number of correct and wrong answer, selected game mode and difficulty level.



四级词汇　5/50　正确✓：0　错误✗：4

Game result

3　Remaining time

neutral

A: n. 口是心非，不诚实，二重性

B: adj.中立的，中性的，空档的，素净色的 n.中立者，空挡位置

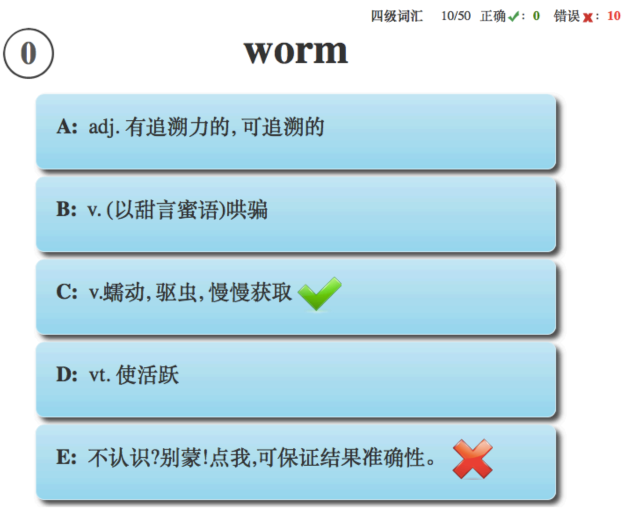
C: v.突出，伸出

D: n. 婚礼 adj. 婚姻的，婚礼的，交配季节的

E: 不认识?别蒙!点我,可保证结果准确性。

Figure 55 A game view by selected 'evaluation based on grades'

When a player selects an answer, the view turns into Figure 56. If a player selected a wrong answer, the system provides the wrong answer (red cross) and the correct answer (green checkmark) respectively. Otherwise, the system only provides the correct answer. If a player does not select an answer before time ending, the selected answer will be treated as option E.


四级词汇 10/50 正确✓: 0 错误✗: 10

0 worm

A: adj. 有追溯力的, 可追溯的

B: v. (以甜言蜜语)哄骗

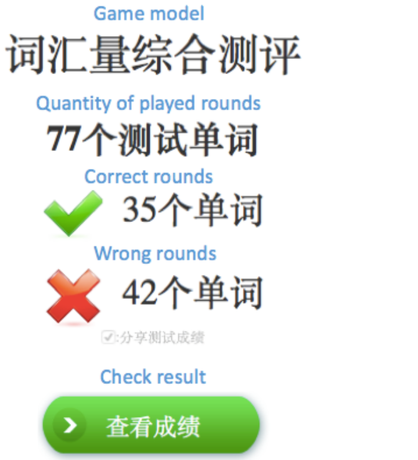
C: v.蠕动, 驱虫, 慢慢获取 ✓

D: vt. 使活跃

E: 不认识?别蒙!点我,可保证结果准确性。 ✗

Figure 56 Result of selection

After played dozens of rounds, the game turns into game result step. In this game, there are two views of game result, view 1 and view 2. Figure 57 and Figure 58 are view 1 for two game modes respectively. It is a conclusion of correctness. Both of them contain the information of the game mode, quantity of played rounds, number of correct/wrong rounds and checking the result button. If a player clicks checking result button, the view is turned into view 2.


Game model

词汇量综合测评

Quantity of played rounds

77个测试单词

Correct rounds

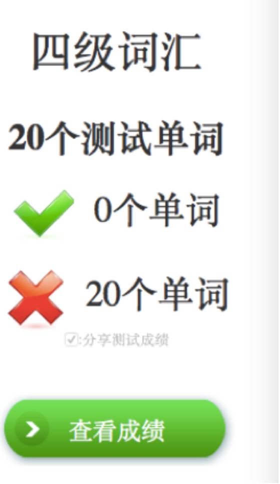
✓ 35个单词

Wrong rounds

✗ 42个单词

☑分享测试成绩

Check result


四级词汇

20个测试单词

✓ 0个单词

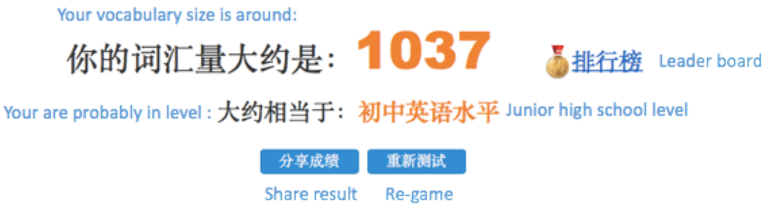
✗ 20个单词

☑分享测试成绩

Figure 57 View 1 of game model 'general evaluation'

Figure 58 View 1 of game model 'evaluation based on levels'

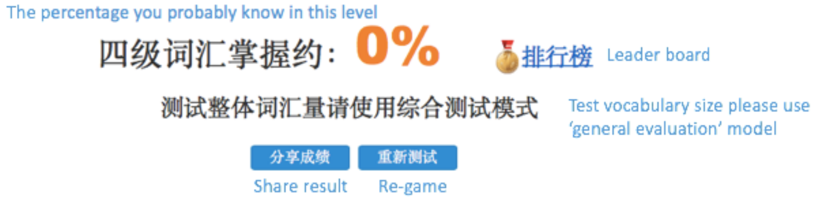
Figure 59 and Figure 60 are view2 of two game modes repsectively. The majority part is the same for both of them, including leader board link, share result and re-game buttons. The difference is coming from the way of expressing the game result. For game mode 'general evaluation', it shows the probable vocabulary size and your current English level. For game mode 'evaluation based on levels', it shows only the knowing presentage of the selected level.


Your vocabulary size is around:

你的词汇量大约是: 1037

Your are probably in level : 大约相当于: 初中英语水平 Junior high school level

Figure 59 View 2 of game model 'general evaluation'



The percentage you probably know in this level

四级词汇掌握约：0%

测试整体词汇量请使用综合测试模式

Test vocabulary size please use 'general evaluation' model

Figure 60 View 2 of game model 'evaluation based on levels'

Here is the summarising of the possible game features we have discussed. At the welcome step, the game features we need to consider are the game mode and leader board (the same one with the result step, so we put it into the result step). For the primary game step, we need to consider several game features, which are 1) Time limitation 2) Game questions 3) Error notification 4) dynamic result board. For the result step, we need to consider the sharing result and the leader board.

5.2.1 Game Features of Welcome Step

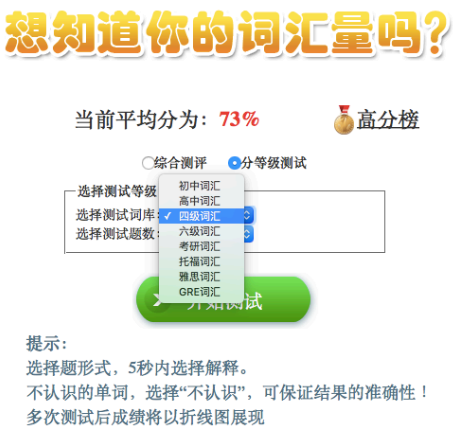
Since we will discuss the leader board in result step, the only game feature we need to consider here is the game model. But we need to think about the quantity of questions for each mode. So we treat the quantity of questions as a game feature in this section as well.

1. Game Mode

In '易记单词'， it has two game modes, 'general evaluation' and 'evaluation based on levels'. Since 'general evaluation' is a test without specifying a difficulty level and 'evaluation based on levels' is a test with specifying a difficulty level, we rename them as 'level unspecified' and 'level specified'. A player has to select a game mode to be in. Not all vocabulary challenge games have these two game models at the same time. Some of applications only has 'level un-specified' game model.

Type	Explanation
Level specified	Before challenging, the player has to select his/her current English skill level, and then the system provides the questions all with the selected English skill level. For example, TOEFL, GRE, High school, or even level 1, level 2, etc. In this game, it contains 8 levels as showed in Figure 61, they are ranked as difficulty level. The translation of these levels are, junior high school level, high school level, CET-4, CET-6, Master level, TOEFL, IELTS, GRE.
Level unspecified	The content of game is unspecified. System decides what to provide to the players. When a player does not know what level he is, he can play with this mode.

Table 12 Two game models



想知道你的词汇量吗?

当前平均分为：73%　　高分榜

○综合测评　●分等级测试

选择测试等级

选择测试词库：✓ 四级词汇

选择测试题数：

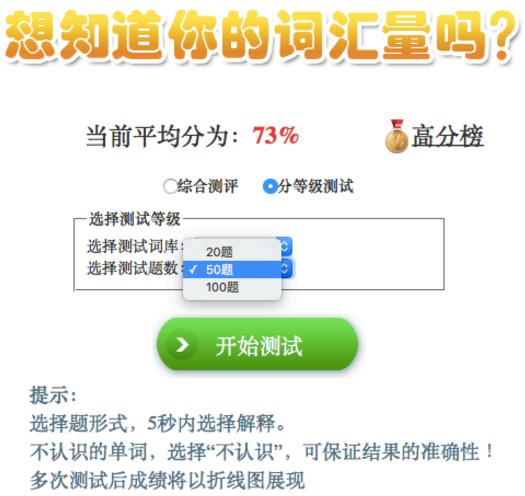
提示：
选择题形式，5秒内选择解释。
不认识的单词，选择“不认识”，可保证结果的准确性！
多次测试后成绩将以折线图展现

Figure 61 game levels for ‘易记单词’

Generally speaking, if a player knows his vocabulary size level, he can just select his corresponding level to test. While, if he has no idea which level he belongs, he can play level unspecified type. Thus, if he selects the first one, he can just play the words that in his level. Otherwise, he needs to play a lot of words that under his level to position his current vocabulary size level. Selection based format is either level specified or level unspecified. A player selects words in his specified level if he is using the type of level specified. Else, he selects from a level unspecified words list. Questionnaire is possible to be level specified, for example, ‘testing how much TOEFL words you know’, or there is a famous testing that ‘testing whether you are on English writing level’, but most of questionnaire format is with level unspecific game mode.

2. Quantity

Quantity refers to the number of rounds to be challenged (or we say the number of rounds in a section). But it is not the more the better. Too many rounds, for example 60 rounds, make game boring. For level unspecified model, it always contains more rounds by comparing to level specified model. That because, level unspecified model needs to locate a player’s skill level first. For level unspecified game mode of ‘易记单词’, the quantity of rounds for each section is not fixed. It depends on the correctness of the answered rounds. For level specified game mode, it contains 3 options, as showed in Figure 62, which are 20, 50 and 100 rounds.



想知道你的词汇量吗?

当前平均分为：73%

高分榜

○综合测评 ●分等级测试

选择测试等级

选择测试词库：

选择测试题数：

20题

✓ 50题

100题

开始测试

提示：

选择题形式，5秒内选择解释。

不认识的单词，选择“不认识”，可保证结果的准确性！

多次测试后成绩将以折线图展现

Figure 62

Above we discussed the quantity of each game model, now we need to think how they arrange the questions sequence of the testing content. It is different depending on the selected game mode. For level unspecified mode, generate speaking, it is increasing from easy to hard. But it can be separated into dynamic or static type. They are two mechanisms to select the difficulty level of the next round.

- ***Dynamic***, in ‘易记单词’, during the testing procedure, the system manages the difficulty level by analysing the correctness of the answered rounds. The finial difficulty level will not exceed to the player’s skill level.
- ***Static***, rounds are sequenced from easy to hard, regardless the correctness of the previous rounds. Static is always aligned with game content closed type in real applications.

In level specified mode, questions are randomly selected from the specified level. That is based on the assumption that a question pool for a specified level is in the same difficulty level. For example, if a player selects TOEFL as the specified level, all the questions will be randomly select from this TOEFL question pool.

5.2.2 Game Features of Main Game Step

In this part, we need to consider game features including time limitation, game questions, error notification after each round, dynamic result board and quantity of the rounds.

1. Time Limitation

It refers to the time limitation for each round. Most games do not have a time limitation. Normally speaking, the average time for each round is around 5-15 seconds, in “易记单词” they set 5 seconds as default for each round, thus, only some easy and very familiar words can be answered in time.

2. Questions

A question contains two parts, question (title) and options. A title is just an English word for all vocabulary challenge games, so we do not take it into consideration. Instead, we need to consider the questions pool size these applications provided in the game. In '易记单词', for each section, the game questions are totally different. But during the investigation process, we found that some games are providing the same questions or almost same questions for each section. For example, when you play a section in a game, it provides questions as 'apple, tee, …, number'. And next time when you play it, it provides the same questions as the last one. So, if we consider the openness of the game question pool (or we say game content), we can use content opened and closed to classify word challenge games.

Content closed: questions are fixed, and a player is playing the same questions for each time. The game content is mainly produced via human manually. It derives from the written form vocabulary test and always with a questionnaire type (it only has 30-100 rounds and can be present within an one-page questionnaire). Researchers and language education workers created these lists of words to test whether a person achieves a certain level. They selected some representative words from each English level and put them together. Since it is used to be a written form, the test content pool is always limited, generally speaking, less than one hundred, and a tester has to finish all of them. For example, a Chinese English education organization named 新东方 produced a list[39] which contains 100 words, that have been divided into 6 levels, to test the English level of learners. To evaluate the English vocabulary size, a player has to finish all these 100 words and use the following formula to evaluate vocabulary size. 180•correct rounds of level 1+280•correct rounds of level 2 …… + 192•correct rounds of level 6.

Content opened: unlike content closed, this type has a large question pool containing all English words, that is, even after several game plays, a player still cannot feel the overlap between each game section. It has to be divided into several difficulty level as well, but for each level, it contains hundreds of words. These words are not the representative ones, but since it cannot feel the duplication, the game life cycle is longer. While, in order to provide the suitable game content to players, the difficulty level has to be arranged reasonable.

Selection based format is possible to design into content closed type, but since the small question pool and selection based format is rarely to be seen itself, there is no such kind of application existing. Since questions of content closed type are sequenced as from easy to hard and covers the representative words of all difficulty levels, it is possible to make the challenge as single round type, but the game life cycle is very short. That because after few sections play, a big duplication will be found.

Options of the question

Now we discuss game options for a question. In general, there are three or four options with a 'don't know' option. While, the content of each option is slightly different. Some games provide

[39] English vocabulary size test: http://www.douban.com/group/topic/19077267/

the full Chinese translation including its entire part of speech (POS) as one option, for example, ‘adj. 平凡的， 陈腐的; n. 常事，老生常谈， 普通的东西’. Some of them only provide a brief meaning with one of its POS, for example, ‘n, 加拿大’. In ‘易记单词’, they used the first one. As showed in Figure 39, the title is apple and the options are with the meanings of ‘downhearted’, ‘minaret’, ‘exile’ and ‘apple’. We see that the words ‘downhearted, minaret, exile’ are totally unrelated with the word ‘apple’ on both morphology perspective and semantic perspective.

The options can be designed or randomly generated. The meaning of designed options is closed to the title by comparing with random options. They have slight differences, that is, the designed answer increased the accuracy of the knowing of this title by providing the comparable answers. It increased the difficulty of a round as well. Figure 63, Figure 64 and Figure 65 are concrete examples of designed options. In these examples, the English word(s) besides each Chinese option is the English translation we marked for understanding.

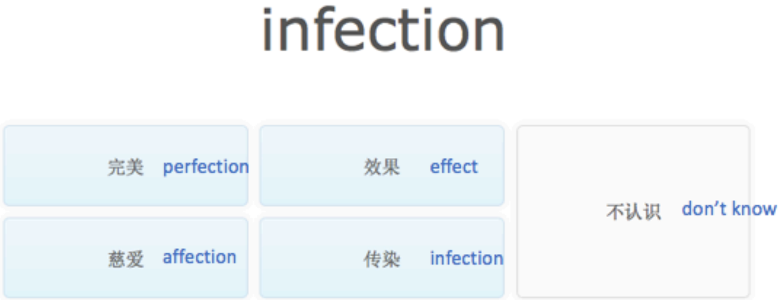


Figure 63 Example 1 of designed options

In Figure 63, the title is infection, four normal options are translated as perfection, effect, affection and infection respectively. We can see that all these words are related to an affix ‘fect’.

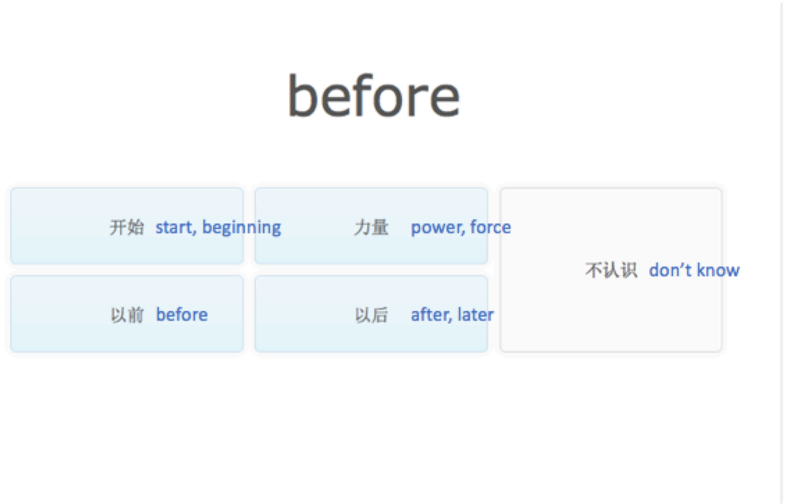


Figure 64 Example 2 of designed options

In Figure 64, title is ‘before’. The meanings of option are ‘start, beginning’, ‘before’, ‘power, force’ and ‘after, later’ respectively. We can see that, except ‘power, force’, the rest options are all related with time, that is, similar meanings.

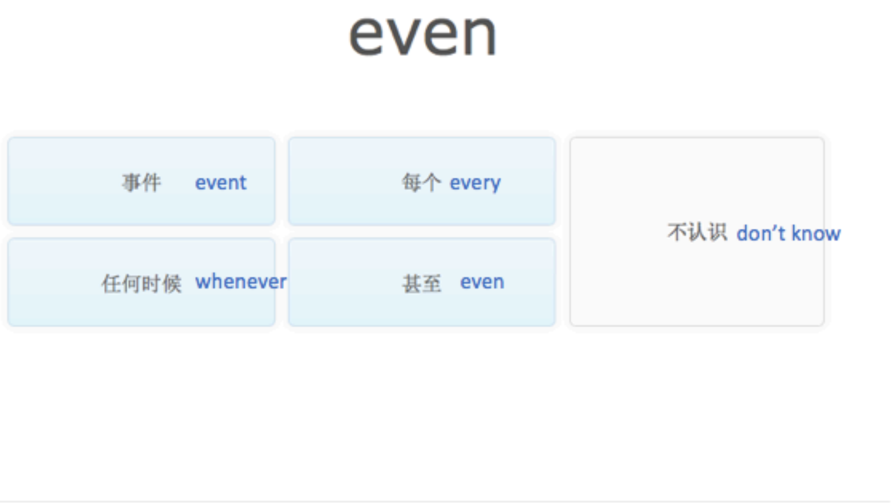


Figure 65 Example 3 of designed options

The situation of Figure 65 is same with Figure 63, all these options are related with an affix ‘eve’. But since designed options are complicated to generate, most applications are using randomly generated options.

3. Error Notification

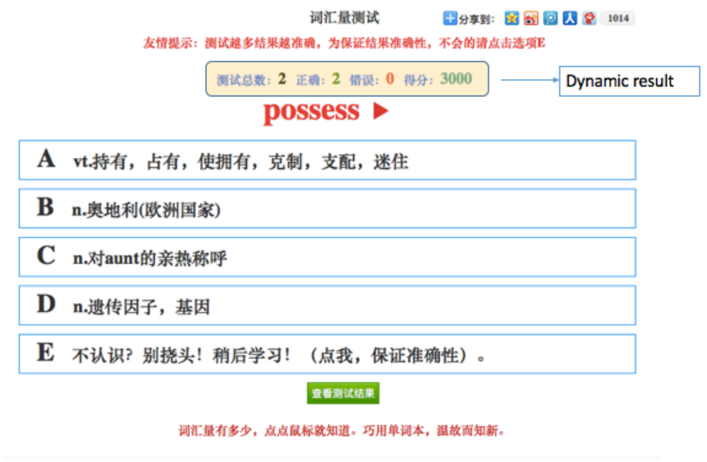
As we discussed above, in Figure 56, the game notifies correctness immediately after each round. That is, a player can know the correctness of his previous selection immediately. After each round, the screen freezes around 1-2 seconds to let the player read the result. It has 3 situations:

- A player selected a right answer. It only shows the right answers with a green checkmark on the screen.
- A player selected a wrong answer. It shows both wrong selection and the right answer.
- Time out. If a player does not select an answer in the previous round, the selected answer is treated as E. ‘don't know’.

While, not all the games have this game feature. Notifying correctness for each round increase the guessing action in some extent, since a player can always get the correct answer immediately. Players would like to guess an answer in order to let the system providing the correct answer. So some games do not provide this game feature. For example, in ‘沪江部落’, they do not notify wrong and correct answer after each round.

4. Dynamic Result Board

‘易记单词’ provides a dynamic result during the game. But it only simply provides the number of correct/wrong answers. In some games, they also provide the current game points. Figure 66 is an example for the dynamic result of game ‘轻松背单词’ and Figure 67 is the zoom of it. Its dynamic result board provides the information of quantity of the answers, number of the correct/wrong answers and the game points. In this example, the game score is the same with the vocabulary size. This number is fluctuating by casting more answers.



词汇量测试

友情提示：测试越多结果越准确，为保证结果准确性，不会的请点击选项E

测试总数：2 正确：2 错误：0 得分：3000

possess

A vt.持有，占有，使拥有，克制，支配，迷住

B n.奥地利(欧洲国家)

C n.对aunt的亲热称呼

D n.遗传因子，基因

E 不认识？别挠头！稍后学习！（点我，保证准确性）。

词汇量有多少，点点鼠标就知道。巧用单词本，温故而知新。

Figure 66 An example of dynamic result

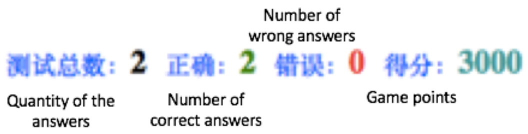
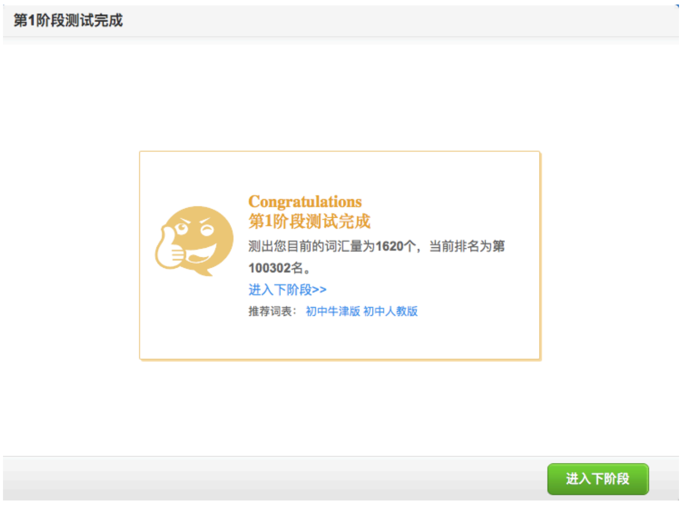


Figure 67 A dynamic result board

5.2.3 Game Features of Final Result Step

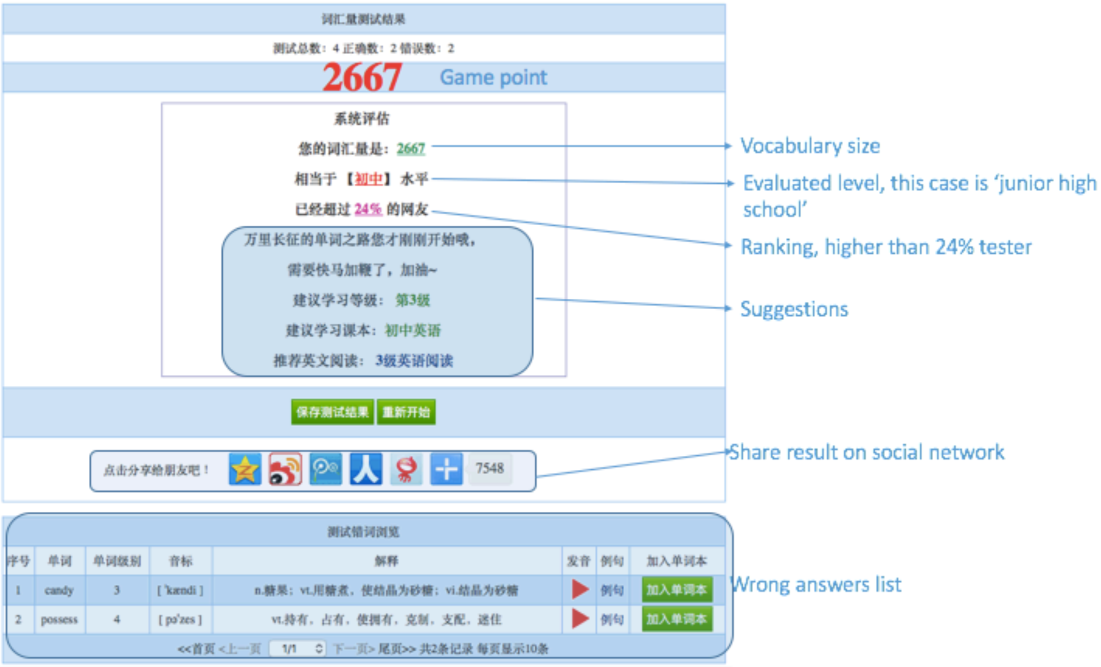
In ‘易记单词’, the final result contains two views, including number of correct/wrong answers and vocabulary size/percentage for each game modes. And it provides the sub game features like share result, leader board and re-game. For the other games, the containing information and provided game functions are different. For example, game ‘沪江部落’, as showed in Figure 68, only provides a simple game result that ‘your current vocabulary size is 1620, and ranking in 100302’. While, in this part, we need a more comprehensive understanding of ‘what a final result should have’. Because the result view of ‘轻松背单词’ contains more sub game features of final result in vocabulary challenge game, we use it as our example to illustrate the sub game features of the final result.



第1阶段测试完成

Congratulations
第1阶段测试完成
测出您目前的词汇量为1620个，当前排名为第100302名。
进入下阶段>>
推荐词表： 初中牛津版 初中人教版

Figure 68 An example of the final result



词汇量测试结果

测试总数：4 正确数：2 错误数：2

2667

系统评估

您的词汇量是：2667

相当于【初中】水平

已经超过 24% 的网友

万里长征的单词之路您才刚刚开始哦，
需要快马加鞭了，加油~
建议学习等级： 第3级
建议学习课本：初中英语
推荐英文阅读： 3级英语阅读

点击分享给朋友吧！ 7548

测试错词浏览

序号	单词	单词级别	音标	解释	发音	例句	加入单词本
1	candy	3	[ˈkændi]	n.糖果；vt.用糖煮，使结晶为砂糖；vi.结晶为砂糖	▶	例句	加入单词本
2	possess	4	[pəˈzes]	vt.持有，占有，使拥有，克制，支配，迷住	▶	例句	加入单词本

<<首页 <上一页 1/1 下一页> 尾页>> 共2条记录 每页显示10条

Game point
Vocabulary size
Evaluated level, this case is ‘junior high school’
Ranking, higher than 24% tester
Suggestions
保存测试结果
重新开始
Share result on social network
Wrong answers list

Figure 69 An example for Share result, leader board and errors list

Figure 70 is the final result view of ‘轻松背单词’, containing the quantity of the game rounds, the number of correct/wrong rounds, vocabulary size, ranking (leader board), suggestions, save result, re-game, share result and error list. But since we discussed the quantity of each mode in section 5.2.1, in this part we do not talk about it. In the following we discuss them respectively.

1. Vocabulary Size/Percentage

This is the purpose of the vocabulary challenge game. For game mode level unspecified, the game result should be the vocabulary size. Although the vocabulary size is just an approximate value, the evaluated result via different applications have a huge difference. For example, a player A tested in ‘易记单词’, his vocabulary size is 8000, but in ‘轻松背单词’ he might only get 5000 as the result. This primarily based on the mechanism that computing the vocabulary size.

2. Suggestions

Suggestion is a short phrase used to summarize the played game, for example, ‘you are just beginning, need to work hard’, and suggested a level, for example, ‘you should play with the high school level’.

3. Sharing Result

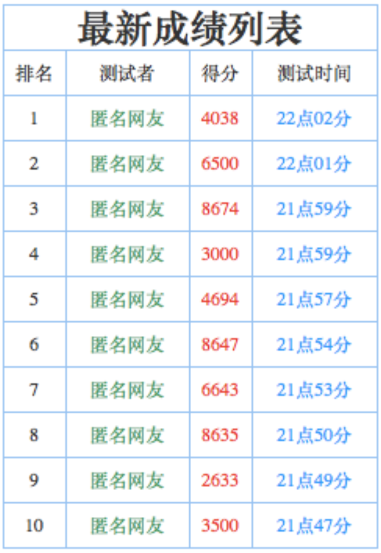
It is a game function that sharing results through social networks. As showed in Figure 69, these icons are the famous social networks in China.

4. Error List

The system provides the wrong answer list or the entire list for the continuing study after the game. As showed in Figure 69, it displays all wrong answer a player played. In this game, for each record, it includes No., word, difficulty level, phonetic symbol, explanation, pronunciation, example sentence and add to vocab.

5. Leader Board

Leader board provides a rank of the result. It could be a rough ranking, a concrete ranking or a recent ranking. Rough ranking is showed as Figure 69, ‘higher than 24% players’. Concrete ranking is showed in Figure 68, ranking in 100302. An example of recent ranking is showed in Figure 70. Actually it is a record board of last 10 testers. It includes 4 columns, which are No., tester name, game points and time respectively. In this example, since the last 10 players are all non-register users, the tester name is all displayed as an anonymous user.



最新成绩列表

排名	测试者	得分	测试时间
1	匿名网友	4038	22点02分
2	匿名网友	6500	22点01分
3	匿名网友	8674	21点59分
4	匿名网友	3000	21点59分
5	匿名网友	4694	21点57分
6	匿名网友	8647	21点54分
7	匿名网友	6643	21点53分
8	匿名网友	8635	21点50分
9	匿名网友	2633	21点49分
10	匿名网友	3500	21点47分

Figure 70 An example of recent ranking

Table 13 provides a summary of discussed game features for each game step.

Welcome step	Main game step	Result step
□ Game model	□ Time limitation	□ Suggestions
○ Quantity	□ Questions	□ Sharing result
	○ Game options	□ Leader board
	□ Error notification	□ Error list
	□ Dynamic result board	

Table 13 Game features for each game step

To compare these game features, we analysed several representative games and illustrated the result in the following table.

Name	Format	Model	Quantity	Game Options	Error notification	Result	Errors list	Share results	Time limitation	Leader board
轻松背单词[40]	Single round	Level unspecified (dynamic)	The more the better	5 options, all meanings	Show errors after every imputing immediately	Dynamic result	Yes	Yes	No	Yes
沪江部落[41]	Single round	Level specified	20 words	5 options, brief meanings	No	Show result at the end	No	No	No	No
爱词霸[42]	Single round	Level unspecified (static)	6 levels, each level has increasingly 10	4 options, brief meanings	Show errors after every input immediately	Show result after each level	No	No	No	Yes

[40] http://test.qsbdc.com

[41] http://bulo.hujiang.com/app/testword/

[42] http://word.iciba.com/?action=level

			to 20 words							
易记词汇[43]	Single round	Level specified/unspecified(dynamic)	the more the better	5 options ,all meanings	Show errors after every input immediately	After all	Yes	No	Yes Each word has 5 seconds	Yes
唯途词汇测试[44]	Questionnaire	level unspecified	60 specific words	4 options ,brief meanings	No	After all	No	No	No	No
扇贝网[45]	select lection based	Level specified/level unspecified	50 words	5 options all meanings	Show errors after every input immediately	After all	No	Yes	No	No

Table 14 Game features

[43] http://www.estudywith.us

[44] http://www.way2english.com/service/chlcs.htm

[45] https://www.shanbay.com

Chapter 6

6 Concept Challenge Game Design

We prefer to reuse an existing game instead of creating a new game to keep the playability. Our game is derived from a game named English-Chinese Language vocabulary challenge and with minor adjustments based on our purpose. In the last Chapter, we introduced the game features of all existing vocabulary challenge game in order to understand this game better. Instead of playing with words, we provide synsets as the title and options, which is an innovative style. A player has to choose the correct Chinese synset from its candidate options. Since a synset contains all English words of this meaning, a synset can be treated as a concept. So, if we provide synsets as the title, the game becomes to a concept challenge game. For a single round, the challenge point is changing from testing whether a player is knowing this English word, to whether a player has the ability to express this concept with English. For example: The concept '***the period of instruction in a school; the time period when school is in session'*** contains 3 words 'school, schooltime, school day'. If we use it as the question, this round will be:

Title: school, schooltime, school day
Option A: Chinese synset A
Option B: Chinese synset B
Option C: Chinese synset C
Option D: Chinese synset corresponding to this title

Option D is supposed to be the right answer, if a player can answer this round correctly, it means that this player knows this word sense with at least one word. It might be any of them. If a player knows at least one word that having this sense in a synset, he has the ability to express this concept in English.

Playing with concepts is an innovation, but we believe that it is a better way to test vocabulary size than playing with words. In vocabulary test[46], which is one of the most famous vocabulary test website, they claimed that the best way to play with words is from its meaning but not a word only. Currently, in word challenge game, the assumption is based on the consensus of knowing a word is defined as 'at least knowing one meaning of this word'. While, knowing a word with only one meaning is not enough to indicate that you know this word precisely. For example, you may know that nuns wear habits, but did you also know that they can fly? 'nun' is a kind of bird as well. So do you really know word 'nun'? Besides, even if one meaning is enough to indicate that you know this word, the meaning you know maybe is not the excepted one. For example, word 'approach', you only know the meaning 'near', but in a certain level, what you really need to know is the meaning 'ideas or actions intended to deal with a problem or situation'. So we think a better way to indicate a player's vocabulary size is that how many concepts he can express in English. If we test with English synsets, we are able to count how many concepts controlled. Further more, testing with meanings gives a possibility to learn more of a word, which is

[46] Test your vocab: http://testyourvocab.com/

more interesting. For example, when a player is playing a synset ‘school, shoal’, he learns that school also has the meaning of ‘a group of fish’. And in some cases can help players to reduce the possibilities of abusing a word meaning.

While, since it is difficult to implement, till now there is no such application existing that playing with meanings. At first, the biggest obstacle is that there's no easy way to organize word senses by frequency using the way we can with words. We can obtain a word frequency list by analysing a suitably large corpus, whereas the situation is different for word senses. If we want to obtain a word senses frequency list from a corpus, every word in this corpus has to mark with its sense. Till now, this work has to be done by human being, and need experts for some hard cases. But even do this by crowdsourcing, which is supposed to be the cheapest way doing this kind of work, it is unacceptable expensive to mark a whole corpus with word senses. Second, in simplified Chinese language, we do not have a reliable, free source English-Chinese linguistic resource that have the capability to be used as the baseline to develop such kind of application. Third, even in the best-known English WordNet, which is always mentioned as the baseline of English word senses disambiguation, has the polysemy issues. The senses in WordNet are overly classified for using in a word challenge game. For instance, School has 10 senses in English WordNet, and most of them are too similar. Even in word challenge game with respect to play with concepts, we do not need such fine classified senses.

Fortunately, UKC is such a linguistic resource where is able to provide English-Chinese data and at the same time it is committed to solving the polysemy issues. So if we can adopt UKC as the background linguistic resource to develop this application, it will not only help us to figure out the existing errors, but also be the first concept challenge game.

The basic idea of figuring out the existing errors is, for a round, if the most answers for players are different with our database record, we recognize this record has some problems. To create our game and ‘hide’ our goal properly, we need to consider not only the discussed game features of vocabulary challenge in previous chapter, but also the general criteria. Thus, we organised this chapter as: At first, in Chapter 6.1 we talk about the general criteria. After that, Chapter 6.2, following the sequence of welcome step, main game step and final result step, which is the same as the sequence of the last Chapter, we discussed the design of each feature based on the general criteria and real situation. After discussing game features designing, we discussed game mechanism in chapter 6.3, for example, how to compute the game result and leader board. At last, in chapter 6.4 we discuss how we adopt the feedback to figure out errors.

6.1 General Design Criteria

The general design criteria need to take three accepts into consideration, game, data and scalability. From the game point of view, we analyse it from game mechanism and interface. From the game mechanism perspective**,** the challenge is that how we can adopt or modify the discussed features to satisfy our objective. Basically, there are two kinds of play modes, single play and multi play. Single play means a player plays only once to test his vocabulary size in the short term. Multi-play means that a certain player plays several times continuously in a period. But since the vocabulary size is growing slowly, he has no need to play more during a short period. That is, if our game challenge point is only to test the vocabulary size, the possibility of multi play will be very rare. In this case, we need to think how to foster more participations. From the interface perspective, of course, a well-designed interface is helpful to attract players. From the data point of view, we should think how we could get the high quality data. And at last, from the scalability point of view, we need to think the reusability of this game.

- **Game**
  - **Playability** from the game mechanism perspective, we have already known that this game is well known and popular, the challenge is that how we can adopt or modify the discussed features to satisfy our objective in details.
    - **Fostering user participation** the more player playing, the more data we collect.
    - **Study** use study as the additional incentive in order to get more multi plays.
    - **Slightly change** We should adopt the most used game features if they do not conflict with our purpose.
  - **An attractive interface** from the game interface perspective**,** an attractive and friendly interface is necessary.
- **Data**
  - **Coverage** the game should cover all the content of UKC.
  - **Cheating** normally speaking, cheating, for instance, a player looks up a dictionary during the game, breaks the fairness of game. But, in our case, cheating is benefit to us. Since it increases the accuracy of collected data.
  - **Preventing unreasonable input** the unreasonable and guessing inputs will decrease the data quality. How to prevent them is a challenge for us.
  - **Quality control** Cheating and unreasonable inputs are the special case of player actions, still, the language skill of players is different. A player who has a higher English language ability is more reliable than these beginners of English language in general. So, distinguishing the input quality from players and select the reliable data from the collected data will be helpful for our purpose.
  - **Effective distribution** to maximize the human labor, the difficulty degree of questions should be appropriate. Providing too easy questions for a player is wasting human labor in our perspective.
- **Scalability**
  - **Reusable** as we mentioned, there are several types of error in the English-Chinese part UKC, so if this project has the ability of maximal reusable to the other objectives will save us more efforts.

6.2 Game Features design

In order to give an intuitive impression, we provide a summary at first. Table 15 is the summery of the designed features for our game. They are designed based on the general criteria we discussed in the last section and the real situation. The discussion of the game features is following the sequence of chapter 5.2. At first we decide the game format and then we design the game features for 3 game steps. After that is the discussion of each feature respectively.

Category	Game feature	Design
Game Classification	Game format	Single round type
Welcome step	Game modes	3 game modes
	Quantity	10 or 20 depending on each model
Main game step	Dynamic result board	No
	Question	Content opened
	Options	Chinese synset and corresponding POS
	Error notification	Do not show errors after each round, but show at the end in English
	Time limitation	Depending on game models
Result step	Errors list	Provide the list of all tested rounds
	Share results	Provide the function to share results
	Suggestions	Yes
	Leader board	Yes

Table 15 the summery of designed features

6.2.1 Game classification

6.2.1.1 Game format

Our game is designed as the single round format as most applications did. The other two types are not appropriate for our case. Questionnaire format is good for written type, but not good on either smart phone or Internet browser. In single round format, the selection of options is the key point to figure out errors from UKC. So we cannot use selection based format either. To make our game as the ‘study’ type, a player is recommended to register to record his study career. We should consider two situations:

□ **Single play** A player only plays once in a short time. Registration does not be too important. And he is no idea which level he is, the selection of level unspecified is in the biggest probability.

□ **Multi play** A certain player plays the game continuously. Registration is recommended after several sessions played. And the initial difficulty level of level unspecified game type should be the one he played last time. So, we should save his current difficulty level in the cache regardless he is registered or not.

By considering this is a non-immersive game, we adopted simplified login from UKC Game Framework as our login system.

6.2.2 Game Features of Welcome Step

6.2.2.1 Game Modes

In chapter 5.2, we mentioned that we have two game modes, level specified and level unspecified. Time pressure is a good element for a game and it is helpful to prevent cheating, but most vocabulary test games are without time limitation. And we have no evidences to prove that which one is better for this game. Besides, cheating benefits us on the data quality perspective. So, we use no time limitation like most vocabulary challenge games did.

As same as the most current applications did, we provide both level specified/unspecified game modes. Since:

i. A player has no idea which level he is when he first time plays this game. So he can select level unspecified to position his level and after that he can choose a certain level to play with.

ii. In level unspecified game type, the game content is setting from easy to hard. Since we have an additional purpose which is finding errors from more than 100,000 synsets. If we only set level unspecified game type, some hard synsets will never be played and some easy ones will be played again and again. That because the average word senses size of players is far less than the size of UKC. In this case, the majority part of UKC has no chance to be played. Thus, in order to increase the coverage of the game content. We use level unspecified type to increase the game coverage of UKC.

We provide two content options in level specified game type, 1) Taking a domain as the content. 2) Selecting a specific difficulty level to play. Domain specific is possible since the UKC games framework. And providing the domain content is not only helpful for taking more fun, but also helpful for cleaning the domain specific information.

Further more, we use 10 or 20 rounds in a game session. In reality, we found that the lowest number round of this game is 10, and the most number is 60 or unlimited. Suppose that the average time for one round is 10 second (our game is playing with synset, the average time is supposed to be longer than a word), 60 rounds will take at least 10 minutes. While, we found that 10 minutes are very long and making game boring as well. Less than 5 minutes will be appropriate for our case. And a short time is helpful for studying in a piece of time. For example, play one game when waiting bus. For level unspecified game type, we use unlimited rounds as the quantity, but a player can stop at any time. To ensure that, we asked around 10 players, all of them think 10-20 rounds for a game session is the best case.

Last, since we add a studying objective, we have to provide the appropriate content to the users. That means the difficulty of questions should not be too easy or too hard. For example, for a language beginner, it makes no sense to provide all hard synsets like '***expiation, atonement, propitiation***'. This synset is neither helpful for studying in the user perspective nor useful for collecting data in our perspective. And for a senior player, providing easy questions not only makes game boring, but also wastes the human labour. So, unless a player is playing level specified game type, we adopted dynamic difficulty level as our game feature. That is, the difficulty level is shifting during the game based on the answered rounds. While, to do that, at first, we need to rank all synsets by difficulty and develop the difficulty level of synset ourselves, as in Chapter 4.2. That because our game content is synsets from UKC and till now we do not have the synset ranks. Second, create a mechanism to provide the next question depending on the rounds of player's answered, as in Chapter 6.3. Furthermore, taking game modes into account. Each model should have its corresponding unique difficulty strategy to arrange the game content, this part will also discuss in Chapter 6.3.

Since the different game functions change the challenge point of the game. So, we discuss the game functions and the possible combinations of them. In this kind of serious game, always, there is only one game pressure, that is pursuing the high known vocabulary size of the player, which will give the high rank on the leader board and in this case fulfil the satisfaction to the player himself. In the other words, the existing word challenge games only have one game model normally, that is, given a question set let the player answer it, and provide an evaluated vocabulary size at the end. There are two drawbacks, at first, from the game point of view, that pressure is only focusing on the game result, but there is no restriction or pressure for the game procedure, game content or game itself. The game only cares about the game result, cheating or guessing activities are not taking into the consideration. Second, in GWAP perspective, provide more game models are benefiting us to collecting feedbacks. To give an intuitive expression of the composed models, at first, we summarized the possible alternatives functions, which we have discussed above, as showed in Table 16.

Category	Functions	Description
Time	Time option	No time limitation
Levels	Level option 1	Specified difficulty level (level 0-6)
	Level option 2	Specified Domain
	Level option 3	Unspecified level
Quantity	Quantity options 1	10 rounds for each session
	Quantity options 2	20 rounds for each session

Table 17 Alternative functions

At first, a player has to select the game mode. For level unspecified game mode, the game content is dynamic and the quantity is 20-25. For level specified game type, there are two kinds of content, of domain or of difficulty level. Each of them has 10-20 rounds. After combination, we get 3 game modes as in the following Table 18.

Type	Level	Quantity	No.	Description
Unspecified	Dynamic difficulty level	20-25	1	Finish around 20 rounds with dynamic difficulty level
Specified	Domain	10-20	2	Finish 10 or 20 rounds with selected domain
	difficulty level 0-6	10-20	3	Finish 10 or 20 rounds with selected difficulty level

Table 18 Possible game models

Two points make our game type has to be the content unfixed type. 1, our goal is to find errors in UKC, so we have to traversal the whole UKC content several times. 2, if we want to use study as an additional incentive, the game question pool should to be large enough.

6.2.3 Game features of main game step

6.2.3.1 Time Limitation

We use no time limitation for a round like most vocabulary challenge games did.

6.2.3.2 Questions

We utilized UKC games framework to generate questions. The English part is used as the game title, Chinese part as the game options. In order to ensure options more precisely, we use 'POS + Chinese synset' format. It has 5 options for each round, in which 4 of them are Chinese synsets, and the last one is 'no correct answer' option. Still, in a vocabulary challenge game, it must have an option named 'don't know'. While, 'no correct answer' and 'don't know' are both special options. Put both of them into options list will cause the option list seems strange. So we make 'don't know' option as a button in contract to 'confirm' as showed in Figure 71.

For more information about why we adopt 'No correct answer' and how to generate Chinese options, we discussed in Chapter 4.3.

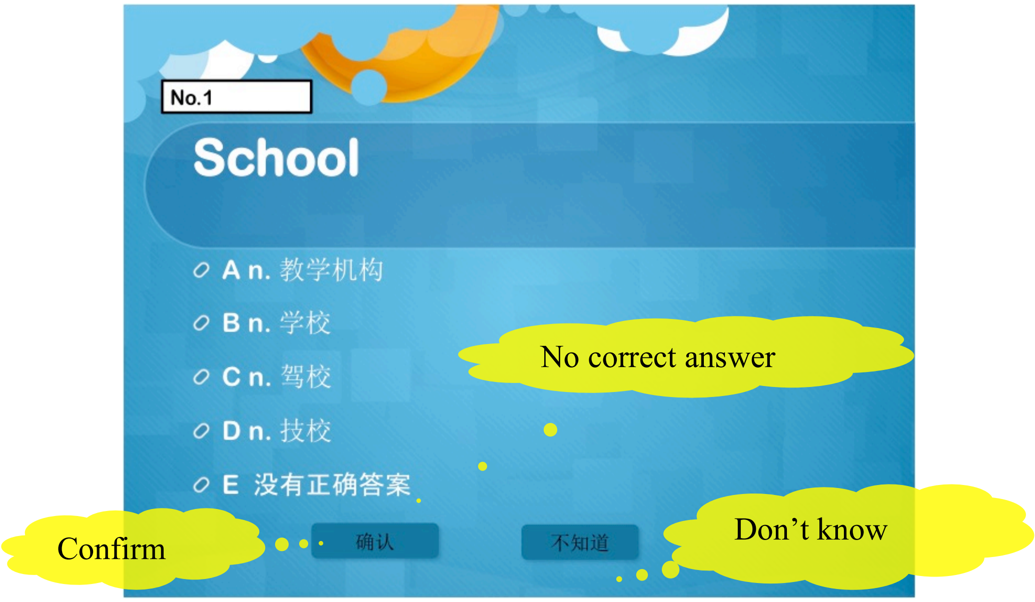
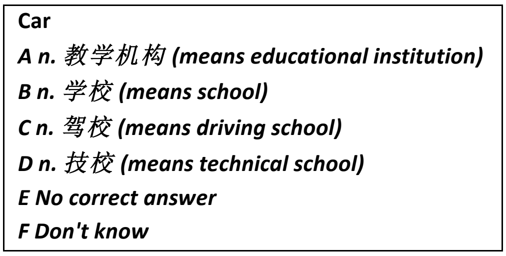


Figure 71 Prototype of the game main screen

6.2.3.3 Error Notification

We chose do not notify the wrong answers to players after each round. At first, if the system informs the error immediately, players always guess since there is no punishment mechanism and he can get notified the correct answer immediately after guessing. Second, the correctness of

English-Chinese UKC is around 90%, as we discussed, it will cause a bug if we inform an error for a correct answer. As in the following example, the answer should be E. While in our system, the default answer is B, which will confuse players. Instead, we provide the errors by marking on the English study list. That will make less confusion even we provide a wrong judgement.



Car

A n. 教学机构 (means educational institution)

B n. 学校 (means school)

C n. 驾校 (means driving school)

D n. 技校 (means technical school)

E No correct answer

F Don't know

6.2.4 Game Features of Result Step

The game result is provided after each game session. There are several factors we need to add to the game result. For players who are playing the different game difficulty level, it is unfair obviously that if we provide feedback without mentioning the difficulty level. So we provide the difficulty level, e.g. Newbie, beginner, talented, skilled, and professional, etc., as one factor in the game result. Accuracy is the main challenge of this game, so we put accuracy as one factor either. Vocabulary challenge with the goal of 'test/challenge' is a game that people only play once during a period, this is harmful for the data collecting. So we defined our game as 'study' type instead of 'test/challenge' in order to make our game as a long-term game. Without continuing study, a vocabulary challenge is a test, and after the test, a player has already evaluated his vocabulary size. Since the vocabulary size cannot change in a short period, he has no motivation to play it once more. While, if we provide a list of wrong/all challenged rounds for the studying, players have enough motivation to play it frequently. Actually, in real life, when we study English words, there are not only the unknown words, but also these words are familiar with, still cannot recognize its meaning even after a long time thinking. After a player played a game, he figures out these words, and in the continuing study, he can get the points to focus, which is more effective for studying. Thus, we provide a list of the tested rounds, including English synsets (and its related English synset is available to check), explanation and examples. We do not provide Chinese synset here.

But, that is not enough yet. Leader board is a good incentive for attracting more game plays. A player can get his rank after each session plays. While, in order to provide a reasonable and fair ranking. We need to find an integrated element (we say it as 'game points') to evaluate that rank. The game points much fulfil three rules, first, that game points must have the ability to demonstrated the integration of accuracy, the difficulty level and the played quantity rounds. Second, it is must fair enough. Third, as we mentioned in general criteria, we also need to consider how to reduce the harmful actions from the players in this game points calculating. So, our result contains these elements, Game score, Quantity or Time cost, Accuracy, difficulty level, study list and leader board. The details of calculating the game points will be introduced in Chapter 6.3.

6.3 Game Generation Mechanisms Design

6.3.1 Content Providing

As we discussed, based on the coverage of the question pool, there are two kinds of content providing mechanisms, opened and closed, and most of the current applications are adopting content opened method. That because, on one hand, in our aspect of view a large game content pool that can cover the more content of UKC, and on the other hand, in the playable aspect of view, a long game life cycle that can attract more game playing. Both of these two reasons led us to utilize open content. In this scenario, closed content is not so much related to our case, a player will not pay much time to play a game with the same content continuously. In order to understand how opened content is working, we did a comprehensive study.

To distinguish the applications with respect to opened/closed, as mentioned in Chapter 5, we played quite a lot times for each game. For a game A, after N sessions playing, we say the game content is closed if we found that the game content for all games are the same or the major part of the game content is the same. For example, there are 100 questions in a questions pool, and providing 30 rounds to a player once as a game session. After 3 or 4 sessions playing, the next session might have a big overlap with what you had just played, with which makes the game no more challenge. In this case, the size of the game question pool somehow is a crucial element for the length of the game life cycle to some extent.

For an opened content game, currently, there are 3 ways to provide game content, ***dynamic***, ***from easy to hard*** and ***level specified***. We have previously described briefly what they are in the former section. Since we chose dynamic and level specified as our game functions. In this part, we will introduce the study that how these two works in details, respectively. And after that we provide our solution based on the study.

6.3.1.1 Existing Content Providing Method Analysing

Level specific

Since level specified can be partly used into dynamic method and it is pretty straightforward to implement, so, we discuss the level specific first. There are 3 applications provide level specified mode, which are 易记单词, 沪江部落 and 扇贝网. First, we see how they classify their level. The classification of levels is different, 易记单词 is classified by famous examinations. It has eight levels, such as IELTS, TOEFL, and GRE. The category of 沪江部落 is based on education level, with 5 levels, which are primary school, junior school, high school, university and upon university. 扇贝网 is consistent with 易记单词, classified with examinations. Second, we need to learn how to arrange the word order in a level. After played, we found that, in these 3 applications, questions are all ordered from easy to hard, the difficulty is still increasing even after making a mistake.

In our case, as we discussed in Chapter 4, since a word sense is no way to judge to which education level or examination it belongs, we use our difficulty level as our classification. Now we discuss how we order the word senses in a level. Because in a level it contains 2000-4000 rounds as the question pool, we need to keep in mind that the coverage of a level should be representative, that is, the question rounds should be evenly picked from the question pool. In our case, we can just use a quantity of level/ quantity of rounds to get the interval between two question rounds. And randomly pick rounds in each interval.

Level unspecified

In order to provide clean, suitable and interesting game content, we use dynamic difficulty level as one of our game feature to evaluate a player's level first. The basic idea is that if the current round is hard, which is performing as a player answered wrong or skipped the current round, then, we provide him an easier one, vice versa. After playing existing games, we found that there are two applications satisfying our requirement, 轻松背单词 and 易记词汇, which provided both dynamic function and at the same time have a opened question pool. The idea for implementing 'dynamic' is to position a player's corresponding difficulty level precisely. While, to position the difficulty level, we need to find the perfect time to increase/decrease the difficulty level. Thus, we divided our study into two branches during analysing these two applications, D+1 and D-1, representing the difficulty level increase and difficulty level decrease respectively.

In 易记单词, each time, the system generates rounds N, after a player finished N rounds, the system will generate another N rounds to him. We can distinguish each N round because there is a significant loading time and loading bar between it. The difficulty level D of next N rounds is calculated by the correctness of the last N rounds. At the beginning, N equals to 5 and D is 1 (difficulty level is showing on the information bar dynamically). As showed in Figure 72, there are two states after a player answered N rounds, positive and negative. Positive means that last N rounds are answered perfectively. It contains two situations, 1) answer these N rounds totally correct; 2) several negative states but each with high accuracy, for example, all negative states with only 1 error, after 3 negative states, the state will swap to positive. Otherwise the state belongs to negative, that is, a state that a player answers 1 or more than 1 wrong. Difficulty level D increases only when the state is positive. When the state is negative, D remains, but the N is increased by the following method until this player gets a positive state.

Suppose that t is the quantity of negative state that a player gets in a level D. *Nc* is the current N number. Next round *Nn* is calculated as:

$$Nn = \begin{cases} 5 & t = 1 \\ 2Nc & 1 < t < 4 \end{cases}$$

When Nc = 40 and a player still got a negative state, the game is forced to the end.

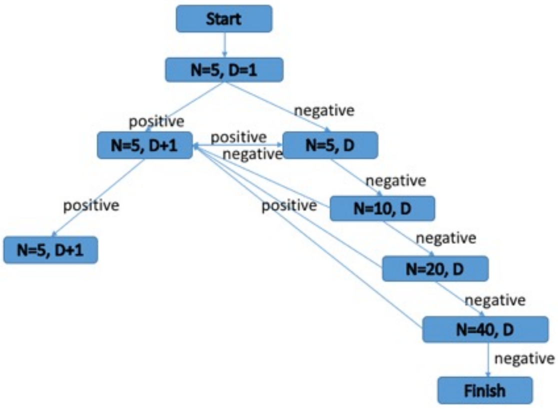


Figure 72 易记单词 dynamic content providing

Let us make a concrete example to illustrate this clearly. Suppose that a player A is playing the game. After he entered the game by selecting dynamic mode, at first, the system arranges him 5 rounds with difficulty level 1, if his answer is completely correct (positive), the system will give him the next 5 rounds with difficulty level 2, and if he can answer these rounds completely correct continually, next 5 rounds with difficulty level 3 is coming. Else, when he got any wrong answers in the current rounds (negative), he will get 5 rounds still with the current difficulty level. If he can answer them all correctly (positive), he will give next 5 rounds with one more difficulty level. Otherwise, the amount of next rounds with the current difficulty level increases to 10. If he keeps making errors, the amount of rounds in this difficulty level is keeping increasing, until: option 1, he makes once totally correct, the amount of rounds back to 5 and difficulty level add 1. Option 2, all these negative states are in high accuracy, after several N rounds, the state is set to positive. Option 3, the quantity of rounds reaches to 40 and if he still cannot get positive, this game is end. That is, his last chance to get to next difficulty level is to answer that 40 rounds 100% correct or high accuracy if last N rounds were in high accuracy either.

Thus, in this application, we see that:
D+1: When the state changes to positive, difficulty level increases.
D-1: they remain the current level with the increasing N rather than providing difficulty level decreasing mechanism.

While, in **轻松背单词,** there are not significant loading time, loading bar between rounds, or difficulty level showing on the title bar. So it is hard to judge the granularity (the exact N) for changing the difficulty level. Fortunately, a dynamic game score is displaying on the game bar. We tested the domain of this game point is [0,12000] by cheating (Since no time limitation of each round, we played by looking up a dictionary). We suppose that each 1000 game points is a difficulty level, for example, 0-1000 is difficulty level 1 and 5000-6000 is difficulty level 6. When the game point shifts in a 1000 period slightly, we assume that it is shifting in the same difficulty level. Still, there is the state of positive and negative. At the beginning, the game point of the first round is 2000 or 3000 (level 2 or 3). If a player answers N rounds correct, the state turns to positive, else, before turning to the negative state, he has N chances to make wrong answers. Here, we use N to represent the quantity of chance for turning a state. This N is changing after every turning of the state, increase or decrease. Since the quantity of N with respect to turning positive and negative is a little bit different, we use ***Np*** to represent the rounds that need to turn to the positive state and ***Nn*** to present the negative. The relative early winning streak is helpful for achieving to a high difficulty level. The game is arranged as the following rules, a player A:

1)Initially, ***Nn***=1 and ***Np***=2. If A answered the first two round correctly, he gets 3000 points, if he answered the first round wrong, he gets 2000 points.

2)***Np*** is increasing after a negative state. Every time after a player got a negative state, he has to answer more rounds to change the state to positive. For example, after A made some mistakes, if he can answer the following 3 questions totally correct, he gets positive state and D+1. And after that he needs to answer 10 questions to get D+2.

3)But, ***Np*** is resettled to 1 after a two difficulty level winning streak. That is, after answering totally correct of all the rounds in difficulty level D+1 and D+2, a player can go to next difficulty level by only answering one round correctly.

4)If now A player in a positive state, it needs to answer 3 rounds wrong continuously to get the negative state (D-1). And it needs 10 round wrong continuously to get next negative state (D-2).

5)After two negative states, ***Nn*** is settled to 2. That is, after two negative states, the each following negative state only needs to answer two rounds wrong until a player gets a positive state.

6)If the answers of a player contain wrong and right, the game point is shifting in this level until he gets a negative or positive state.

Thus, in this application:

D+1: when a player gets a positive state.

D -1: when a player gets a negative state.

After studying, we find that both of these two applications have the ability to precisely evaluate the player’s English vocabulary level. There are two main differences in the perspective of dynamic mechanism,

1)the decrease of difficulty level D. In 易记单词, difficulty level D is not able to decrease. When a player reaches level D, this D remains no matter how much errors he made in the following rounds.

2)the rounds needed to reach a player’s English level. For example, suppose that the game point 10000 of the second application is mapping to the difficulty level 10 of first application and the English vocabulary level of a player A is 5, that is, around 5000 game point. The best situation, where the assumption is that a player can answered all words below his level correctly, (for example, in real, a player is in level D, the possibility of answering wrong with level D-2 or lower than D-2 is pretty low. But, he might make some mistake in D-1) is representing as the following chart (Figure 73), in which red line represents 轻松背单词 and blue one represents 易记单词. In 轻松背单词, a player only needs 6 rounds to reach his English vocabulary level, while, in 易记单词, he needs at least 25 rounds. Even A made some mistakes in level 4, in 易记单词, he needs 5-75 rounds to get a positive state, but, in 轻松背单词, he needs 3 rounds.

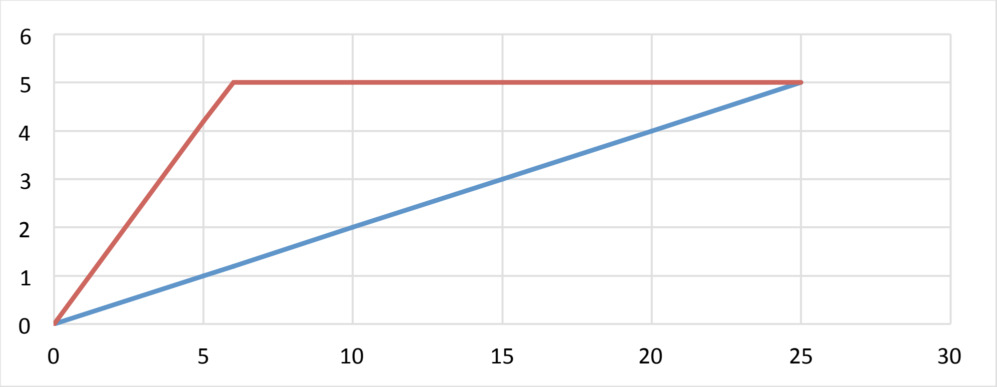


Figure 73 Steps need to reach difficulty level 5

The basic idea of D+1 for both two applications is: if a player could answer N rounds of current level in a very high accuracy, his English vocabulary size level is supposed to be higher than this level in a very high probability. The basic idea of D-1 is: if a player makes mistakes continuously, his English vocabulary skill is supposed to have not achieved current level yet.

易记单词	轻松背单词
Each D+1 operation needs more rounds answered, the accuracy of positioning level is higher. Since more correct rounds are needed to up the level, decreasing a level is supposed to be unnecessary. Players are playing more on positioning their approximate level.	It needs fewer rounds to increase the difficulty level, a player can achieve his level faster, but it is taking the higher risk of lower positioning accuracy. A word in higher difficulty level could represent it is harder, but that is not in 100% percentage. Besides, a player may know some words beyond his difficulty level. While, in the winning streak, the operation of only 1 round correct upping the level is dangerous. Since the increasing level is in low accuracy, the decrease level is inevitable when a player is making errors continuously. The positioning process is shorter; in this case, players are playing more near his difficulty level.

Table 19 Difference between two content providing methods

In Table 19, we discussed the pros and cons of two methods. In our aspect of view, for long-term games, the difference is not much, since after the first time positioning, next time the system will record this player's English level. But for a short-term game, the first method can collect more reliable answers. As we discussed, if a player is in level D, his answer below D-2 should be in high accuracy. So, based on the reason the higher accuracy rounds player answers, the better result quality we get. The first method will be benefit us more. In this situation, we choose the idea of the first application as our solution.

6.3.1.2 Content Providing Method Design

As we studied, to implement dynamic difficulty level function, we need to make sure the determination condition of swapping positive and negative states and the quantity of rounds N. Assuming the average time for finishing a round is around 10 second. In this case, the average rounds finished in 5 minutes are 30. We do not expect that a player cannot position his current level even after 30 rounds. In our case, since we have 7 levels, we have to make sure, at least, a professional player should have a chance to achieve level 7 and play some rounds before time ending. When N=3, we need at least 21 rounds to get level 7, and there left 9 rounds of level 7 to play. So we use N=3 as the basic rounds for upping level.

For the determination condition of D+1, we simply use an accuracy ***a*** as the threshold value, rather than the complicated algorithms of increasing N. If a player can get this accuracy for a level, we think his level is upon it. It is noteworthy that skipping action in the game points part is not taking punishment, but it is treated as a wrong answer at here. The accuracy is calculated by the following formula.

$$\boldsymbol{a} = \frac{Correct}{\mathrm{All}}$$

The ideal accuracy for increasing level should be around 100% percentage. But, on one hand, we have two unsatisfiable points which makes 100% is impossible, the first one is the accuracy of English-Chinese UKC. The accuracy of UKC is around 90%, it happens that even a player answered all correct but he cannot up level. Second, even for difficulty level of a word is an approximate thing evaluated by the word frequency, which has distinctions between corpuses. Our case is of word senses, which is more complicated and vague to classify the difficulty. On the other hand, each level contains 2500-4000 senses, the requirement of around 100% accuracy is too high to achieve. Thus, we follow two rules to decide the threshold of accuracy *a*.

1)Make sure game enter into N+1 rounds when a player made one or more than one mistake. Since we use 3 as the basic rounds, a player answered all 3 rounds correct in level D, his accuracy in level D is 100%, D+1. But if he made a mistake, his accuracy is 66%. While, we think 66% is not enough to judge his level is upon current level. So this accuracy a should be higher than 66% to ensure we can provide him more questions to evaluate his level.

2)An appropriate number of rounds to achieve this accuracy after making an error in the first 3 rounds. The following Table 20 is a simple table of accuracy by making 1-3 errors, which is used to find out the possible number of rounds to up a level. It is without thinking where the error makes. The error, in real, could be made between any two rounds. For example, a play made two errors in first 6 questions his accuracy is 4/6=66%, if he makes one more, his accuracy is 4/7=57%. In this table, the first row is the accuracy with making 1 error, the second is 2 and the third is 3. We think accuracy of 70%-80% is the good choice. At first, accuracy increasing slower after 70%. Second, the difficulty of level 70% is not that hard to achieve even there are some errors.

One error	2/3 66%	3/4 75%	4/5 80%										
Two errors	1/3 33%	2/4 50%	3/5 60%	4/6 66%	5/7 71%	6/8 75%	7/9 77%	8/10 80%					
Three errors	0/3 0%	1/4 25%	2/5 40%	3/6 50%	4/7 57%	5/8 63%	6/9 66%	7/10 70%	8/11 73%	9/12 75%	10/13 77%	11/14 79%	12/15 80%

Table 20 Simulation of making 1-3 errors for upgrading a difficulty level

So, we take 75% as our condition. When a player who made one mistake in the initial 3 rounds, need to answer two more rounds to up a difficulty level.

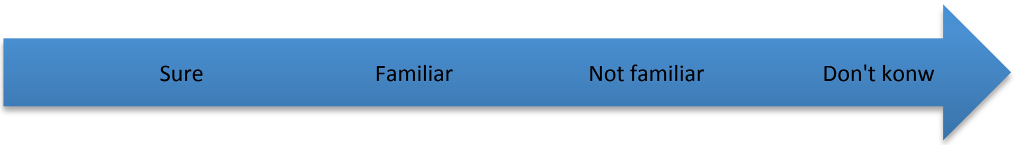
And our determination condition for swapping D is:
D+1: Accuracy > 75%
D remain: Accuracy =< 75%

For example, when a player made one mistake, the system will provide him an additional round with current level D. If he could answer the following two rounds correctly, he gets a positive state. Depending on the number of mistakes make, the following needed rounds are 5, 9, 13 and so on.

6.3.2 Game result and Leader board Computing

Since our game has several game modes, and the different game mode changes the challenge point of the game, selecting different mode will cause the different game result and leaderboard. So, we discuss the game result and leaderboard for each model. Besides, the game result reflects the player actions of a game. An appropriate game result will be helpful to reduce the probability of bad actions from players. At this point, we need to analyse the possible actions from the players at first. After that, we discuss the appropriate game result and leaderboard for each model.

Since our target player is Chinese, the understanding of the English synset (the title of a question) is the key problem of a question, in which the candidate options are assumed to be 100% understandable for a player. For an English synset, the understand state of a player is from 0 to 100%, in which 0% is “don’t know” and 100% is “sure”. We suppose that a player has four possible knowing states. From sure of a synset to ‘don't know’ of a question.



- **Sure**: a player knows current English synset perfectly, even no need to read a question clearly. Could be a question that he just played in previous games or a very easy question. For example, English synsets ‘school’, ‘time’, and ‘car’.
- **Familiar:** a player knows this synset, but needs to read/think a little bit before selecting.
- **Not familiar**: a player saw this synset before, but he cannot recognize its meaning even after thinking. He should select ‘don't know’ options instead of guessing.
- **Don’t know** a player does not know this question, never saw it before. He should select ‘don't know’ option.

As we discussed above, the ideal situation we expected is that a player only answers the questions that he is sure and familiar. When a question for a player is on the state of ‘Not familiar’ or ‘Don’t know’, the possible bad action for him should be cheating, guessing or even unreasonable input. Basically, there are four possible actions when a player in the state of Not familiar and don't know.

1, **unreasonable** or we say unintentional input. That is, a player answers a question without thinking or intention, instead, he just randomly select an answer.
2, **Cheating** a player uses additional tools during the game. His answer is more valuable.
3, **Guessing** a player answer the question by guessing. This answer is in low valuable.
4, **Honest** a player answers the questions honestly, that is, select 'Don't know' option.

In these four actions, honest answer is the required action; the other three actions are all bad actions that we want to filter out. Unreasonable answer is easy to remove since the correctness of unreasonable input for a session should be very low. So, at this part we should keep in mind how to reduce the possibility of guessing and cheating. Before that, we analyse the challenge point for the game mode.

Mode	Challenge	Bad action
Play with domains	The quantity of concepts knew in a domain	Cheating Guessing
Play with difficulty levels	The quantity of concepts knew in a difficulty level	Cheating Guessing
Dynamic difficulty level	The quantity of concepts knew totally	Cheating Guessing

Table 21 Challenges of game modes

There is no time limitation, for a round, a player can play a round as long as he wants. In this case cheating exists. To design an appropriate game result and leaderboard, the rule we need to follow is fairness in the player perspective and preventing bad action in our perspective. Since cheating is benefiting us, we discuss it in the fairness part. In the following discussion, we at first think how to prevent guessing and then consider how to keep fairness.

6.3.2.1 Game Score

To prevent guessing, we encourage players to do more honest actions. We should keep in mind that a player would like to choose the low cost action instead of the higher one. That is, we need to design our game result as by comparing to 'Don't know', whereas guessing is more harm than good. On one point, we have to reduce the cost of choosing 'don't know' action. The punishment of skipping 'not familiar' and 'don't know' states by selecting 'don't know' option will decrease the possibility of it, in the other words, it increases guessing actions. So, it is a benefit that the game result is to encourage or no punishment when player select 'don't know' option. Since skipping is negative, obviously, the design of no punishment of a negative action is better than encouraging. But even choosing 'don't know' has no punishment, we still can not make sure it is the lower cost action by comparing to guessing in the situation of time unlimited. Answering a round correctly is a positive action, which can get some game points, either by guessing or answer honestly. Thus, for a wrong answer, we use minus game points to increase the risk of guessing, which is helpful to augment the guessing cost.

The general game score format:

$$\textbf{Game points} = \sum_{i=0}^{6} (1+i)(\text{correctanswer}_i - \text{wronganswer}_i) \times 100$$

*In which **i** is the difficulty level. **Correctanswer**$_i$ is the correct answer of the difficulty level i. And **wronganswer**$_i$ is the wrong answer of the difficulty level i. The parameter 100 means each correct answer has 100 basic points. And we set that if game points < 0, we use 0 as the result.*

While in game of domains, we still take difficulty level into consideration. That because we only create one leaderboard for domain game, and the difficulty between domains is different. In a dynamic game, supposed that a player finished 30 rounds in 5mins, in which 5 rounds are skipped. In the remaining 25 rounds, 20 rounds are correct and 5 rounds are wrong. His state for each difficulty level is Correct [5,5,5,4,1,0,0], Wrong [0,0,0,1,4,0,0]. In this case, his final game point is 1480.

Suppose a player reaches level 6, depending on the content providing mechanism, means that in last 6 levels his accuracy is higher than 80%. We suppose he is in the best situation, which is the last 6 level were in 100% accuracy. In this case, an approximate game points for each level is in Table 22.

	level 0	level 1	level 2	level 3	level 4	level 5	level 6
Basic points for each level	0	500	1500	3000	5000	7500	10500
Each round points	100	200	300	400	500	600	700

Table 22 Basic points for each level in dynamic difficulty level game

Our ideal model for game point is for a senior player should get more points than a junior player in the situation that both of them are answered honestly. So we use 100 as the difference for the smoothed point between two adjacent difficulty levels in order to avoid the situation like in the following.

> A player intended on keeping in a low level and answered very fast. For example, the parameter was initially conceived to $\left(1+\frac{i}{10}\right)$, a player can try to stay in level 0 by skipping some rounds on purpose. In the best situation, a player who in level 5 and answered 30 rounds totally in 5 minutes. His game point is 3800. While, if he stays in level 0 by skipping some rounds without wasting time on reading, he has a chance to answer more than 40 rounds, since level 0 is very easy. In this case, he can get 4000 points. That does not constitute an expected situation. But for the parameter of (1+i), in the same situation with the last example, a player who in level 5 gets 11000 point. He needs to answer more than 110 rounds in level 0 to get the same point, which is impossible.

6.3.2.2 Leader Board

We provide one leader board for each game mode for all languages. And in game menu, we provide an additional leader board access for each leaderboard as in Figure 74. Figure 75 shows each leader board respectively.

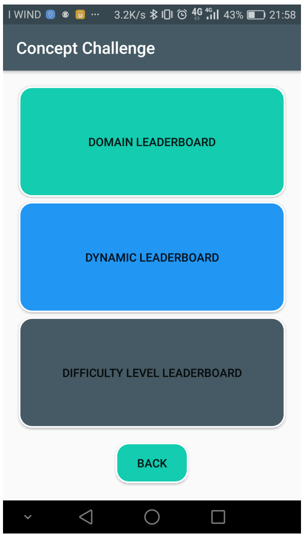
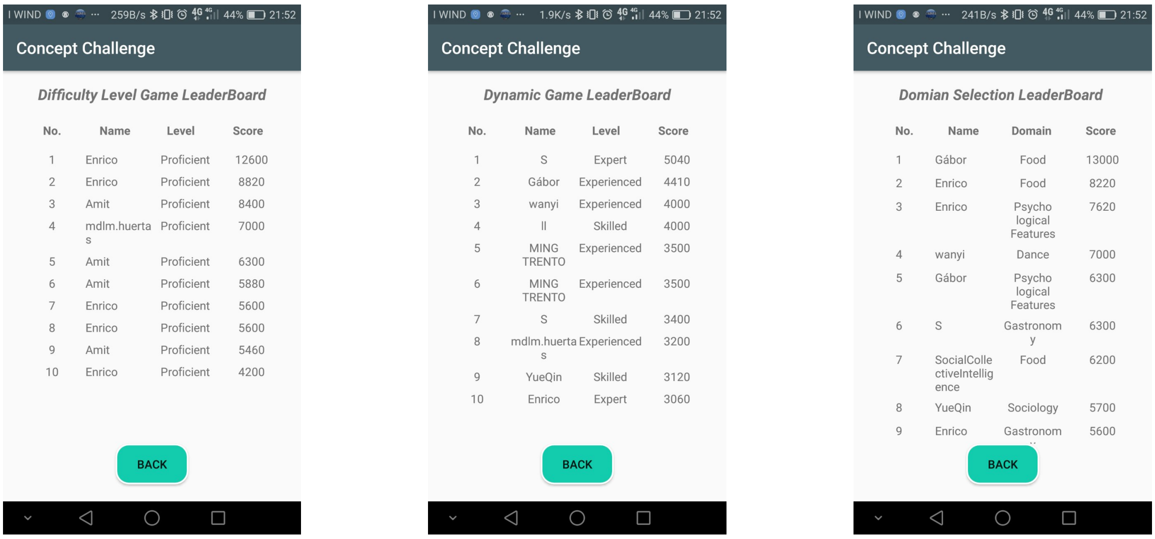


Figure 74 Leader board access



Difficulty Level Game LeaderBoard

No.	Name	Level	Score
1	Enrico	Proficient	12600
2	Enrico	Proficient	8820
3	Amit	Proficient	8400
4	mdlm.huertas	Proficient	7000
5	Amit	Proficient	6300
6	Amit	Proficient	5880
7	Enrico	Proficient	5600
8	Enrico	Proficient	5600
9	Amit	Proficient	5460
10	Enrico	Proficient	4200

Dynamic Game LeaderBoard

No.	Name	Level	Score
1	S	Expert	5040
2	Gábor	Experienced	4410
3	wanyi	Experienced	4000
4	ll	Skilled	4000
5	MING TRENTO	Experienced	3500
6	MING TRENTO	Experienced	3500
7	S	Skilled	3400
8	mdlm.huertas	Experienced	3200
9	YueQin	Skilled	3120
10	Enrico	Expert	3060

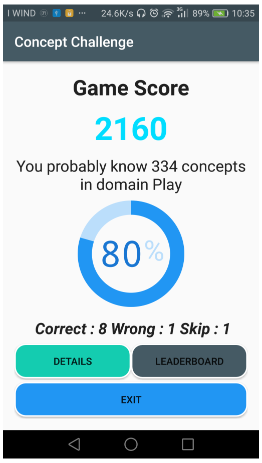
Domian Selection LeaderBoard

No.	Name	Domain	Score
1	Gábor	Food	13000
2	Enrico	Food	8220
3	Enrico	Psychological Features	7620
4	wanyi	Dance	7000
5	Gábor	Psychological Features	6300
6	S	Gastronomy	6300
7	SocialCollectiveIntelligence	Food	6200
8	YueQin	Sociology	5700
9	Enrico	Gastronom	5600

Figure 75 Three leader boards

6.3.2.3 Game Result

In addition to leaderboard of current playing game mode, we also provide accuracy, game score, recommendation and answered questions details as showed in Figure 76 and Figure 77. Recommendation information slightly changes based on each mode. In answered questions details, we provide correct tab, wrong tab and skip tab, which displays corrected questions, wrong questions and skipped questions respectively, including English synsets, English gloss and Examples.



Game Score

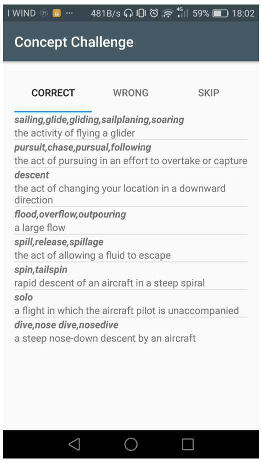
2160

You probably know 334 concepts in domain Play

80%

Correct : 8 Wrong : 1 Skip : 1

Figure 76 Game result



sailing,glide,gliding,sailplaning,soaring
the activity of flying a glider

pursuit,chase,pursual,following
the act of pursuing in an effort to overtake or capture

descent
the act of changing your location in a downward direction

flood,overflow,outpouring
a large flow

spill,release,spillage
the act of allowing a fluid to escape

spin,tailspin
rapid descent of an aircraft in a steep spiral

solo
a flight in which the aircraft pilot is unaccompanied

dive,nose dive,nosedive
a steep nose-down descent by an aircraft

Figure 77 Details of answered questions

6.4 Game Feedback Analysing Design

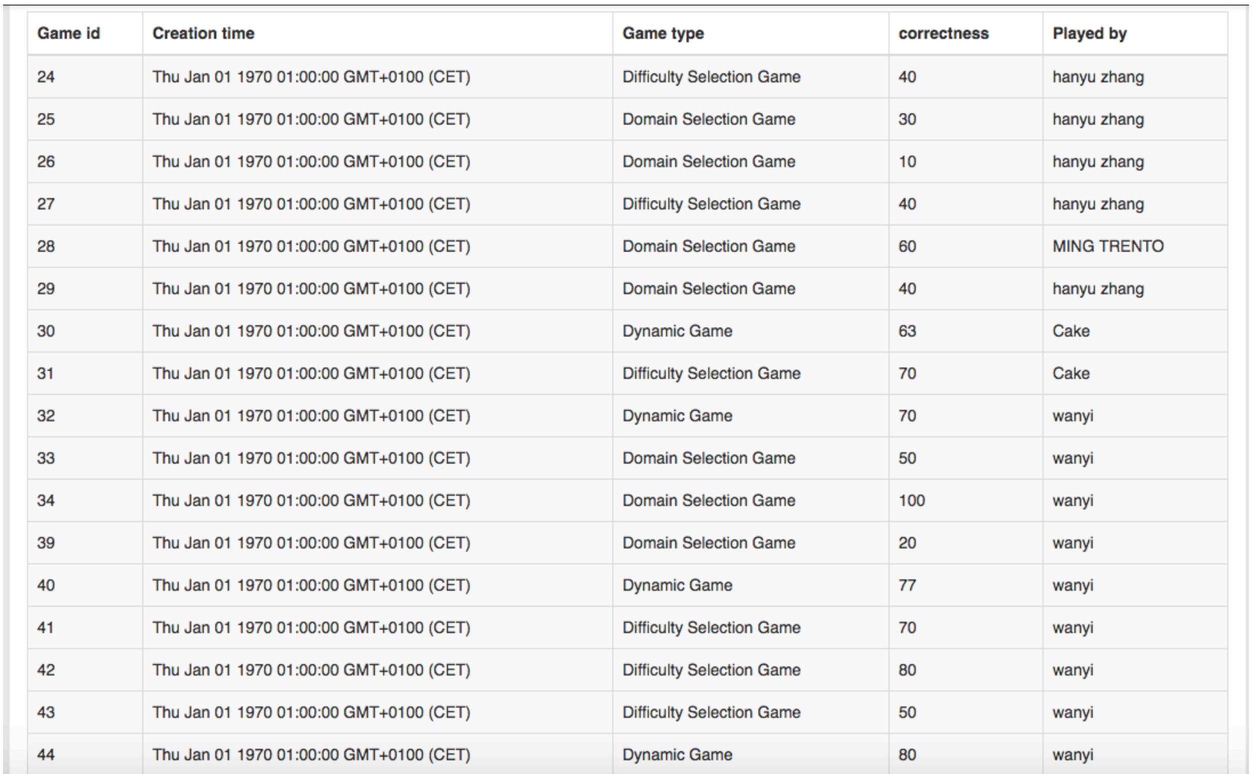
Our primary goal is to figure out the translation errors in UKC (English - Chinese pattern as the case study) and at the same time evolving our difficulty level system. The Concept challenge game is designed as implicit contribution, that is, feedback data are totally depending on the normal game playing rather than asking a player to submit some explicit feedbacks. As some games are asking people to report errors. When players are playing the game, the playing data is generated and collected, and these game feedback data will be used to analyse the translation errors in UKC. The basic idea of utilizing the game feedback is based on the assumption that the majority answers are the correct answer.

In this section, we use what, when and how to introduce the game feedback design. In section 6.4.1, we will introduce what feedback will be collected and when we collect it. In the following section 6.4.2, we will discuss how to utilize the collected data.

6.4.1 Data Collecting

The game requests a player answer questions honestly, that is, the player has to select the "Don't know" option when he is not with a strong confidence for a question. So, in the ideal situation, the answer we got should be only the confidence one, and the correctness of each game should be 100% when getting rid of the unknown rounds. In this situation, each played answer is wrathful for us. For example, in Test your vocab[47], the test result computed by the assumption that a user is 100% honest in terms of knowing at least one definition of answered words. But there is a distance between ideality and reality. When we inspect our game results as showed in Figure 78, 100% correctness is bare to see. In this example, 100% only happened only once.

[47] Test your vocab website: http://testyourvocab



Game id	Creation time	Game type	correctness	Played by
24	Thu Jan 01 1970 01:00:00 GMT+0100 (CET)	Difficulty Selection Game	40	hanyu zhang
25	Thu Jan 01 1970 01:00:00 GMT+0100 (CET)	Domain Selection Game	30	hanyu zhang
26	Thu Jan 01 1970 01:00:00 GMT+0100 (CET)	Domain Selection Game	10	hanyu zhang
27	Thu Jan 01 1970 01:00:00 GMT+0100 (CET)	Difficulty Selection Game	40	hanyu zhang
28	Thu Jan 01 1970 01:00:00 GMT+0100 (CET)	Domain Selection Game	60	MING TRENTO
29	Thu Jan 01 1970 01:00:00 GMT+0100 (CET)	Domain Selection Game	40	hanyu zhang
30	Thu Jan 01 1970 01:00:00 GMT+0100 (CET)	Dynamic Game	63	Cake
31	Thu Jan 01 1970 01:00:00 GMT+0100 (CET)	Difficulty Selection Game	70	Cake
32	Thu Jan 01 1970 01:00:00 GMT+0100 (CET)	Dynamic Game	70	wanyi
33	Thu Jan 01 1970 01:00:00 GMT+0100 (CET)	Domain Selection Game	50	wanyi
34	Thu Jan 01 1970 01:00:00 GMT+0100 (CET)	Domain Selection Game	100	wanyi
39	Thu Jan 01 1970 01:00:00 GMT+0100 (CET)	Domain Selection Game	20	wanyi
40	Thu Jan 01 1970 01:00:00 GMT+0100 (CET)	Dynamic Game	77	wanyi
41	Thu Jan 01 1970 01:00:00 GMT+0100 (CET)	Difficulty Selection Game	70	wanyi
42	Thu Jan 01 1970 01:00:00 GMT+0100 (CET)	Difficulty Selection Game	80	wanyi
43	Thu Jan 01 1970 01:00:00 GMT+0100 (CET)	Difficulty Selection Game	50	wanyi
44	Thu Jan 01 1970 01:00:00 GMT+0100 (CET)	Dynamic Game	80	wanyi

Figure 78 Part of game results

The quality of the collected feedback is influencing the quality of our goal directly. The low quality of the collected feedback will cause our result in low accuracy either. As we mentioned in the design rules, we need to filter the collected data to enhance the overall quality. Basically, if we analyze the data quality in the player action point of view, since the different motivations, there are four possible actions, which are unreasonable, cheating, guessing and honest answer, when a player plays a round of a game. It reflects a state of being that how a player answers a question. In the other words, it shows that how much we can trust a feedback.

If we can distinguish unreasonable and guessing actions, we could improve our data quality. In order to monitor the feedback quality, our feedback model is designed as player-games-answers as in Figure 79.

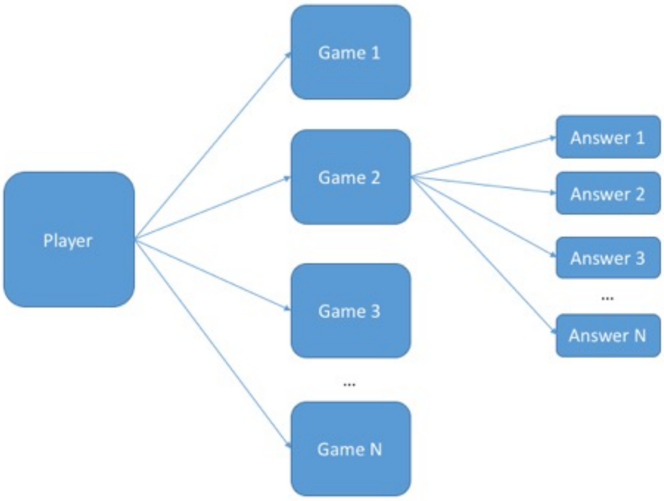


Figure 79Feedback model

The feedback model has three basic entities, Player, Game and Answer. For each player, he/she can play a lot of games, and each game has from 10 to 30 answers. The advantage is that we can easily control the feedback quality. For example, if a game is in low correctness, or a player is always in low correctness, or even a player always finishes a game too fast, we can assume that he is always with guessing actions and his feedbacks is worthless. After some tests, we found that the game accuracy of unreasonable action is always lower than 20%. We discuss each entity respectively in the following. For each entity, we need to consider when to create it temporarily (create in memory) and persist permanently (save in server), and also what to be collected for each of them. Firstly, let us discuss when we create and save them.

Player: the information of a player is gathered when a player first time opens the Word Challenge Game. If the first time opening action is detected, the system will simply ask a player to input a nickname instead of a complicated registration process. We use the Android secure ID as the unique identifier and simplified login to distinguish the players.

Game: a game entity is created when a game is starting, but it only be persisted when the game is completely finished. In this case, a game will not be recorded if the player quit in the middle of a game. The quit action here is indicating "click Back button twice". The game will show up a message "click again to quit the current game" when the back button has clicked once. So, the disruption by the phone call or game crash will not lose the current game process, a player can play this game continually later on. And after finishing all rounds, this game will be saved. The game records are persisted on both server and android phone side in order to implement the game function 'Check Game History' conveniently and decrease the server load.

Answer: an answer entity is created when a round is starting and temporarily saved in Game entity. It will be persisted with Game entity together.

Secondly, let us discuss what we collected for each entity. In Table 23, we list all the crucial content that need to be collected in the feedback. After that we will explain each of them respectively.

Entity	Content
Player	Player nickname, Android secure ID
Game	Game type, Option type, Player ID, Correctness
Answer	Selected option ID, Answer status, Question ID, Game ID, Time Cost

Table 23 Collected game content

For a player, we collect:

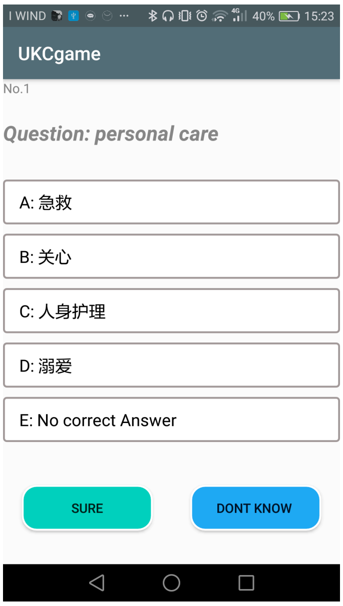
- **Player nickname**: a name used to show on greetings and leaderboard. And it is a non-strict human readable identifier in the server.
- **Android secure ID**: a strict identifier for each user. It has the advantages that easy to use, no duplicate, a player could avoid a complicate register procedure. But it is also taking the risk that this Id is changed if a smart phone is reset to the factory setting.

A game is used to encapsulate a group of answers. It contains the common information of these answers as following:

- **Game type**: used to record which game type is played. We have three game types, which are play dynamically, play with domains and play with difficulty levels.
- **Option type ID**: in order to record the selected option type. The option type can be swapped in game menu, including four types, related options, semi-related options, domain-related options and random options.
- **Player**: to record the person who played this game.
- **Correctness**: is the correctness of the entire game.

While, for an answer, although we have 3 game modes, for all kind of modes only one kind of feedback is collected during the game, As showed in Figure 80, each round has seven options. A,B,C,D,E (all wrong) and “Don’t know” and “Sure” ,in which “Don’t know” as a single button in contrast to ’Sure’.

- **Answer status**: The answer statuses are in three conditions, correct, wrong and skip. Correct means the player answer is the same with our supposed answer. Wrong is opposite with correct. Skip indicates a player skip this round by clicking “Don’t know” button.
- **Time cost**: the time used for this round.



No.1

Question: personal care

A: 急救

B: 关心

C: 人身护理

D: 溺爱

E: No correct Answer

Figure 80 Screenshot of a game round

- **Selected option**: we assigned a unique ID for each option in the server part. To understand a player selection clearly, we record the selected option ID directly instead of the label of each option (like A, B, C, D or E).
- **Question id**: to record which question has been played.
- **Game id**: to record which game it belongs.

6.4.2 Finding An Error

After we collected some feedbacks, we need to think about how we utilize it to find an error. Our assumption is that, for a round, the majority answer is the correct answer. We use ***assumed answer*** to indicate the right answer from KB, ***real answer*** to indicate the answer that it should be and ***player answer*** means the answer from players. For example, for a round, there are 10 answers. If most of player answers are the same with an assumed answer, we think that assumed answer is the real answer. At first, before talking about the majority, let us discuss the situation of collected feedback for a single round. For an answer, it has the two possibilities, positive and negative. While, not all inputs are useful to find an error.

- **Situation 1: positive**
  - A player chose the assumed answer
- **Situation 2: negative**
  - A player chose ***the other answer***
  - A player chose ***all wrong***
  - A player chose ***don’t know*** *(worthless)*

The positive situation is that player answer equals to assumed answer. And the negative situation is the case that player answer is not the assumed answer. It could be divided into three situations, a player chooses options of ABCD, a player chooses option all wrong and a player choose option don't know. For example:

Title: school

A: 学校 （school, school institution）
B: 教育机构 (education institution)
C: 驾校 (driving school)
D: 学院 (academy)
E : No correct answer
F : Don’t know

Table 24 An example of a round

In this example, option A is the assumed answer. Our game is a single selection game; a player can select only one option at a round. A player selects the option ***‘don't know’*** means that he is no idea or weak knowing of this question. That is, he is abstained from this round. Thus, the skip action is worthless for us in this task. While, if a player selected BCD, means he has some confidence to answer this right (wrong answer will be punished). Since we use similar options in our game, it has a small possibility that the option BCD is the real answer. In this example, option D is correct in some cases, but not precise enough. In a Chinese – English dictionary, school can be translated into ‘学校, 学院’. And academy can be translated into ‘学院’. In this case, we have to collect BCD as the feedback also. A player selects ‘***all wrong’*** refers to that he has some confidence of this question also. So, for a player, we count when he selects the assumed answer or ‘no correct answer’ option. And we also collect option BCD, which are not the assumed answer.

As above, we discussed the situation of a single answer, but before talking about multiple answers, we need to consider a situation that a player answered a question more than once. In word challenge game, it has the probability that two games have a small overlap. So, we need to consider, how we deal with the multiple answers from the same player for a question. In Table 25, we get 9 answers for question ‘long haul’ (number in brackets is the amount of answers). Since ‘Don't know’ is worthless here, we have 6 worthiness answers totally. Two of the A selection are from a player named Hanyu. A player named wanyi selected A and D in two different games. So, the problem is, how we count the amount for the voting system?

Title: long haul
A: 持久 [3]
B: 带狗撬在雪上的旅行 [1]
C: 旅程，旅行，历程 [0] ‘
D: 驾车，搭便车，乘车[2]
E : No correct answer []
F : Don’t know [3]

Table 25 A summary of a round

A lot of reasons can cause a player selected different answers by playing different games. For instance, he selected D at the first time playing. After game he found he made a mistake by checking the game details. So the second time he selected A. But it also could be guessing or unreasonable input. In this case we do not know which answer can be used to indicate his opinion. Fortunately, since we have a huge question pool, the probability of overlapping synsets is very limited. For example, in difficulty selection game model, each difficulty level has thousands of synsets. But for a game we just randomly select 10-30 synsets from thousands, it is hard to selected the same synsets. So, we can just discard this situation and treat an answer as a unit in the voting system without concerning with the same player.

For multiple answers, it becomes complicated. The ideal situation is that the feedbacks are all in situation 1, which means all the feedbacks from the players are the same with the assumed answer; or all in situation 2, means all the feedbacks from the players show that the assumed answer is wrong. However, in reality, the feedback comes from both of them. Table 25 is an example of related options. The question is synset 'long haul'. For a word 'long haul' has two related synsets: 1) {long haul} means a journey over a long distance. 2) {long run, long haul} means a period of time sufficient for factors to work themselves out. In our database, the assumed answer is '持久' (in real game, the options are in the random sequence). In some dictionaries, the Chinese word '持久' can be used to indicate the second synset. But, it cannot be used to indicate the first one. In this case, this record is wrong in our database. In this example, till now it has 9 answers and 3 of them are skipped. So we have 6 useful answers. In these 6 answers, 3 of them are the same with the assumed answer and the rest are not. In this case, we need to consider how to judge the result properly.

Chapter 7

7 Evaluation

The primary goal of our task is to figure out different kinds of errors in UKC via games. To achieve that, we developed a game named Concept Challenge game. In this Chapter, we will illustrate our games results in details. The evaluation work can be divided into two parts, where the first part is the user experience, and the second part is to see whether it is possible to figure out these errors via this game by marking both correct and wrong for each record in UKC. The following chapter is organized as, 1) introduce the dataset used to evaluate, 2) user experience, 3) game results evaluation. 4) we extended our work to Italian language to check whether the game is extendable.

7.1 Evaluation Data Settings

The basic idea to find an error is voting by answers, that is, for each record in the game database we need to collect a certain number of answers in order to decide whether it is correct or not. But there are around 100,000 Chinese records in Chinese LKC, which means it is difficult to find the overlaps of answers for each record in a short period if the entire Chinese LKC is used as the game question pool. Thus, to evaluate whether people can find different kinds errors, we selected 3000 questions out of Chinese LKC. According to the reason that game mode 'Play with domains' needs domain data, the dataset was selected by selecting a seed randomly and then parse 3000 possible questions around. Otherwise, the randomly selected dataset may not contain enough domain data. After several days of trail, we found that 604 questions were played, thus, in order to get more overlaps for each question, we decreased our dataset to these 604 questions in the further evaluation.

In Concept Challenge Game, 'Play with domain' is ranged from 10-20 rounds depending on a player selection. In this case, the minimum dataset for each option set is 20. After decreasing the dataset from 3000 to 604, around 10 domains were extracted, in which 5 of them are satisfied the minimum 20 requirement. We removed domain 'Factotum' since it is used to category these Sysnets that cannot find a domain in WordNet Domains. Table 26 shows the calculation of the rest four domains for each kind of option set. The difficulty for playing an option set is Related option > Semi-related option > Domain-related option > Random option. Since retrieving a related option set has the most limitations, the sum of the related option set should be the least. The difficulty for retrieving the option sets should follow the sequence of random option, domain-related option, semi-related option and related option, from easy to hard. And random option can be obtained always.

	Related option	Semi-related option	Domain-related option	Random option
Sport	46	47	50	50
Sociology	22	28	28	28
Play	37	43	44	44
Dance	31	34	34	34
Sum	136	152	156	156

Table 26 Chinese domain dataset

Table 27 is distribution for each difficulty level (DL) with respect to each option set. Since the chance to generate random option set is 100%, the sum of the random option set is 604, which is the total number of data set.

	Related option set	Semi-related option set	Domain-related option set	Random option set
DL 0	136	143	127	147
DL 1	74	80	65	83
DL 2	56	59	52	60
DL 3	50	51	47	52
DL 4	34	34	33	35
DL 5	34	39	40	40
DL 6	158	185	175	188
Sum	542	591	539	604

Table 27 Chinese difficulty level dataset

The game is further extended to Italian language. The dataset summary of domain and difficulty level for Italian is shown in Table 28 and Table 29, respectively.

	Related option	Semi-related option	Domain-related option	Random option
Gastronomy	225	195	376	379
Food	58	82	82	92
Factotum	120	141	183	183
Psychological	106	206	213	214
Sum	509	588	854	868

Table 28 Italian domain dataset

	Related option	Semi-related option	Domain-related option	Random option
DL 0	46	50	70	72
DL 1	37	47	54	55
DL 2	34	44	53	54
DL 3	59	71	80	84
DL 4	64	78	104	106
DL 5	60	76	101	104
DL 6	431	511	737	758
Sum	731	877	1199	1233

Table 29 Italian difficulty level dataset

7.2 Collected Data

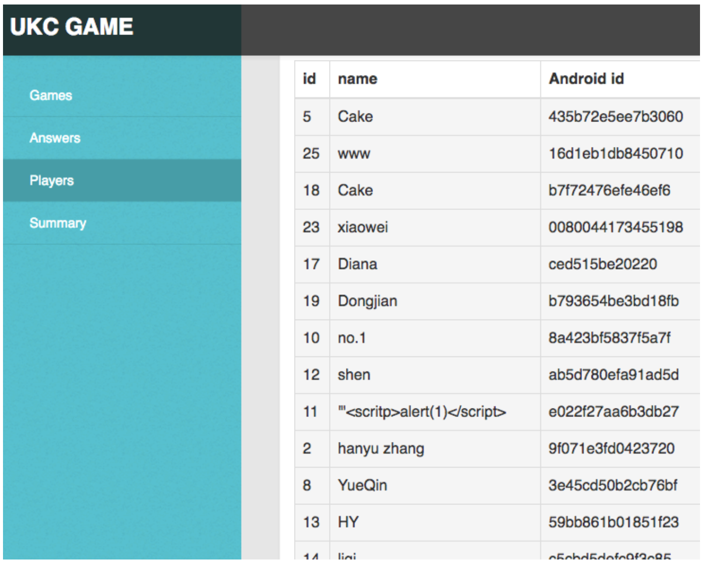
To test our game, we invited some players from both Jilin University[48] and Trento University[49]. 38 players were invited to test our game. Figure 81 is the part of the players view from UKC games framework. For each player we collected their Android system id as the identifier. In these 38 players, out of which 26 of them were active participants, in which 16 were Chinese players and 10 were Italian players. The English level for these Chinese testers are ranged from College English Test (CET) 4 to 6 and the majority of them are Master students. While for Italian

[48] Jilin university website: http://www.jlu.edu.cn/

[49] Trento university website: http://www.unitn.it/en

testers, most of them are PhD students, their English skill level is supposed to be higher than Chinese testers.

For Chinese language, in duration of 10 days, players spent 672 minutes playing the game in total. The average mean playtime for each player is 42 minutes. 241 games were played in total, out of which 148 games were domain based, 49 were difficulty level and remaining 44 games were played with dynamic difficulty level. For Italian language, in duration of 7 days, 150 games were played, in which most of them are domain based as well. The active players spent 234 minutes for playing the game in this week, and the average time for each player is around 24 minutes.



id	name	Android id
5	Cake	435b72e5ee7b3060
25	www	16d1eb1db8450710
18	Cake	b7f72476efe46ef6
23	xiaowei	0080044173455198
17	Diana	ced515be20220
19	Dongjian	b793654be3bd18fb
10	no.1	8a423bf5837f5a7f
12	shen	ab5d780efa91ad5d
11	'"<scritp>alert(1)</script>	e022f27aa6b3db27
2	hanyu zhang	9f071e3fd0423720
8	YueQin	3e45cd50b2cb76bf
13	HY	59bb861b01851f23
14	liqi	c5cbd5defc9f3c85

Figure 81 UKC games framework for players

Figure 82 shows the users' playing pattern for Chinese games. On average, the honesty level for each player ranged from 19 to 88%. Their average accuracy ranged from 19 to 82% and the game played for each player ranged from 0 to 47 games. The evaluation results showing the average honesty level, average accuracy, game played and English competence level can be found elsewhere[50]. Figure 83 is the corresponding users' playing pattern for Italian language. Whereas the average honesty level for each player ranged from 53% to 93%. Accuracy is ranged from 45% to 92%. And the total of games for each player is ranged from 0 to54. Comparably, the average accuracy and honesty level of collected Chinese games are 57% and 62%, which Italian part is much higher, which are 71% and 80% respectively.

The reason for that is not only because the players' performance, but also the reason of language itself. Chinese language is character based, whereas Italian and English are alphabet based, which means that Italian and English are similar sometimes. The similarity of the languages gives the chance that answering without understanding the corresponding English question. For example, question is 'aperitif' in English and answer is 'aperitivo' in Italian, and question 'public' for answer 'publico' are very similar. For some cases, English and Italian are even in the same appearance, for example 'vermouth', which means 'any of several white wines flavored

[50] https://goo.gl/GOCoeK

with aromatic herbs; used as aperitifs or in mixed drinks', are the same for both Italian and English.

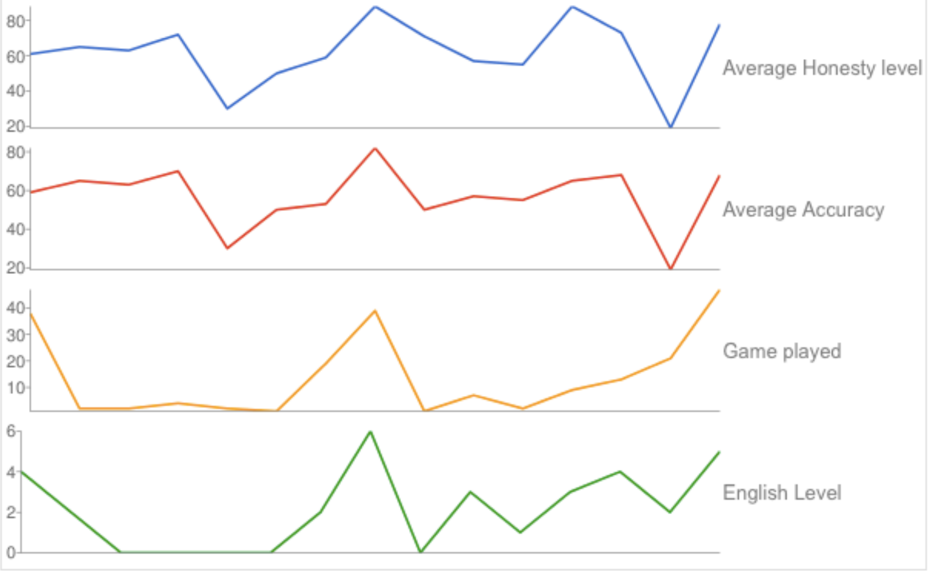


Figure 82 Chinese user results

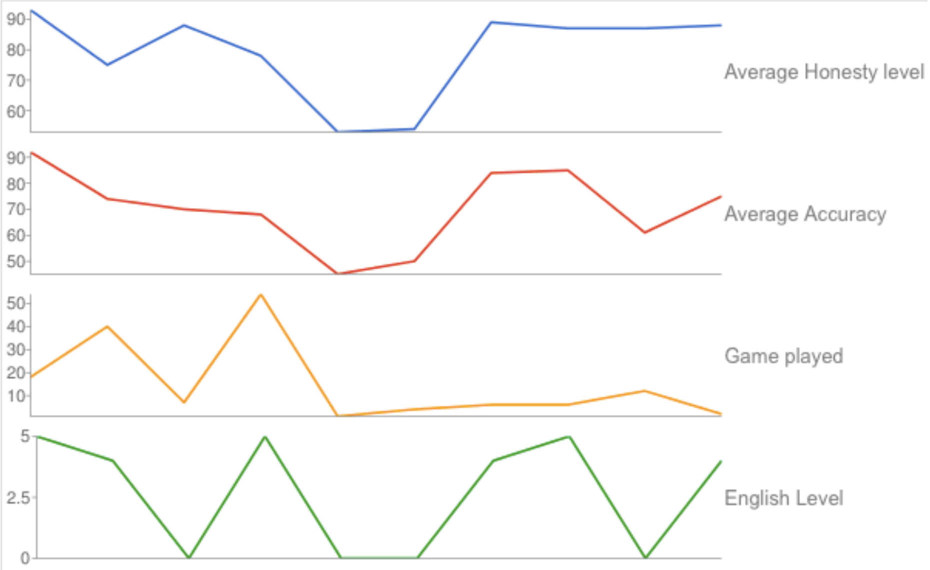


Figure 83 Italian user results

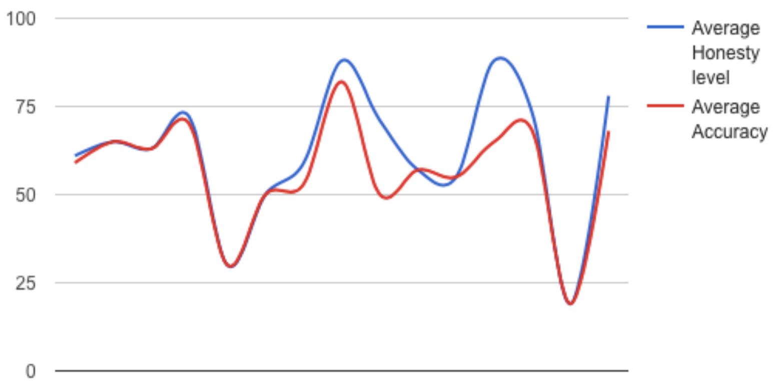


Figure 84 Average honesty level and average accuracy comparison for Chinese players

Average honesty level should be higher or equal to the average accuracy. Unless the situation that the accuracy is around 100%, the average honesty level is higher than the accuracy means that a player answered questions honestly. The difference between honesty level and accuracy indicates how likely a player selecting 'Don't know' option. In the other words, it shows how likely a player guessing a non-familiar or unknown question. When a game's accuracy is very low and its honesty level is the equal to the accuracy, the answer from this game is worthless. Figure 84 illustrates the comparison between the average honesty level and average accuracy with respect to Chinese players. Blue line is average honesty level and red one is the average accuracy. In this case, honesty level and accuracy are overlapping in most part of it, which means that people are more likely to guess instead of clicking 'Don't know' honestly.

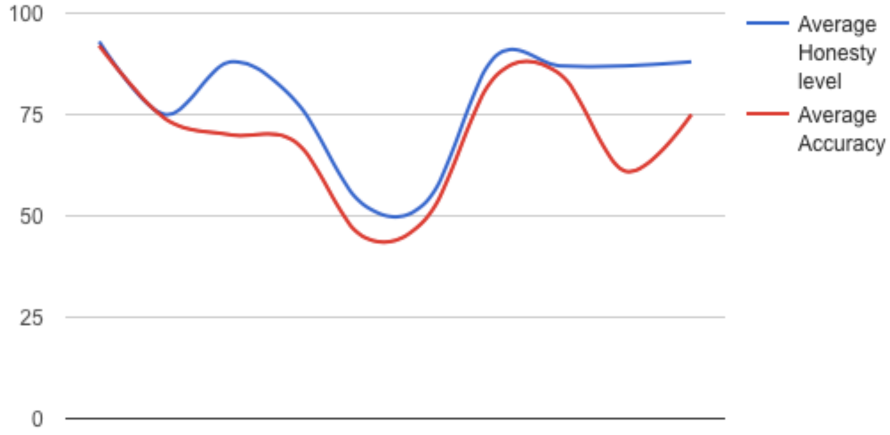


Figure 85 Average honesty level and average accuracy comparison for Italian players

Figure 85 indicates the same comparison for the Italian players. In this case, we can see that average honesty level is higher than average accuracy for the most part, which means that people would like to select 'Don't know' option honestly in most cases.

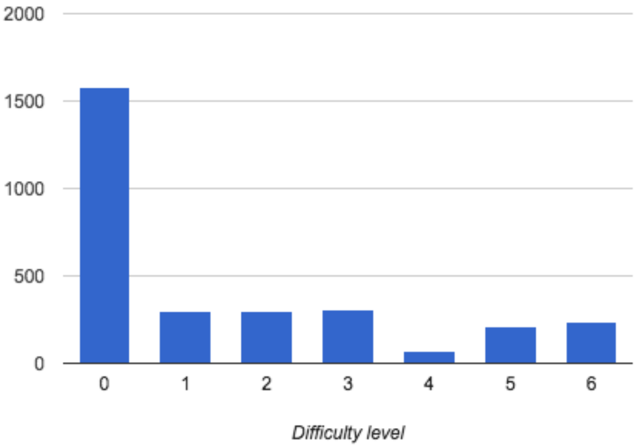


Figure 86 Difficulty level distribution for Chinese collected answers

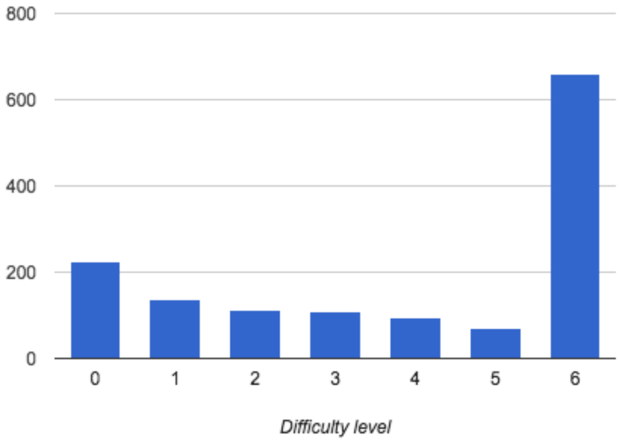


Figure 87 Difficulty level distribution for Italian collected answers

Almost 5000 answers were collected during the evaluation, in which 3000 around are Chinese answers and 1600 around are Italian. Figure 86 and Figure 87 are difficulty level distributions of the collected answers for Chinese and Italian language respectively. From these two figures we can see that the most played games for Chinese concentrated in difficulty level 0 but for Italian is concentrated in difficulty level 6. This basically fulfills the real situation. Also 80% collected answers came for related option, which means related option is the most selected option type. Our game is used to evaluate a players English concept vocabulary, the average result of Chinese players is English level 2 and for Italian players is 4, satisfying the real situation as well.

7.3 Game Result Evaluation

To understand whether the users were finding errors, for the Chinese part, we randomly sampled 140 questions out of the 604 questions set (23.17%). And we manually validated these 140 questions as the gold standard by experts. In these 140 questions, 24 of them are wrong and the rest are correct. Several methods could be helpful for selecting the final results. The easiest one is treating the feedbacks from the game as a plurality single voting system. Finding the result via comparing how many players answered a question as correct or wrong. If all players answered a question as wrong, it might either the question is too hard or this question is negative. Two reasons can cause a negative question, the wrong assumed answer and bad corresponding option sets. While, skip in a voting system can be treated as an abstention action, we do not take it into account. To find the ultimate result in such a voting system, we can simply calculate either relative majority (plurality) or super majority (called absolute majority and qualified majority as well) for correct and wrong. However, there are several possible algorithms that can help us selecting the majority either.

7.3.1 Gold Standard

In order to assess the validation result from the game, at first, we created a gold standard. 140 out of 604 questions were chosen as the gold standard set. A manually validation by three experts was proceeded. Fleiss's kappa was employed to evaluate inter-annotator agreement, resulting in ***k = 0.795***, which achieved the substantial agreement and near ***almost perfect agreement***. The percentage of fully agreement for each question is around 67%. While, after that, experts discussed together in order to further get the full agreement for the entire dataset.

7.3.2 Possible Algorithms to Select The Majority

In the relative majority method, the winner is the one with the largest number of votes received out of the entire group of options. While, in a super majority, the winner is the option with more than a threshold (for example 60%) votes out of entire votes. Table 30 is randomly selected from the game feedbacks. In this table, gloss means the gloss of a concept. Chinese synset and English synset are the related synset that attached to this concept. Question ID indicates the question record belongs to. Correct, wrong and skip are the sum of correct answers, wrong answers and skipped respectively. For instance, for question '58', we have collected 6 answers already (which can be seen as 6 players have answered this question as well), in which 3 players answered wrong and 3 of them selected 'Don't know' options. Since 3 people abstain from voting, the totally vote number for question 58 is 3 and wrong get the majority of this voting. Here, there are still two things we need to take into consideration. A) How many of the total vote number for a question is enough. B) The threshold for super majority vote.

Correct	Wrong	Skip	Gloss	Chinese Synset	English Synset	Question ID
0	5	3	any frame in which a bowler fails to make a strike or spare	未能击倒全部木瓶的一局	break,open frame	58
3	2	1	a sustained effort	不断的努力	pull	2151

9	2	0	several exercises intended to be done in series	锻练设定	set,exercise set	2153
11	0	0	exercise designed to strengthen the arm muscles	臂力锻炼	arm exercise	2171
8	0	0	exercise designed to strengthen the back muscles	背部肌肉锻炼	back exercise	2175
9	0	0	exercise designed to strengthen the leg muscles	腿部锻炼	leg exercise	2176

Table 30 An example of game feedbacks

In these two methods, however, we did not consider the possibility of correctness for each answer. Obviously, the contribution for each answer is different. For example, a player who has several English spoken experiences and a player who just started to learn English, their contribution is different. In normal sense, a senior player should be trusted more. But, there is no evidence that a senior player must have a larger weight than a junior player for all rounds. It happens that a junior player knows the question of a round that a senior player does not know in some specific domains or difficulty levels. By examining this scenario, we need a value that can fairly represent a player's weight. In Concept Challenge Game, each game is a related group of rounds and these rounds are always related. For example, a player play domain selection game mode, the questions are all from the selected domain, and in difficulty selection, all questions are the same difficulty level. Since the relatedness of the questions, we can simply adopt the correctness of a game to indicate how we can trust an answer. The correctness of a game perfectly represents how much a player knows an area of words. But in some case, for example, a player played 10 games, in which 5 of them as correct and 5 of them are selected 'Don't know' option honestly. In this case, his correctness is 50%, but his answer is trustable indeed. To solve this problem we introduced honesty level as in the following. We introduced honesty level as the correct rounds with respect to the answered rounds without skipped rounds.

$$Honesty\ level = \frac{correct}{sum - skipped}$$

We propose two methods by considering the trust of each answer. The first one is simply adding a filter into the two voting system to find out the high quality answers. For example, all answers' game correctness needs to be higher than 50%. The second one is DS evidential theory. DS theory is competent to reason with uncertainty. In order to adopt DS theory, we suppose that the probability of an answer is equivalent to the correctness of the game and the rest probabilities are uniformly distributed. For example, a player selected answer A, and in this game, his correctness is 60%. The probability distribution for this question is in the following:

Option	A	B	C	D
Probability	60%	13%	13%	14%

However, two important points are ignored above, which are the total number of answers and the number of possible answers. We need a measure to determine whether an answer is significantly selected more than the rest. To do this, we can exploit Pearson's chi-square test, which is the most commonly utilized test in chi-squared tests. If we take the distribution of options over the set of all choices, we can say that only those questions for which this distribution significantly differs from a uniform distribution ($p < 0.05$) can be considered providing an acceptable answer.

To understand how many votes for a question is enough to assess the correctness, we use relative majority, which is the simplest method, as the selection method and use 50% game accuracy and 50% honesty level as the thresholds to select candidate answers. 5, 7 and 9 votes are used as the parameter, while, '3' is not considered since fault tolerance is too small. To figure out the best parameter, we use precision, where to compute the accuracy of validated results from the game, and recall, where to calculate the validating ability. In this case, precision is the proportion of the correct validated questions with respect to the total validated questions. And recall is the proportion of the correct validated questions in terms of the entire dataset. The entire dataset is the gold standard dataset where the number of voting is higher than 5.

$$\boldsymbol{precision} = \frac{Correct\ validations}{\text{All validated questions}} \qquad \boldsymbol{recall} = \frac{Correct\ validations}{\text{Entire dataset}}$$

	Precision	Recall	F1 measure
Relative majority (5)	0.8440	**0.8141**	**0.8288**
Relative majority (7)	**0.8727**	0.4247	0.5774
Relative majority (9)	0.8518	0.2035	0.3285

Table 31 Comparison of voting numbers

Table 31 is the evaluation result for each casted vote number. The best precision of voting number is 7, whereas the precision did not keep increasing when voting number increases. Voting number 9 decreases the accuracy instead. That is because the bias of the players, for example, question is an isolate synset 'run' and the assumed answer is '短途旅程', all players are selecting 'No correct answer' or the rest options instead of selecting the assumed answer. In this case, this question will be validated improperly as wrong even though the number of voting is increasing. But, we can see that recall is decreasing significantly from voting number 5 to 7. By considering F1 measure, finally, we chose voting number 5 as our threshold to decide the correctness of a record since it has a outstanding F1 measure vs. the other two thresholds.

	Precision	Recall	F1 measure
Relative majority	0.8440	**0.8141**	**0.8288**
Super majority (60%)	0.8823	0.6757	0.7653
Super majority (75%)	0.8833	0.3698	0.5213
Super majority (100%)	**0.9655**	0.2522	0.3999
Pearson's chi-square	0.9234	0.4623	0.6160
DS evidential	*N*	*N*	*N*

Table 32 Comparison of algorithms

Table 32 shows the comparison of the algorithms mentioned above. For super majority, we use 60% (three-fifths vote), 75% (three fourths) and 100% respectively, which are commonly used thresholds. After evaluation, since DS evidential theory cannot help us calculating reasonable validation results in our setting, its precision and recall are super low. Thus, we put 'N' instead of a precise number. The best precision of these methods is super majority with 100% as the threshold, which means all the cast votes need to fully concentrate on one option. While, since it has this very strong restriction, its recall is very low. Person's chi-square has the similar precision, and the restriction is relatively loose. For example, in a super majority, 'correct' casted 9 votes. To make a decision, the number of voting ‘wrong’ needs to be 0. While, for the Person’s chi-square, to make a decision, the number of voting ‘wrong’ can be 1 or 2 based on the number of ‘correct’ votes. In this scenario, the recall is higher than 100% threshold super majority. Relative majority, super majority with 60% and 75% threshold have the similar precision, but since the last two methods have a relatively strong restriction, their recall is lower. Thus, as the result, because relative majority has the best F1 measure, we are going to use it as our algorithms to validate game result.

7.3.3 Validation Result for Chinese

We set a threshold of the average honesty level and the average accuracy level as 50% respectively and then further used relative majority referring to the number of votes is higher than the rest, which had been proved has the best F1 measure for our game. We use 5 votes as the threshold to decide the correctness where had been provided as the best number. When an assumed answer of a question gets a majority, we consider this Chinese synset of the question as correct and vice versa. By implementing this method, we found 28 errors, in which 12 of them were false negative.

Error Type	Definition	Example	Total
Wrong Translation	An unrelated Chinese translation	job – 假公济私	8
Imperfect mapping	The English synset has this meaning, but not the specified one	snap – 猛咬	3
Not a word	The Chinese synset is not a Chinese word	field game –领域比赛	2
Partially correct	Not all Chinese words in the synset are related	stage,leg –阶段，旅程的一段，舞台	2
Typo	Wrongly written or mispronounced characters	promotion –晋什	1

Table 33 Different kinds of errors

We discovered errors like ‘wrong translation’, ‘imperfect mapping’, ‘not a word’, ‘partially correct’ and ‘typo’ as showed in Table 33. I) ‘Wrong translation’ indicates when the words used in the Chinese synset means something else entirely. For example, word job itself has multiple meaning but none of the meaning can be translated to ‘假公济私’ (practical jobbery). II) ‘Imperfect mappings’, for example, the word ‘snap’ has the meaning of ‘猛咬’ in some English-Chinese dictionaries, but here it should be ‘the act of snapping the fingers; movement of a finger from the tip to the base of the thumb on the same hand’. The most of false positive synsets we found are belonging to imperfect mapping error type. III) ‘Not a word’ is where a synset is mapped to a phrase or a short sentence. IV) ‘Partially correct’, for example ‘stage, leg’ is correctly mapped to ‘阶段，旅程的一段’ but ‘ 舞台’ is not related with this meaning. V) We also found typo errors. It should be ‘晋升’ instead of ‘晋什’.

English synset	Chinese Synset	The required meaning
break	（台球）开球	the opening shot that scatters the balls in billiards or pool
run	短途旅行	a short trip
run	定期旅行	a regular trip
job,task,chore	零活	a specific piece of work required to be done as a duty or for a specific fee
end	边锋	a position on the line of scrimmage

round trip	来回旅程的	a trip to some place and back again

Table 34 False negative Examples

We list some representative false negative examples in Table 34. Generally, we can divide them into 2 types based on the restriction degree of the used majority algorithm. Since we use a relative loose restriction algorithm, some of these false negative can be eliminated by increasing the limitation or further voting. This is the drawback of the loose restriction algorithm. For example, ‘job, task, chore’ vote casts as ‘correct 5 vs. wrong 6’, and ‘round trip’ was ‘correct 3 vs. wrong 4’. By further voting it might become ‘correct 10 vs. wrong 4’, or eliminate by adopting super majority with 60% as the threshold. But some of them cannot be eliminated even tough waiting for the further voting or increasing restriction. For example, ‘run’-‘短途旅行’ and ‘run’-‘长途旅行’ were casted as ‘correct 0 vs. wrong 7’ and ‘correct 2 vs. wrong 9’. They cannot be figure out even though waiting or adding more restrictions.

The second type false negative shows that the players faced difficulty while selecting the most appropriate option from the polysemous word of rarely used senses. In fact, English WordNet is full of such polysemous words. For example, run has 56 senses in WordNet. These senses are too fine classified to recognize. Some senses are even hard to find in English dictionaries. As the above example ‘run’ as ‘a shot trip’ and ‘a regular trip’, Chinese people rarely use run as ‘a short trip’ and ‘a regular trip’ so they thought the answers were incorrect and selected ‘No answer option’. Actually, these two senses do not exist even in some English- Chinese dictionaries; there are many cases like this. In the same situation, some senses for a specific domain are hard to recognize. For example, if a player plays ‘Play with domain’ game mode and selected ‘Sport’ domain to play, he can understand ‘break’ synset in the question is indicating we talk about is ‘the opening shot that scatters the balls in billiards or pool’ and ‘end’ as ‘a position on the line of scrimmage’. Otherwise, it is hard for him to select correctly unless it is commonly used for Chinese people. It was difficult for the players to recognize the correct sense of the word without a gloss for an isolated synset with multiple senses. So, from the player’s perspective, any Chinese synset related to that English synset will be the correct option. This kind of error was hard to find.

Similarly, 80 were correct, in which 5 of them were false positive. Since the number of false positive is relatively small, we list them all in Table 35. Actually, the false positive were made by the same reason we mentioned before, it is hard for players to recognize isolate synset with too fine classified meanings. As in this table, the relevance between Chinese and English synset is strong. For example, ‘death’ to ‘死亡’ is correct in most scenarios. But at here, it is mentioning ‘the act of killing’.

English synset	Chinese synset	The required meaning
death	死亡	the act of killing
activity	活力，活性	any specific behavior
match play	比赛计分法	golf scoring by holes won
long haul	持久	a journey over a long distance
musical chairs	抢座的游戏	a rearrangement that has no practical effect or significance

Table 35 False positives

Notably, when the casted votes for the correct and wrong are equal or similar, it could be a semantic error as well. As shown in Figure 88, it is a part of semantic structure of synset ‘crossing'. When we generate a related option, we use a synset's direct corresponding Chinese parent and Chinese children as the options. In this case, Chinese part of 'crossing' and 'traversal, traverse' is in the same option set. When we check the English-Chinese pairs 'crossing'-'横越，交叉' and 'traversal, traverse'- '横过，横贯' with respect to the meaning of 'travelling across' and 'taking a zigzag path on skis' respectively, they are correct. But when we put them together, the semantic structure should be modified for the Chinese part.

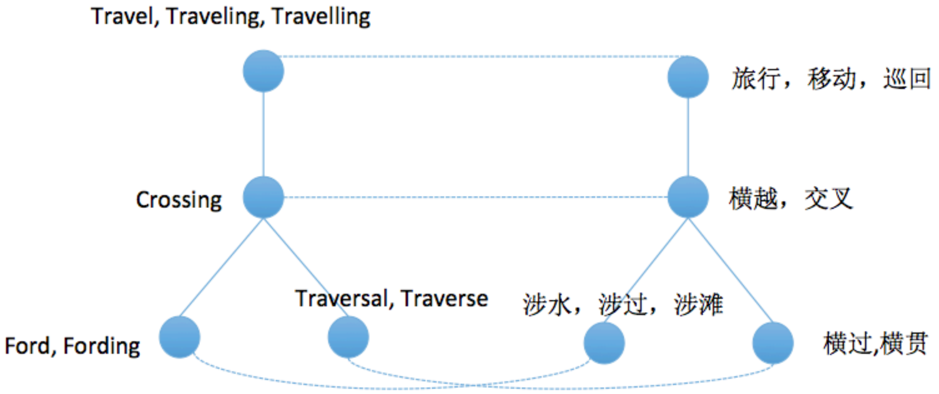


Figure 88 Semantic error example

7.3.4 Validation Result for Italian

We further extended our work to the Italian language as well to evaluate our game is extendable to the other languages. We use the same setting to select possible answers, which are honesty level as 50% and accuracy as 50%. The decision vote number is 5 and the algorithm is relative majority, still. A gold standard with 60 questions was created in which only one was considered as error, which means Italian LKC quality is very high. After evaluation, 57 questions were validated as correct and 3 as wrong, in which 2 of them are false negative. As shown in Table 36, the first one marked with green is the error one, the last two marked with red are false negative. 'potion' with respect to the meaning of 'a medicinal or magical or poisonous beverage' can be translated into 'filtro, pozione'. But 'beveraggio' which means 'beverage' is a larger scope, should not be the part of this synset. These two false negatives because of the same reason discussed above, that is, an isolate synset is hard for players to recognize.

Italian synset	English synset	The required meaning
beveraggio, filtro, pozione	potion	a medicinal or magical or poisonous beverage
coppia	pair	two people considered as a unit
vendemmia	vintage	a season's yield of wine from a vineyard

Table 36 Validation result of errors

The result shows that players spend a significant amount of time playing the game in a short time span. A major constraint for the players was the type of the phone they were using to play the game. Many participants were unable to play as they owned an iPhone, since the system was developed for Android phones. The implementation of the game indicates that the game players with very limited linguistic background can also be involved in finding different kinds of error in a multilingual lexico-semantic resource which can be a major help while building a high quality lexico-semantic resource. The proposed game model can also be extended to support many different languages.

Chapter 8

8 Game Architecture And Implementation

In this chapter we will introduce the architecture and the development of UKC games framework entirely, including data integration framework backend, and a game client named Concept Challenge. In order to guarantee that someone can continue or understand this work conveniently, we provide applied technics for each part of it. Moreover, we will briefly introduce another games framework named Entitypedia Games Framework that we were working for, which is similar to our work. The chapter is structured from a top-down point of view. At first, we introduce the big picture of UKC games framework. After that we will introduce each part in details.

Figure 89 is illustrating UKC games framework from a macroscopic point of view. In this framework, all questions of UKC games framework are transformed automatically from UKC. Currently, UKC contains 109,942 concepts, but not all of them are assigned with all languages. For example, 98,000 of them are mapped with Chinese language and 33,000 of them are mapped with Italian. In this case, UKC games framework can provide around 98k Chinese questions and 33k Italian questions at most. To make word games funnier, we append information of domain and word frequency to UKC games framework, which were retrieved from two independent Databases respectively. Besides, we try to retrieve 4 kinds of options for each question in order to understand the best kind of option and also try to provide more different possibilities of an option set. Moreover, to improve the game performance, we also did a quality check before preparing each question. These three points cause time cost for loading each question extending to around 1 second. While, 1 second for preparing a question is not acceptable. Imagine a situation that a game needs randomly to select 20 Chinese questions, in addition to the time cost for randomly selecting from 98k question pool and time cost of information transfer between server and client, it still needs more than 20 seconds to load the requested questions. To solve this, a cache is utilized between UKC games framework and UKC. We preloaded all questions of a language and then save them in game database, which explicitly decreases the loading time for games. To keep UKC and game database synchronized, we use UKC concept ID, which is a unique ID, as the connection. More information of how we integrated data is in Section 8.1.

UKC games framework was developed by Spring[51] MVC web framework, which is the one of the most popular open source framework for developing enterprise applications, and RESTful web service, which is helpful for organizing a very complex application into simple resources. We utilized hibernate4 + Spring4 + maven3 as the basic development set up, which has been seen as the basic configuration. We discuses this in details is in Section 8.2

A game named Concept Challenge is developed on Android platform. While, it can be seen as an interface merely. All calculation works are processing on the server part. We will talk Android part in Section 8.3.

[51] https://projects.spring.io/spring-framework/

As the extension and to keep in sync with Entitypedia game framework, which is our pervious work, we reuse some APIs of Entitypedia games framework. APIs are used as the medium to interact between UKC games framework and Entitypedia games framework.

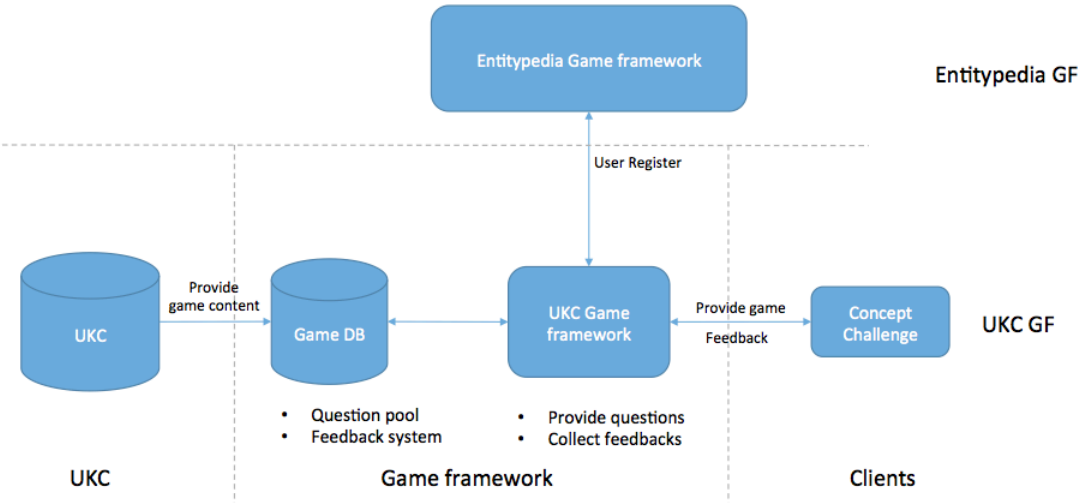


Figure 89 UKC Game Framework architecture

8.1 Data Integration

To provide game-like data and decrease the server loading time significantly, at the beginning, we integrate 3 independent databases and save the integrated data in the game database as a cache. The basic data were imported from UKC. Domain information was imported from WordNet Domains. And word frequency data was extracted from British National Corpus. In the following, we will introduce each of them briefly.

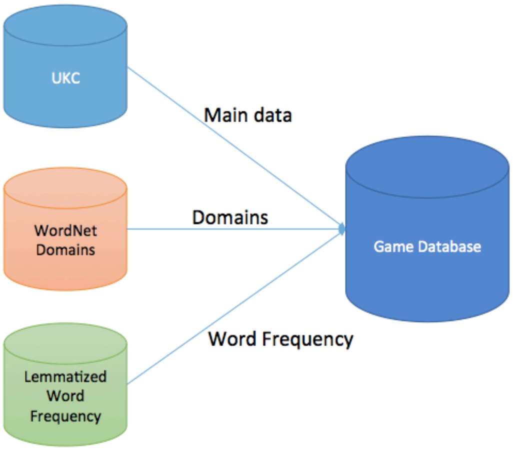


Figure 90 Data integration

8.1.1 WordNet Domains

Domains of UKC GF are imported from WordNet domains developed by FBK. More information in terms of WordNet domains is introduced in Chapter 4.4. This part code is related to project *UKC-domains*.

8.1.2 Word Frequency

We adopted the lemmatised frequency list from this link[52], which contains 6318 lemmatised words with more than 800 occurrences, to compute the synset frequency. Lemmatised words means that all inflected forms had been converted back to its original form. This work was extracted from British National Corpus (BNC)[53] where is a large size (100M word) and large spoken component. The list is arranged as the format of sort-order, frequency, word, word-class. For example,

```
   5 2186369 a det
2107 4249 abandon v
 5204 1110 abbey n
```

8.1.3 Data Structure

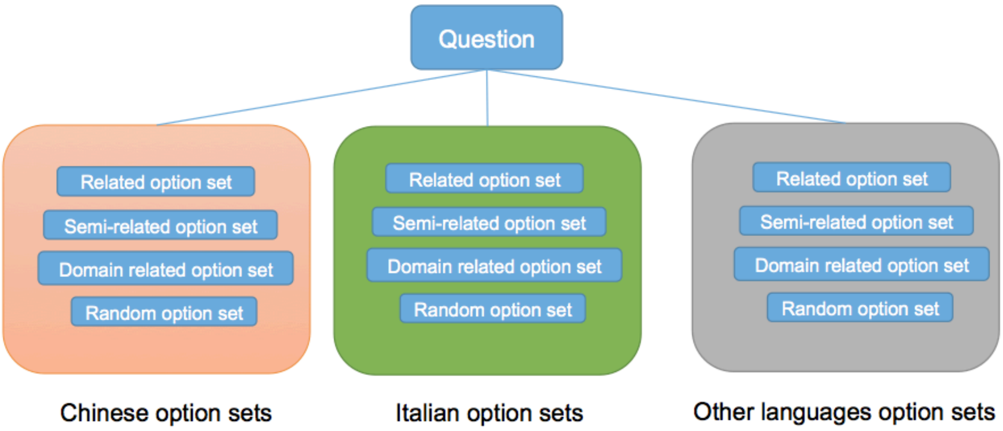


Figure 91 Question Structure

In UKC games framework, a question is an English synset in UKC. A question is designed to connect with multiple-language option sets. Till now, Chinese and Italian option sets are imported. A new language option set is easy to import by identifying UKC language code, e.g. 'zh', in *UKC game-framework-importer*. Cast time is depending on the language size and machine performance. For example, Chinese language has around 95,000 synsets, it costs around 20 hours on a 4 GB RAM computer.

Because synsets between different languages are not one to one mapping, some questions are connected with only Chinese and some of them are only connected Italian option set. English synsets connected with neither Chinese nor Italian option set are not recorded in the game framework. That is, a question in games framework has at least one language option set.

[52] http://www.kilgarriff.co.uk/bnc-readme.html

[53] http://www.natcorp.ox.ac.uk

In a language, each kind of option set has 5 candidate options composed by 1 correct option and 4 wrong options. A question and its correct answer is a language pair connected to one concept in UKC. The other 4 options are randomly selected based on requirements. Each question is supposed to have 4 types of options, which are related options, semi-related options, domain related options and random options, arranged from hard to easy. For example, for a synset 'school, schoolhouse', its related options are random selected from a Chinese synset set, composed by Chinese synset corresponding to "building, edifice", which is the hypernym of synset "school, schoolhouse", "conservatory, conservatoire" and "day school" which are its hyponym, and "skyscraper", "telecom hotel, telco building", "theater, theatre, house", etc., which are the siblings.

Sometimes, because the related synsets of a synset are less than 5, cannot provide 5 Chinese candidate synsets as options. They did not provide related option set. For example, leaves of a tree structure and with few siblings connected. "Art school" is such a case, it is a leaf of the current tree, and only has one sibling "music school". Because of the same reason, some questions are not provided semi-related option set. Basically, a question has a domain-related option set, unless the synset of this question has no domain. A question is always connected with a random option set.

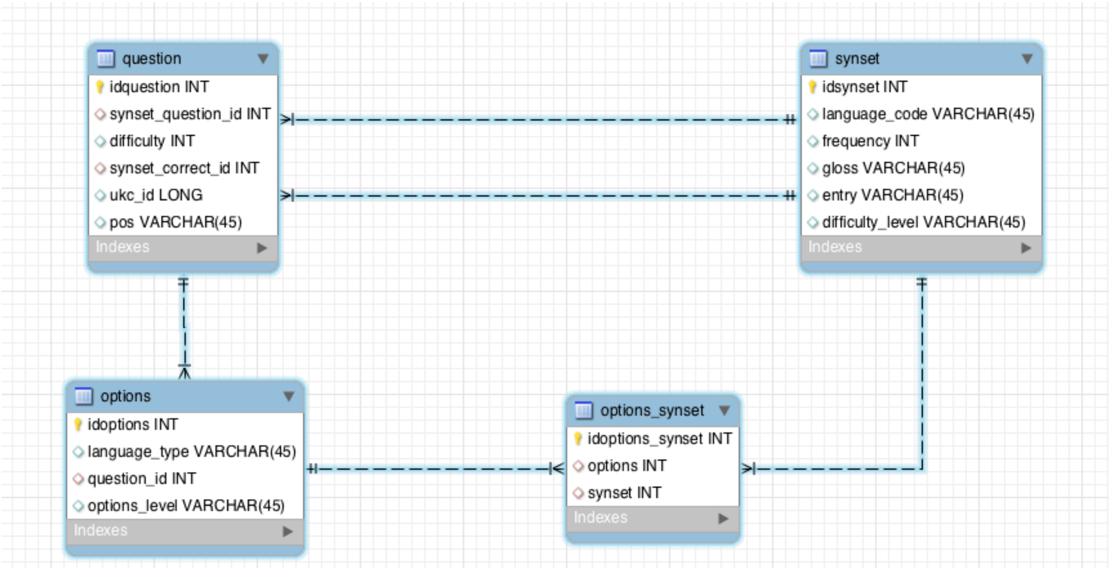


Figure 92 Question Database structure

Figure 92 is a part of game framework database showing how questions are saved in database. The fundamental unit is synset. Both question itself and its mapped correct answer are saving in the synset table. Each question is connected with multiple options mapping to synset table also.

8.1.4 Questions Generation

We need 3 steps to generate questions and its option sets, as in Figure 93. To prepare our game database, we need to parse all concepts in UKC. In general, it is in three steps. At first, since not all concepts can be transferred into a game format, we need to figure out all possible concepts by some checks. For instance, we need to check whether a UKC ID is empty. And also, we need the correct answer from the other languages. Thus, checking whether a concept is connecting with the target language is necessary.

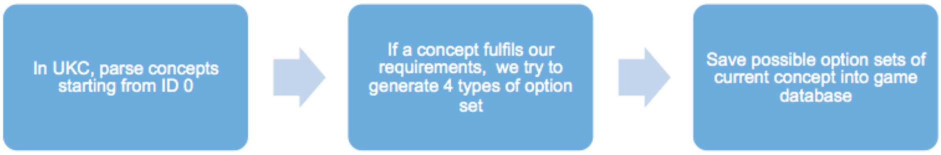

In UKC, parse concepts starting from ID 0
If a concept fulfils our requirements, we try to generate 4 types of option set
Save possible option sets of current concept into game database

Figure 93 Main steps of question generation

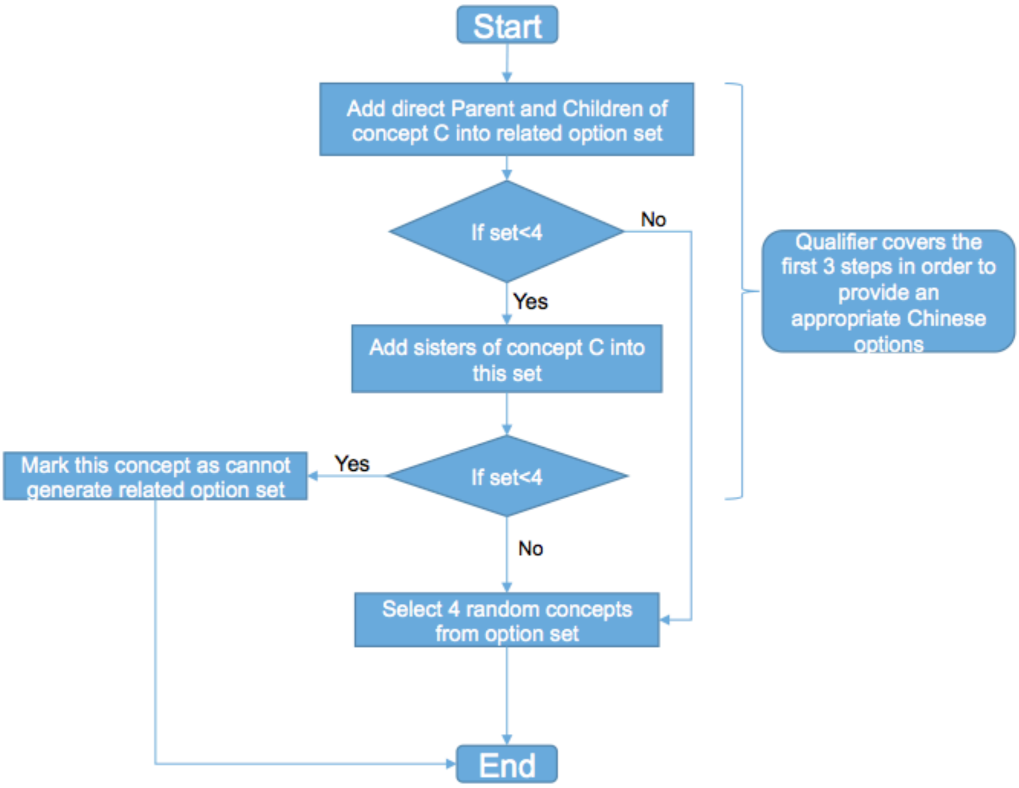

Start
Add direct Parent and Children of concept C into related option set
If set<4
No
Yes
Add sisters of concept C into this set
Mark this concept as cannot generate related option set
Yes
If set<4
No
Select 4 random concepts from option set
End
Qualifier covers the first 3 steps in order to provide an appropriate Chinese options

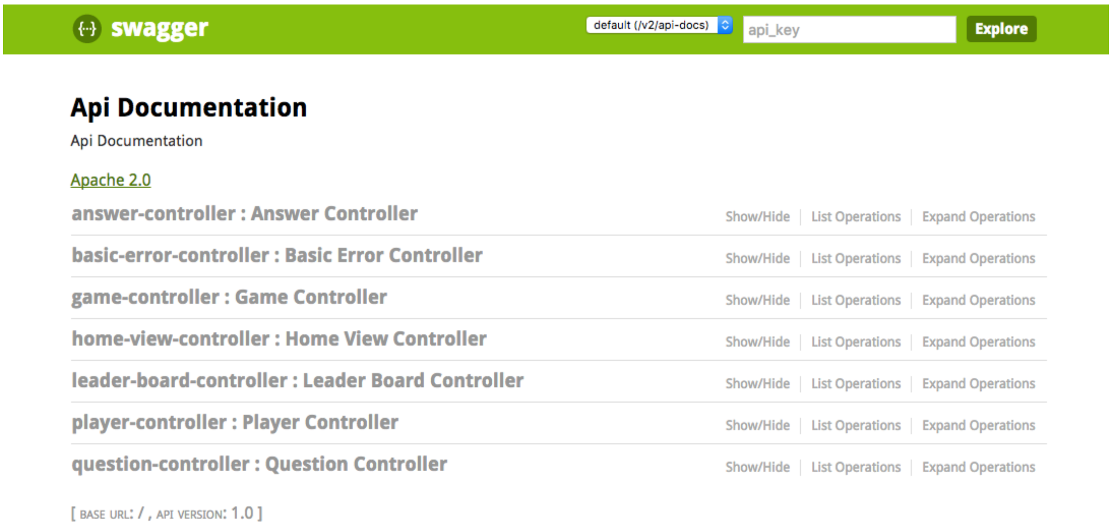
Figure 94 Related option set generation

After we found all possible concepts, we need to generate 4 types of option set for each of them. The procedure of generating a related option set for concept C is showed in Figure 94. For a concept, if we cannot generate its related option set, we marked a negative state for its related option set. All target language synsets need to pass a qualifier in order to prevent bad options. For example, duplicated options when Chinese as target language. It happens that two similar English synsets translated into two identical Chinese synsets where we need to remove. Otherwise, duplicated options will be in one option set. Also, some bad formats like symbol abuse, space abuse, etc. need to be removed. Since different language resource has different problems, new conditions can be conveniently added by overwriting the qualifier method. The rest kinds of option set are generated by the similar method with modifying the selection area of the candidates.

8.2 Games Framework Development

UKC games framework was developed by the popular development pattern ‘Maven + Spring + Hibernate’ and strictly developed in accordance with Model + View + Controller Spring design pattern. Game database is using PostgreSQL 9.5. UKC game framework was deployed on

DreamOne[54] server of the University of Trento. UKC game framework backend is an interface displaying collected data and required information, including users, collected games and collected answers, etc. It was developed by Node.js and transfer data via APIs from the games framework.



Api Documentation

Api Documentation

Apache 2.0

answer-controller : Answer Controller — Show/Hide | List Operations | Expand Operations

basic-error-controller : Basic Error Controller — Show/Hide | List Operations | Expand Operations

game-controller : Game Controller — Show/Hide | List Operations | Expand Operations

home-view-controller : Home View Controller — Show/Hide | List Operations | Expand Operations

leader-board-controller : Leader Board Controller — Show/Hide | List Operations | Expand Operations

player-controller : Player Controller — Show/Hide | List Operations | Expand Operations

question-controller : Question Controller — Show/Hide | List Operations | Expand Operations

[BASE URL: / , API VERSION: 1.0]

Figure 95 Swagger 2.0 interface

All functions of the framework are provided via REST service managing by Swagger 2.0. The Framework Web API adheres to the following common conventions:

- Interactive API documentation is available at link[55].
- Application Root: all URLs start with a common prefix, application root, for the moment: http://dreamone.disi.unitn.it:8090/;
- HTTP Verbs: read only API calls use HTTP GET request, change API calls use HTTP POST request;
- JSON for objects: simple parameters (simple types like date or number) go via request prameters, more complex objects (like Player) go via request body encoded in JSON format.
- URL scheme: API and Web URLs follow the scheme:
  /{object-name}/{verb}?{params}
  For read methods "read" is omitted and usually object ID follows.
- HTTP response status codes follow the protocol convention.

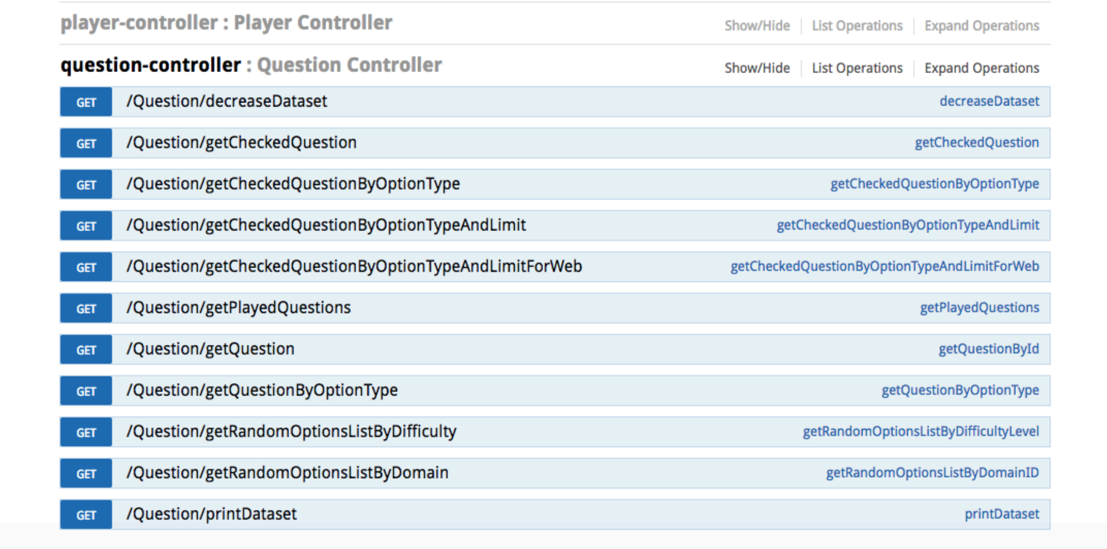
Figure 95 is the screenshot for Swagger interface. It is an interactive interface where a user can test and read all APIs. Functions are divided into 7 categories. A developer can browse all included APIs by unfolding the category as shown in Figure 96. By one more step clicking, the details of an API, including required model schema, parameter, parameter data type, etc., are displayed as shown in Figure 97. In this example, API **/Question/getRandomOptionsListByDifficult** is a GET service, designed to request random questions from the framework via difficulty levels. It has 4 parameters,

[54] https://dreamone.disi.unitn.it:8090

[55] http://dreamone.disi.unitn.it:8090/swagger-ui.html

- Id means the required difficulty level.
- Type means required option set type, related option is 0 and random option set is 3.
- Number means the quantity of required questions. For instance, 10, 20 or 30.
- Language Type is used to specify the requiring language. Till now, ‘zh’ and ‘it’ are included.

A developer can conveniently test this API by click ‘Try it out!’ button. The response is shown in Figure 98. The response body is the requested randomly selected questions in Json format. In case of error, response headers will notify the error details.

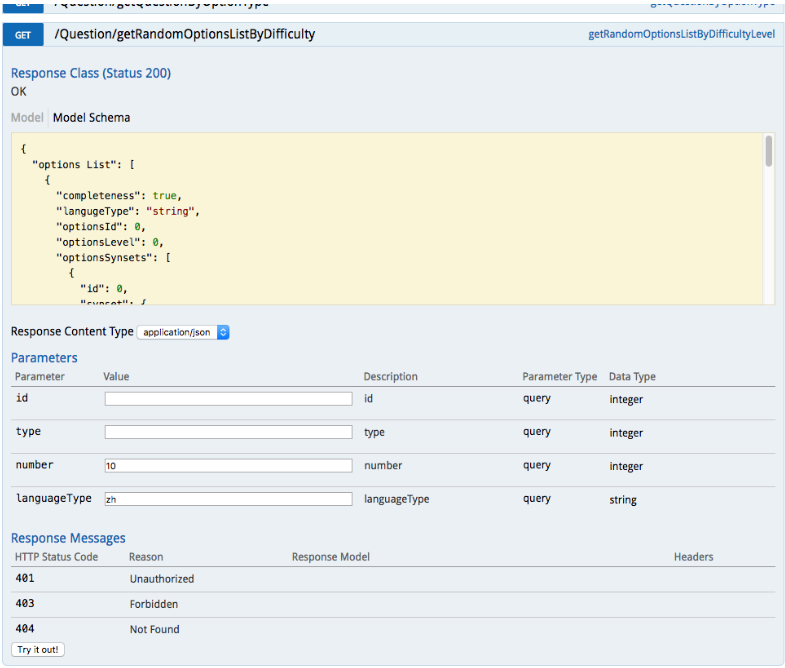


player-controller : Player Controller — Show/Hide | List Operations | Expand Operations

question-controller : Question Controller — Show/Hide | List Operations | Expand Operations

GET	/Question/decreaseDataset	decreaseDataset
GET	/Question/getCheckedQuestion	getCheckedQuestion
GET	/Question/getCheckedQuestionByOptionType	getCheckedQuestionByOptionType
GET	/Question/getCheckedQuestionByOptionTypeAndLimit	getCheckedQuestionByOptionTypeAndLimit
GET	/Question/getCheckedQuestionByOptionTypeAndLimitForWeb	getCheckedQuestionByOptionTypeAndLimitForWeb
GET	/Question/getPlayedQuestions	getPlayedQuestions
GET	/Question/getQuestion	getQuestionById
GET	/Question/getQuestionByOptionType	getQuestionByOptionType
GET	/Question/getRandomOptionsListByDifficulty	getRandomOptionsListByDifficultyLevel
GET	/Question/getRandomOptionsListByDomain	getRandomOptionsListByDomainID
GET	/Question/printDataset	printDataset

Figure 96 unfolded an API category



GET /Question/getRandomOptionsListByDifficulty — getRandomOptionsListByDifficultyLevel

Response Class (Status 200)

OK

Model | Model Schema

```
{
  "options List": [
    {
      "completeness": true,
      "langugeType": "string",
      "optionsId": 0,
      "optionsLevel": 0,
      "optionsSynsets": [
        {
          "id": 0,
```

Response Content Type: application/json

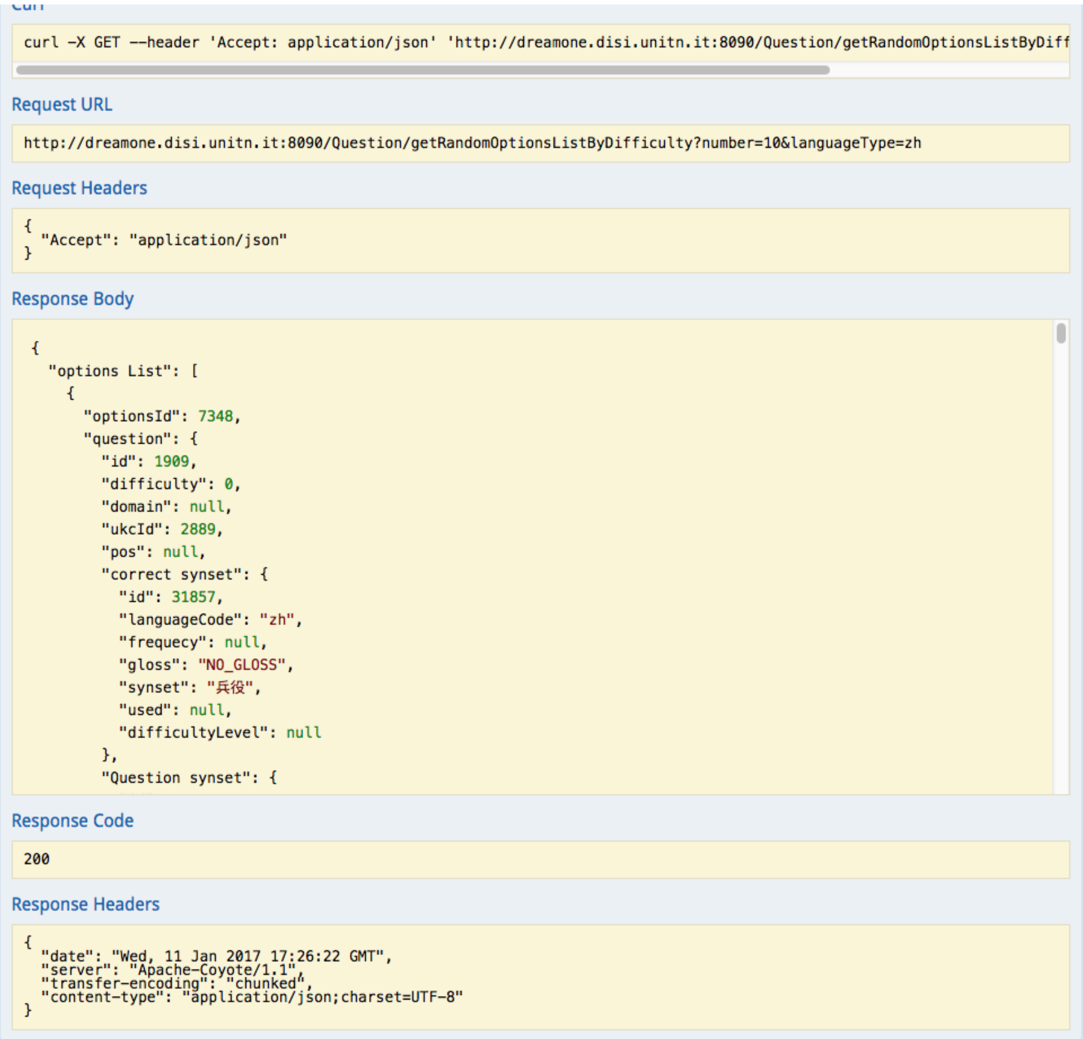
Parameters

Parameter	Value	Description	Parameter Type	Data Type
id		id	query	integer
type		type	query	integer
number	10	number	query	integer
languageType	zh	languageType	query	string

Response Messages

HTTP Status Code	Reason	Response Model	Headers
401	Unauthorized		
403	Forbidden		
404	Not Found		

Figure 97Details of an API (1)



```
curl -X GET --header 'Accept: application/json' 'http://dreamone.disi.unitn.it:8090/Question/getRandomOptionsListByDiff
```

Request URL

```
http://dreamone.disi.unitn.it:8090/Question/getRandomOptionsListByDifficulty?number=10&languageType=zh
```

Request Headers

```
{
  "Accept": "application/json"
}
```

Response Body

```
{
  "options List": [
    {
      "optionsId": 7348,
      "question": {
        "id": 1909,
        "difficulty": 0,
        "domain": null,
        "ukcId": 2889,
        "pos": null,
        "correct synset": {
          "id": 31857,
          "languageCode": "zh",
          "frequecy": null,
          "gloss": "NO_GLOSS",
          "synset": "兵役",
          "used": null,
          "difficultyLevel": null
        },
        "Question synset": {
```

Response Code

```
200
```

Response Headers

```
{
  "date": "Wed, 11 Jan 2017 17:26:22 GMT",
  "server": "Apache-Coyote/1.1",
  "transfer-encoding": "chunked",
  "content-type": "application/json;charset=UTF-8"
}
```

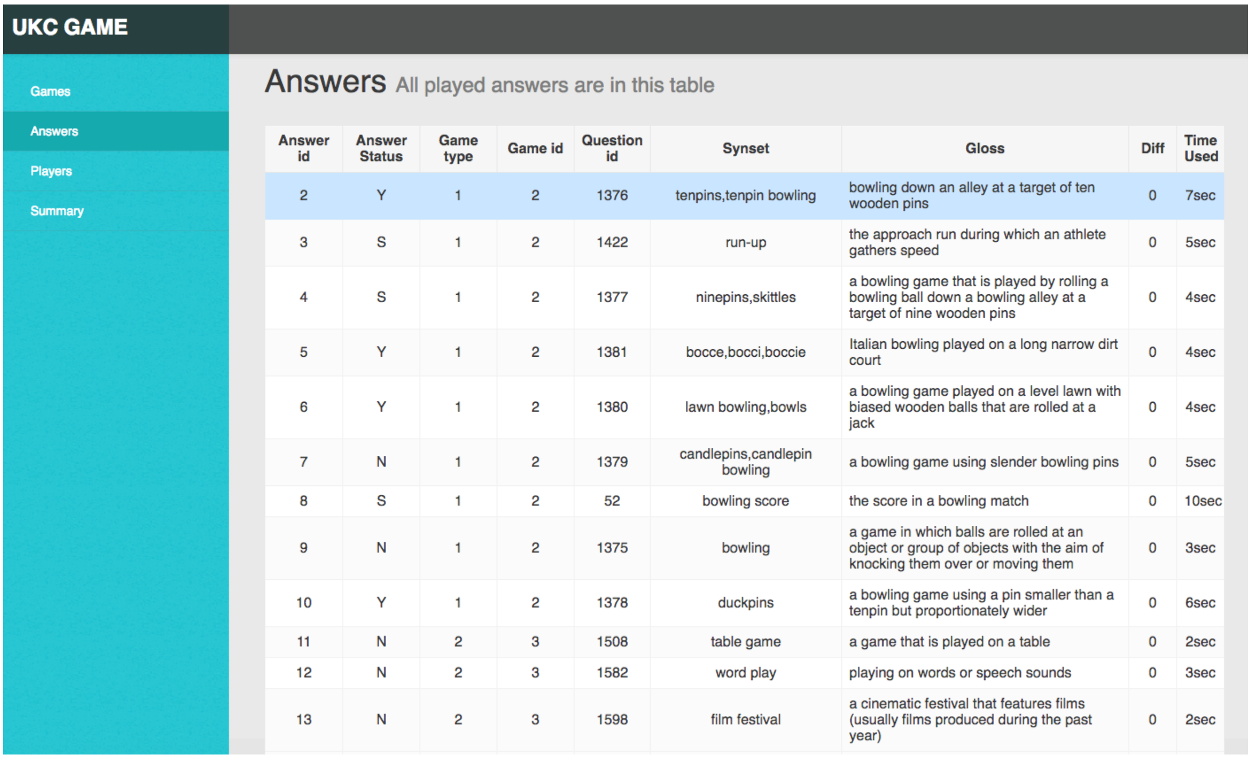
Figure 98 Details of a API(2)

As we mentioned above, we have 7 API categories. In the following, we briefly introduced one by one in the following.

- Answer controller is utilized to display collected answers from games. For example, get all answers of a player or display all answers.
- Basic error controller is used to manage errors.
- Game-controller is used to collect feedbacks from game client. This game indicates a played game from a game client, for instance, Word Challenge Game. A game contains 10-30 rounds. A game client can post a new played game back to the server. Also, it provides functions like compute correctness of a player, return all played games of a player, etc.
- Home view controller is mapping to the main web page.
- Leader board controller is used to provide APIs associated with game leaderboards.
- Player controller is providing functions related to game players.
- Question controller provides the game content. For example, get 10 random rounds for a specific domain; get 20 random rounds from a difficulty level, etc.

Notably, even though we use the game database as a cache to reduce time consuming, due to the size of game database, randomly selecting from more than 10,000 records is still time consuming.

We did an approximate evaluation that the average time cost for selecting a 30 rounds game from a 55,000 records randomly is 12000 milliseconds. We think let players wait around 12 seconds is too long, thus, in the framework we use fake random algorithm instead. In this fake random algorithm, we only random 1 question as a seed, then select a number of questions around this seed. This method costs around 6000 milliseconds for a 30 rounds game. But a disadvantage is that, the randomly selected result belongs to the similar domains in a big probability.



Answers All played answers are in this table

Answer id	Answer Status	Game type	Game id	Question id	Synset	Gloss	Diff	Time Used
2	Y	1	2	1376	tenpins,tenpin bowling	bowling down an alley at a target of ten wooden pins	0	7sec
3	S	1	2	1422	run-up	the approach run during which an athlete gathers speed	0	5sec
4	S	1	2	1377	ninepins,skittles	a bowling game that is played by rolling a bowling ball down a bowling alley at a target of nine wooden pins	0	4sec
5	Y	1	2	1381	bocce,bocci,boccie	Italian bowling played on a long narrow dirt court	0	4sec
6	Y	1	2	1380	lawn bowling,bowls	a bowling game played on a level lawn with biased wooden balls that are rolled at a jack	0	4sec
7	N	1	2	1379	candlepins,candlepin bowling	a bowling game using slender bowling pins	0	5sec
8	S	1	2	52	bowling score	the score in a bowling match	0	10sec
9	N	1	2	1375	bowling	a game in which balls are rolled at an object or group of objects with the aim of knocking them over or moving them	0	3sec
10	Y	1	2	1378	duckpins	a bowling game using a pin smaller than a tenpin but proportionately wider	0	6sec
11	N	2	3	1508	table game	a game that is played on a table	0	2sec
12	N	2	3	1582	word play	playing on words or speech sounds	0	3sec
13	N	2	3	1598	film festival	a cinematic festival that features films (usually films produced during the past year)	0	2sec

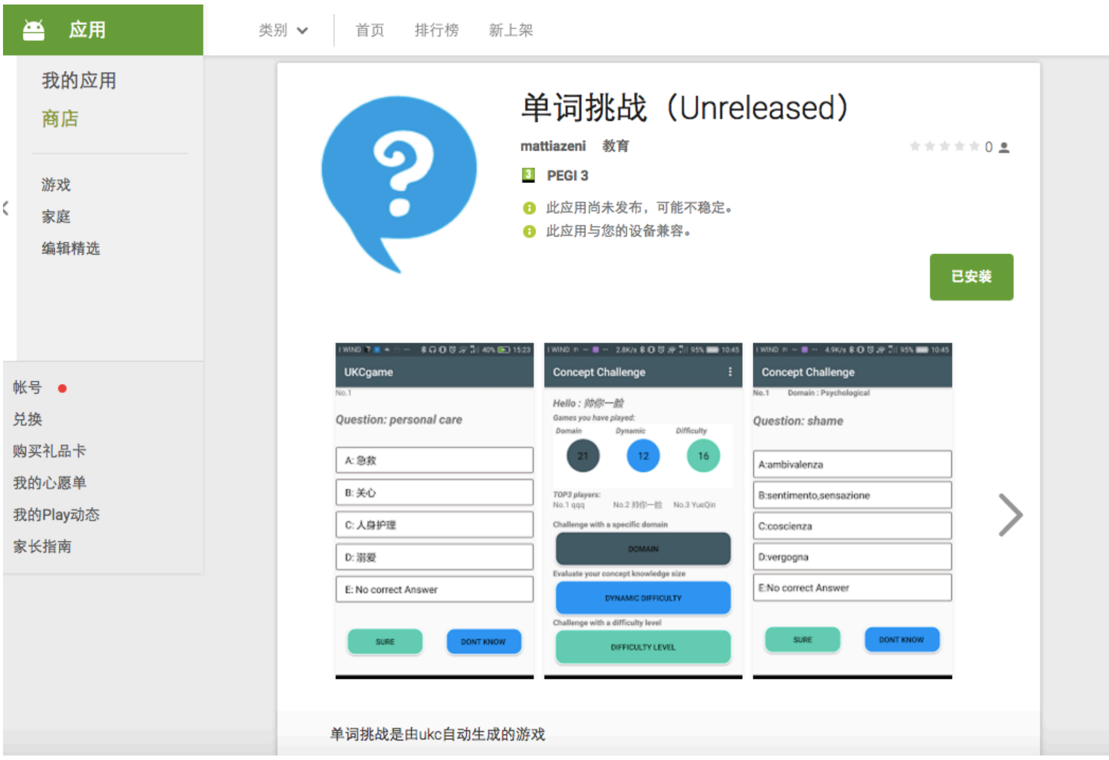
Figure 99 UKC Game Framework backend screenshot

Neither Json files nor database records are convenient for human being to read, we developed a UKC GF backend where retrieving data via APIs to make it human readable. As showed in Figure 99, it has 4 tabs to view the related information.

8.3 Concept Challenge Game

An Android client named Concept Challenge is developed in order to test our idea. The mininum software development kit (SDK) version is 14 and target SDK version is 24. The game has been uploaded to Google Play[56] with respect to Chinese, English and Italian languages. Players are allowed to download and play it without any payment.

[56] https://play.google.com/store/apps



单词挑战（Unreleased）

mattiazeni 教育

PEGI 3

此应用尚未发布，可能不稳定。

此应用与您的设备兼容。

UKCgame

No.1

Question: personal care

A: 急救

B: 关心

C: 人身护理

D: 溺爱

E: No correct Answer

Concept Challenge

Hello : 帅你一脸

Games you have played:

Domain	Dynamic	Difficulty
21	12	16

TOP3 players:

No.1 qqq No.2 帅你一脸 No.3 YueQin

Challenge with a specific domain

Evaluate your concept knowledge size

Challenge with a difficulty level

Concept Challenge

No.1 Domain : Psychological

Question: shame

A:ambivalenza

B:sentimento,sensazione

C:coscienza

D:vergogna

E:No correct Answer

单词挑战是由ukc自动生成的游戏

Figure 100 Screenshot from Anroid Play of Concept Challenge Game

The application was developed via Android Studio, which is a newly released Android integrated development environment (IDE). In addition to Android project structure, to provide a more readable and reusable programming code, the development is strictly followed View, Core, API and Model structure mapping to view, core, api and model package in java code respectively. As shown in Figure 101, each component has its own responsibility. The explanation of each component is in the following.

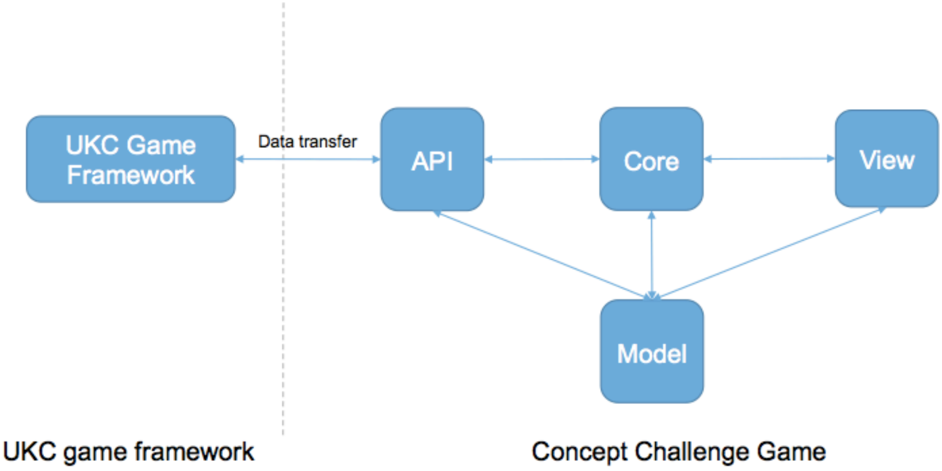


Figure 101 Android application structure

API component: All codes in terms of web transfer belong to API layer. This component has two assignments: the first one is receive/send data to the game framework. Since Json is used as communication data format for the REST services but Java model is required in the application, the second task is parse/prepare the json file, all Json files are parsed into Java models and vice versa. There are several options to implement data transfer and Json parser. For data transfer we utilized Spring for Android[57] project as our solution due to the game framework was developed via Spring framework, which has a better compatibility. The other options are Android Http Client and Android Asynchronous Http Client, but they are not as convenient as Spring one when working with Spring framework. We utilized Gson as our parser in this application, which is a mature parser for Android platform. Furthermore, Gson is more accurate, flexible and extensible comparing with the other parsers. During parsing, we can modify the target Java model object easily.

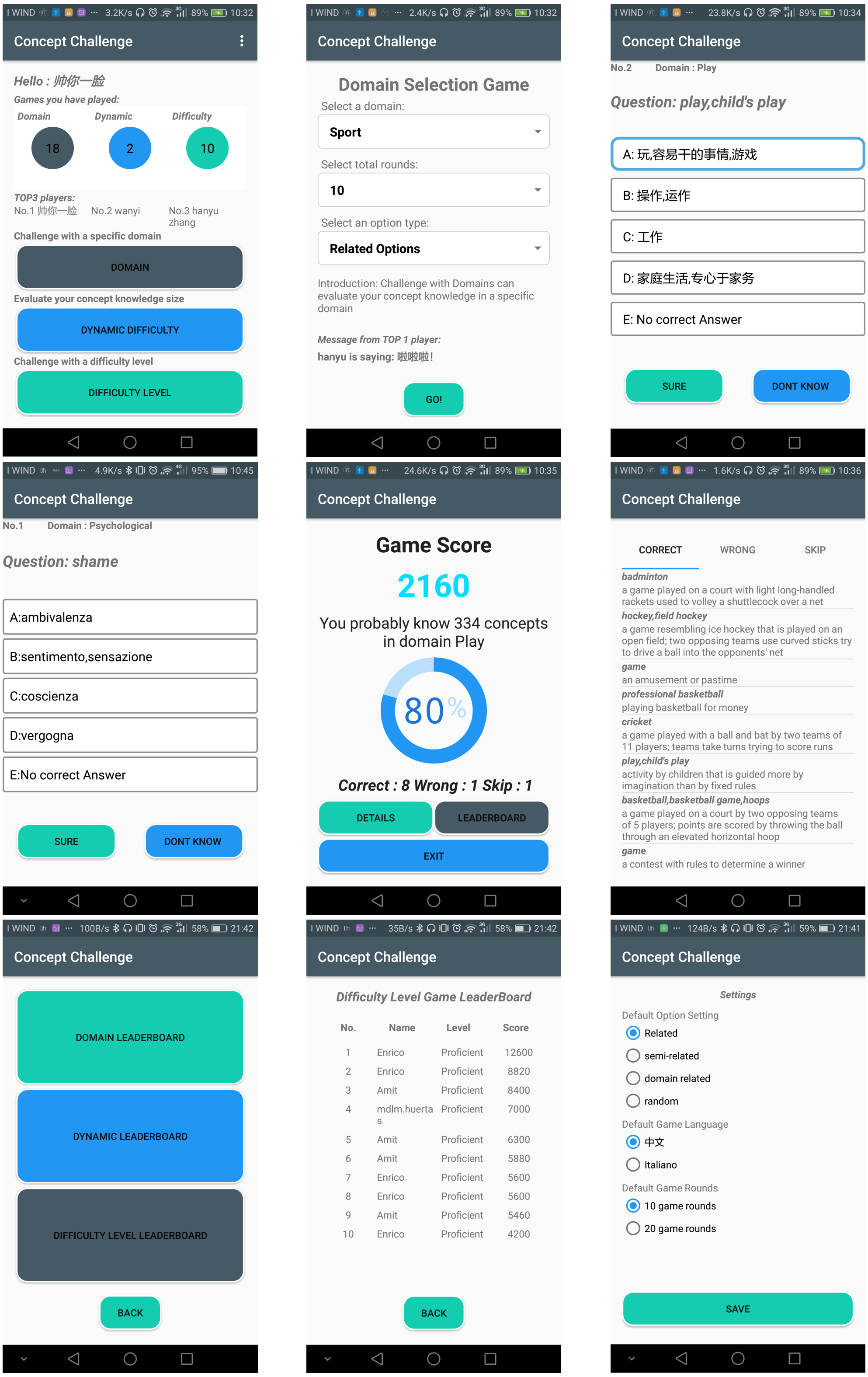
Model component: Data models are saved in this component. Data model inherits project ukc-game framework-common. When develop a new game, a developer can add dependence of this project via Maven or Gradle and use this model in a parser.

Core component: All computations and functionalities are settled in this component. For example, compute game results, prepare game feedback where need to send back to the game framework, etc. Views access this component in order to get content to show on screen.

View component: View component controls all Android activities including human computer interaction, logic processing between activities, screen layout and so on. This component has no core computation.

Figure 102 is the screenshots of Concept Challenge Game. This game has 3 elementary game functionalities. Play with a domain, play with a difficulty level and play with dynamic difficulty levels. In this figure, we use ‘play with a domain’ as the example. After a player choosing this functionality, he has ability to choose with which domain, how many rounds and what kind of option type he wants to play. A favourite pattern can be set as default via game settings (as the last screenshot in Figure 102). After finishing all questions, a summary will be provided, including game score, correctness and the controlled knowledge in the current domain. For different game functionalities, the game result is different. A list of correct, wrong and skip rounds are provided, associating with English glosses in order to help player learning the played game. Leader Boards are available for all game functionalities. When a player gets a highest score on the leader board, he can leave a message to the other players.

[57] http://projects.spring.io/spring-android/



Concept Challenge

Hello : 帅你一脸

Games you have played:

Domain	*Dynamic*	*Difficulty*
18	2	10

TOP3 players:

No.1 帅你一脸 No.2 wanyi No.3 hanyu zhang

Challenge with a specific domain

DOMAIN

Evaluate your concept knowledge size

DYNAMIC DIFFICULTY

Challenge with a difficulty level

DIFFICULTY LEVEL

Concept Challenge

Domain Selection Game

Select a domain: Sport

Select total rounds: 10

Select an option type: Related Options

Introduction: Challenge with Domains can evaluate your concept knowledge in a specific domain

Message from TOP 1 player:

hanyu is saying: 啦啦啦!

GO!

Concept Challenge

No.2 Domain : Play

Question: play,child's play

A: 玩,容易干的事情,游戏

B: 操作,运作

C: 工作

D: 家庭生活,专心于家务

E: No correct Answer

SURE DONT KNOW

Concept Challenge

No.1 Domain : Psychological

Question: shame

A:ambivalenza

B:sentimento,sensazione

C:coscienza

D:vergogna

E:No correct Answer

SURE DONT KNOW

Concept Challenge

Game Score

2160

You probably know 334 concepts in domain Play

80%

Correct : 8 Wrong : 1 Skip : 1

DETAILS LEADERBOARD EXIT

Concept Challenge

CORRECT WRONG SKIP

badminton
a game played on a court with light long-handled rackets used to volley a shuttlecock over a net

hockey,field hockey
a game resembling ice hockey that is played on an open field; two opposing teams use curved sticks try to drive a ball into the opponents' net

game
an amusement or pastime

professional basketball
playing basketball for money

cricket
a game played with a ball and bat by two teams of 11 players; teams take turns trying to score runs

play,child's play
activity by children that is guided more by imagination than by fixed rules

basketball,basketball game,hoops
a game played on a court by two opposing teams of 5 players; points are scored by throwing the ball through an elevated horizontal hoop

game
a contest with rules to determine a winner

Concept Challenge

DOMAIN LEADERBOARD

DYNAMIC LEADERBOARD

DIFFICULTY LEVEL LEADERBOARD

BACK

Concept Challenge

Difficulty Level Game LeaderBoard

No.	Name	Level	Score
1	Enrico	Proficient	12600
2	Enrico	Proficient	8820
3	Amit	Proficient	8400
4	mdlm.huertas	Proficient	7000
5	Amit	Proficient	6300
6	Amit	Proficient	5880
7	Enrico	Proficient	5600
8	Enrico	Proficient	5600
9	Amit	Proficient	5460
10	Enrico	Proficient	4200

BACK

Concept Challenge

Settings

Default Option Setting

● Related
○ semi-related
○ domain related
○ random

Default Game Language

● 中文
○ Italiano

Default Game Rounds

● 10 game rounds
○ 20 game rounds

SAVE

Figure 102 Screenshots of Concept Challenge Game

Chapter 9

9 Conclusion

In this work, firstly, we bootstrapped UKC by importing a Chinese linguistic resource named Chinese WordNet, which developed by Southeast University, China. Before importing, we did a comprehensive investigation of Chinese semantic-linguistic resources to select an appropriate resource to import. There are 3 popular Chinese multilingual semantic-linguistic resources, which are HowNet, Chinese Concept Dictionary (CCD) and the imported Chinese WordNet (CWN). By considering quality, cost and importable, HowNet and CCD are not suitable to import since even though HowNet was built manually and has a very high quality, it is a net structure instead of WordNet-like and CCD is WordNet-like structure but its coverage is relative low. Furthermore, both of them are commercial produces, expensive to use even by education purpose. By consider this scenario, CWN was imported.

Secondly, in order to understand CWN in details, we did a study of CWN including construction process and quality. CWN was built by three automatic algorithms, which are Minimum Distance (MDA), Intersection (IA) and Words Co-occurrence (WCA), from scratch following the English WordNet semantic structure and ultimately a manually validation was made by some undergraduate students. Nevertheless, the quality of CWN is still not high enough. Thus, to further understand the quality in details, we did a manually evaluation of it. We randomly select 1000 records and after evaluation, we found that its accuracy is around 80%-90%, which is better than a fully automatic creation linguistic resource in general.

Obviously, the quality of a semantic linguistic resource influences its further applications significantly, which means the quality of a linguistic resource should be the higher the better. Thus, finding errors in a linguistic resource is inevitable. Since most of the errors related to semantic, they are hard to find by automatic algorithms. Crowdsourcing is a possible method but because 1) a linguistic resource consists of millions of records and the data size keeps increasing; 2) multilingual and structured data requires a relatively high-level educated person to validate, making the cost of crowdsourcing is still expensive.

Thus, thirdly, we intented to find a method that can further reduce the cost of finding errors in a large-scale multilingual semantic linguistic resource, which is also the main focus of this work. Our solution was inspired from an idea named Game with a Purpose (GWAP), proposed by Luis von. The basic idea is since many people spent a huge number of hours to play games each year, if hiding tasks in a game is possible, players can benefit us when they are playing the game as the side effect. Thus, to test our idea, UKC and Chinese language was adopted as our case study. To understand Chinese games better, we did a survey of all eminent Chinese knowledge games. We found that the question-answer pair is commonly utilized for most of Chinese knowledge games. While, a lot of kinds of errors exist in a linguistic resource, one game may not be enough to find all of them. To maximize reusable components, we create a UKC games framework, transferring UKC into game-like data and collecting feedback from games, which providing 2 kinds of questions, 4 kinds of options, and more than 100 domains.

Fourthly, Concept Challenge Game was developed as the first try to evaluate our idea. The game derived from a game named Word Challenge Game, which is to measure a player's vocabulary size. We use this game mode and hide our purpose perfectly by switching an English word to an English synset as a question. And, similarly, this game is to evaluate a player's English concept size and at the same time players can learn new concepts. It has 3 game modes, 'play with domains', 'play with difficulty level' and 'play dynamically'. In addition, we add leaderboards and No.1 player messages as entertainment elements to encourage people to get higher game scores.

Finally, as a result, our evaluation shows that players spend a significant amount of time playing the game at a short time span. A major constraint for the players was the type of the phone they were using to play the game. Numerous participants were unable to play as they owned an iPhone, but the system was developed for Android phones. The implementation of the game indicates that the game players with very limited linguistic background can also be involved in finding different kinds of error in a multilingual lexico-semantic resource which can be a major help while building a high quality lexico-semantic resource. We further extended our game to Italian language to verify the performance of the game for the other languages. A promising result shows that our game has the ability to find errors for the other languages.

As we discussed in the dissertation, basically, there are 3 kinds of errors which are semantic structure errors, sense errors and word errors. Our game has a good performance in finding errors with respect to the last two kinds, which means the semantic structure errors still need to be focus on. Furthermore, Concept Challenge Game also has some problems in terms of an isolate synset with multiple rare meanings, for example, run has 56 meanings in WordNet, a lot of them cannot be found even in some dictionaries. Thus, as the future work, we are going to develop more games for UKC games framework by concentrating on these two kinds of errors.

Bibliography

1 *Bresciani, Paolo, Anna Perini, Paolo Giorgini, Fausto Giunchiglia, and John Mylopoulos. "A knowledge level software engineering methodology for agent oriented programming." In Proceedings of the fifth international conference on Autonomous agents, pp. 648-655. ACM, 2001.*

2 *Giunchiglia, Fausto, Vincenzo Maltese, and Biswanath Dutta. "Domains and context: first steps towards managing diversity in knowledge." Web Semantics: Science, Services and Agents on the World Wide Web 12 (2012): 53-63.*

3 *Kilgarriff, Adam, and Christiane Fellbaum. "WordNet: An Electronic Lexical Database." (2000): 706-708.*

4 *Miller, A. G. "WordNet: a lexical database for English Communications of the ACM 38 (11) 3941." Niemela, I (1995).*

5 *Giunchiglia, Fausto, Biswanath Dutta, and Vincenzo Maltese. "From knowledge organization to knowledge representation." (2013).*

6 *Giunchiglia, Fausto, Biswanath Dutta, Vincenzo Maltese, and Feroz Farazi. "A facet-based methodology for the construction of a large-scale geospatial ontology." Journal on data semantics 1, no. 1 (2012): 57-73.*

7 *Vossen, Piek. "EuroWordNet: a multilingual database for information retrieval." In Proceedings of the DELOS workshop on Cross-language Information Retrieval, pp. 5-7. 1997.*

8 *林杏光等，1994，《现代汉语语述语动动机机器词典》，北京语言学院出版社*

9 *陈小荷. "一个面向工程的语义分析体系." 语言文字应用 2 (1998): 73-78*

10 *Dong, Zhendong, and Qiang Dong. HowNet and the Computation of Meaning. Singapore: World Scientific, 2006.*

11 *詹卫东, and 刘群. "词的语义分类在汉英机器翻译中所起的作用以及难以处理的问题 ⊖." (1997).*

12 *Jiangsheng, Yu, and Yu Shiwen Liu Yang Zhang Huarui. "Introduction to Chinese Concept Dictionary." ICCC2001 (Singapore).*

13 *Yu, J. S., and S. W. Yu. "The Structure of Chinese Concept Dictionary, accepted by Journal of Chinese Information Processing, 2001." (2001).*

14 *Yu, Jiangsheng, Yang Liu, and S. W. Yu. "The Specification of Chinese Concept Dictionary." Journal of Chinese Language and Computing 13, no. 2 (2003): 176-193.*

15 *Xu, Renjie, Zhiqiang Gao, Yingji Pan, Yuzhong Qu, and Zhisheng Huang. "An integrated approach for automatic construction of bilingual Chinese-English WordNet." In Asian Semantic Web Conference, pp. 302-314. Springer Berlin Heidelberg, 2008.*

16 *Giunchiglia, Fausto, Vincenzo Maltese, and Biswanath Dutta. "Domains and context: first steps towards managing diversity in knowledge." Web Semantics: Science, Services and Agents on the World Wide Web 12 (2012): 53-63.*

17 *Bentivogli, Luisa, and Emanuelle Pianta. "Looking for lexical gaps." InProceedings of the Ninth EURALEX International Congress, EURALEX 2000: Stuttgart, Germany, August 8th-12th, 2000, pp. 663-669. 2000.*

18 *Von Ahn, Luis. "Human computation." In Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on, pp. 1-2. IEEE, 2008.*

19 *Yuen, Man-Ching, Ling-Jyh Chen, and Irwin King. "A survey of human computation systems." In Computational Science and Engineering, 2009. CSE'09. International Conference on, vol. 4, pp. 723-728. IEEE, 2009.*

20 *Von Ahn, Luis, Manuel Blum, Nicholas J. Hopper, and John Langford. "CAPTCHA: Using hard AI problems for security." In International Conference on the Theory and Applications of Cryptographic Techniques, pp. 294-311. Springer Berlin Heidelberg, 2003.*

21 *Howe, Jeff. "The rise of crowdsourcing." Wired magazine 14, no. 6 (2006): 1-4.*

22 *Brabham, Daren C. "Crowdsourcing as a model for problem solving: An introduction and cases." Convergence 14, no. 1 (2008): 75-90.*

23 *Brabham, Daren C. "A model for leveraging online communities." The participatory cultures handbook 120 (2012).*

24 *Zesch, Torsten, and Iryna Gurevych. "Wisdom of crowds versus wisdom of linguists-measuring the semantic relatedness of words." Natural Language Engineering 16, no. 1 (2010): 25.*

25 *Quinn, Alexander J., and Benjamin B. Bederson. "Human computation: a survey and taxonomy of a growing field." In Proceedings of the SIGCHI conference on human factors in computing systems, pp. 1403-1412. ACM, 2011.*

26 *Scekic, Ognjen, Hong-Linh Truong, and Schahram Dustdar. "Incentives and rewarding in social computing." Communications of the ACM 56, no. 6 (2013): 72-82.*

27 *Von Ahn, Luis. "Human computation." In Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on, pp. 1-2. IEEE, 2008.*

28 *Von Ahn, Luis, Ruoran Liu, and Manuel Blum. "Peekaboom: a game for locating objects in images." In Proceedings of the SIGCHI conference on Human Factors in computing systems, pp. 55-64. ACM, 2006.*

29 *Von Ahn, Luis, Mihir Kedia, and Manuel Blum. "Verbosity: a game for collecting common-sense facts." In Proceedings of the SIGCHI conference on Human Factors in computing systems, pp. 75-78. ACM, 2006.*

30 *Law, Edith LM, Luis Von Ahn, Roger B. Dannenberg, and Mike Crawford. "TagATune: A Game for Music and Sound Annotation." In ISMIR, vol. 3, p. 2. 2007.*

31 Orkin, Jeff, and Deb Roy. "The restaurant game: Learning social behavior and language from thousands of players online." Journal of Game Development 3, no. 1 (2007): 39-60.

32 Yuen, Man-Ching, Ling-Jyh Chen, and Irwin King. "A survey of human computation systems." In Computational Science and Engineering, 2009. CSE'09. International Conference on, vol. 4, pp. 723-728. IEEE, 2009.

33 Ho, Chien-Ju, and Kuan-Ta Chen. "On formal models for social verification." In Proceedings of the ACM SIGKDD Workshop on Human Computation, pp. 62-69. ACM, 2009.

34 Eiben, Christopher B., Justin B. Siegel, Jacob B. Bale, Seth Cooper, Firas Khatib, Betty W. Shen, Barry L. Stoddard, Zoran Popovic, and David Baker. "Increased Diels-Alderase activity through backbone remodeling guided by Foldit players." Nature biotechnology 30, no. 2 (2012): 190-192.

35 Grant, Lyndsay, Hans Daanen, Steve Benford, Alastair Hampshire, Adam Drozd, and Chris Greenhalgh. MobiMissions: the game of missions for mobile phones. ACM, 2007.

36 Vannella, Daniele, David Jurgens, Daniele Scarfini, Domenico Toscani, and Roberto Navigli. "Validating and Extending Semantic Knowledge Bases using Video Games with a Purpose." In ACL (1), pp. 1294-1304. 2014.

37 Navigli, Roberto, and Simone Paolo Ponzetto. "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network." Artificial Intelligence 193 (2012): 217-250.

38 Herdagdelen, Amac, and Marco Baroni. "The Concept Game: Better Commonsense Knowledge Extraction by Combining Text Mining and a Game with a Purpose." In AAAI Fall Symposium: Commonsense Knowledge. 2010.

39 Venhuizen, Noortje, Kilian Evang, Valerio Basile, and Johan Bos. "Gamification for word sense labeling." In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013). 2013.

40 Seemakurty, Nitin, Jonathan Chu, Luis Von Ahn, and Anthony Tomasic. "Word sense disambiguation via human computation." In Proceedings of the acm sigkdd workshop on human computation, pp. 60-63. ACM, 2010.

41 Bentivogli, Luisa, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. "Revising the wordnet domains hierarchy: semantics, coverage and balancing." In Proceedings of the Workshop on Multilingual Linguistic Ressources, pp. 101-108. Association for Computational Linguistics, 2004.

42 Magnini, Bernardo, and Gabriela Cavaglia. "Integrating Subject Field Codes into WordNet." In LREC, pp. 1413-1418. 2000.

43 Pianta, E. "Bentivogli L. Girardi C.(2002) Multiwordnet: developing an aligned multilingual database." In First International Conference on Global WordNet, Mysore, India.

44 *Mahdisoltani, Farzaneh, Joanna Biega, and Fabian Suchanek. "Yago3: A knowledge base from multilingual wikipedias." In 7th Biennial Conference on Innovative Data Systems Research. CIDR Conference, 2014.*

45 *De Melo, Gerard, and Gerhard Weikum. "Towards a universal wordnet by learning from combined evidence." In Proceedings of the 18th ACM conference on Information and knowledge management, pp. 513-522. ACM, 2009.*

46 *Navigli, Roberto, and Simone Paolo Ponzetto. "BabelNet: Building a very large multilingual semantic network." In Proceedings of the 48th annual meeting of the association for computational linguistics, pp. 216-225. Association for Computational Linguistics, 2010.*

47 *Howe, Jeff. "The rise of crowdsourcing." Wired magazine 14, no. 6 (2006): 1-4.*

48 *Zaidan, Omar F., and Chris Callison-Burch. "Crowdsourcing translation: Professional quality from non-professionals." In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 1220-1229. Association for Computational Linguistics, 2011.*

49 *Brabham, Daren C. "Crowdsourcing as a model for problem solving: An introduction and cases." Convergence 14, no. 1 (2008): 75-90.*

50 *Brabham, Daren C. "A model for leveraging online communities." The participatory cultures handbook 120 (2012).*

51 *Zesch, Torsten, and Iryna Gurevych. "Wisdom of crowds versus wisdom of linguists-measuring the semantic relatedness of words." Natural Language Engineering 16, no. 1 (2010): 25.*

52 *Quinn, Alexander J., and Benjamin B. Bederson. "Human computation: a survey and taxonomy of a growing field." In Proceedings of the SIGCHI conference on human factors in computing systems, pp. 1403-1412. ACM, 2011.*

53 *Scekic, Ognjen, Hong-Linh Truong, and Schahram Dustdar. "Incentives and rewarding in social computing." Communications of the ACM 56, no. 6 (2013): 72-82.*

54 *Bollacker, Kurt, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. "Freebase: a collaboratively created graph database for structuring human knowledge." In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pp. 1247-1250. AcM, 2008.*

55 *Tawfik, Ahmed, Fausto Giunchiglia, and Vincenzo Maltese. "A Collaborative Platform for Multilingual Ontology Development." World Academy of Science, Engineering and Technology 8, no. 12 (2014): 1.*

56 *R.Xu,Z.Gao,Y.Pan,Y.Qu,andZ.Huang.An integrated approach for automatic construction of bilingual chinese-english wordnet. In Asian Semantic Web Conference, pages 302–314. Springer, 2008*

57 *Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum. "Yago: a core of semantic knowledge." In Proceedings of the 16th international conference on World Wide Web, pp. 697-706. ACM, 2007.*

58 *Bizer, Christian, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. "DBpedia-A crystallization point for the Web of Data." Web Semantics: science, services and agents on the world wide web 7, no. 3 (2009): 154-165.*

Appendix A

A list for all imported domains

```
id:1    domain:Chess    :::    number:28
id:2    domain:Rugby    :::    number:6
id:3    domain:Jewellery    :::    number:108
id:4    domain:Fencing    :::    number:8
id:5    domain:Entomology    :::    number:593
id:6    domain:Psychology    :::    number:363
id:7    domain:Badminton    :::    number:8
id:8    domain:Commerce    :::    number:502
id:9    domain:Optics    :::    number:161
id:10    domain:Table Tennis    :::    number:22
id:11    domain:Furniture    :::    number:389
id:12    domain:History    :::    number:288
id:13    domain:Wrestling    :::    number:33
id:14    domain:Vehicles    :::    number:75
id:15    domain:Cricket    :::    number:24
id:16    domain:Telegraphy    :::    number:11
id:17    domain:Free Time    :::    number:206
id:18    domain:Sculpture    :::    number:32
id:19    domain:Psychoanalysis    :::    number:45
id:20    domain:Surgery    :::    number:32
id:21    domain:Topography    :::    number:5
id:22    domain:Psychological Features    :::    number:1109
id:23    domain:Botany    :::    number:3
id:24    domain:Book Keeping    :::    number:28
id:25    domain:Gastronomy    :::    number:2439
id:26    domain:Racing    :::    number:100
id:27    domain:Meteorology    :::    number:204
id:28    domain:Sub    :::    number:1
id:29    domain:Bowling    :::    number:33
id:30    domain:Banking    :::    number:98
id:31    domain:Mountaineering    :::    number:8
id:32    domain:Plants    :::    number:6221
id:33    domain:Social Science    :::    number:9
id:34    domain:Card    :::    number:129
id:35    domain:Exchange    :::    number:199
id:36    domain:Baseball    :::    number:145
id:37    domain:Pharmacy    :::    number:460
id:38    domain:Sexuality    :::    number:198
id:39    domain:Mechanics    :::    number:495
id:40    domain:Transport    :::    number:900
id:41    domain:Person    :::    number:1841
id:42    domain:Roman Catholic    :::    number:23
id:43    domain:Art    :::    number:520
id:44    domain:Time Period    :::    number:557
id:45    domain:Biology    :::    number:14911
id:46    domain:Archaeology    :::    number:49
id:47    domain:Athletics    :::    number:21
id:48    domain:Economy    :::    number:851
id:49    domain:Pure Science    :::    number:46
id:50    domain:Sociology    :::    number:p
id:51    domain:Computer Science    :::    number:469
id:52    domain:Linguistics    :::    number:1234
id:53    domain:Agriculture    :::    number:275
id:54    domain:Body Care    :::    number:179
id:55    domain:Cinema    :::    number:37
id:56    domain:Food    :::    number:502
id:57    domain:Military    :::    number:1169
id:58    domain:Occultism    :::    number:31
id:59    domain:Administration    :::    number:629
id:60    domain:Humanities    :::    number:25
id:61    domain:Cycling    :::    number:8
id:62    domain:Nautical    :::    number:476
id:63    domain:Soccer    :::    number:15
id:64    domain:Fashion    :::    number:899
id:65    domain:Philately    :::    number:4
id:66    domain:Religion    :::    number:1069
id:67    domain:Dentistry    :::    number:23
id:68    domain:Grammar    :::    number:155
id:69    domain:Statistics    :::    number:4
id:70    domain:Physics    :::    number:749
id:92    domain:Acoustics    :::    number:104
id:93    domain:Oceanography    :::    number:9
id:94    domain:Artisanship    :::    number:143
id:95    domain:Paranormal    :::    number:10
id:96    domain:Tourism    :::    number:320
id:97    domain:Industry    :::    number:962
id:98    domain:Sport    :::    number:568
id:99    domain:Betting    :::    number:39
id:100    domain:Archery    :::    number:4
id:101    domain:Genetics    :::    number:22
id:102    domain:Physiology    :::    number:416
id:103    domain:Swimming    :::    number:53
id:104    domain:Factotum    :::    number:15871
id:105    domain:Town Planning    :::    number:382
id:106    domain:Astronautics    :::    number:20
id:107    domain:Skating    :::    number:18
id:108    domain:Boxing    :::    number:52
id:109    domain:Golf    :::    number:92
id:110    domain:Literature    :::    number:481
id:111    domain:Volleyball    :::    number:4
id:112    domain:Architecture    :::    number:125
id:113    domain:Color    :::    number:214
id:114    domain:Rowing    :::    number:6
id:115    domain:Ethnology    :::    number:29
id:116    domain:Anatomy    :::    number:1759
id:117    domain:Philology    :::    number:35
id:118    domain:Radio+Tv    :::    number:94
id:119    domain:Chemistry    :::    number:2082
id:120    domain:University    :::    number:134
id:121    domain:Theatre    :::    number:189
id:122    domain:Applied Science    :::    number:25
id:123    domain:Metrology    :::    number:1190
id:124    domain:Radiology    :::    number:29
id:125    domain:Environment    :::    number:17
id:126    domain:Philosophy    :::    number:201
id:127    domain:Drawing    :::    number:95
id:128    domain:Railway    :::    number:59
id:129    domain:Aviation    :::    number:121
id:130    domain:Zoology    :::    number:1
id:131    domain:Anthropology    :::    number:249
id:132    domain:Mathematics    :::    number:527
id:133    domain:Electrotechnology    :::    number:79
id:134    domain:Basketball    :::    number:43
id:135    domain:Fishing    :::    number:60
id:136    domain:Earth    :::    number:50
id:137    domain:Money    :::    number:585
id:138    domain:Engineering    :::    number:40
id:139    domain:School    :::    number:227
id:140    domain:Publishing    :::    number:454
id:141    domain:Electricity    :::    number:221
id:142    domain:Atomic Physic    :::    number:61
id:143    domain:Number    :::    number:139
id:144    domain:Quality    :::    number:93
id:145    domain:Telephony    :::    number:52
id:146    domain:Diving    :::    number:14
id:147    domain:Folklore    :::    number:29
id:148    domain:Home    :::    number:123
id:149    domain:Theology    :::    number:38
id:150    domain:Tennis    :::    number:44
id:151    domain:Hunting    :::    number:151
id:152    domain:Hockey    :::    number:22
id:153    domain:Geology    :::    number:635
id:154    domain:Health    :::    number:45
id:155    domain:Buildings    :::    number:1631
id:156    domain:Astrology    :::    number:13
id:157    domain:Diplomacy    :::    number:16
id:158    domain:Skiing    :::    number:24
id:159    domain:Post    :::    number:52
id:160    domain:Finance    :::    number:149
id:161    domain:Telecommunication    :::    number:246
```

```
id:71    domain:Paleontology    :::    number:3
id:72    domain:Medicine    :::    number:2167
id:73    domain:Photography    :::    number:130
id:74    domain:Geometry    :::    number:166
id:75    domain:Play    :::    number:417
id:76    domain:Pedagogy    :::    number:207
id:77    domain:Graphic Arts    :::    number:55
id:78    domain:Music    :::    number:822
id:79    domain:Insurance    :::    number:49
id:80    domain:Enterprise    :::    number:283
id:81    domain:Dance    :::    number:121
id:82    domain:Mythology    :::    number:136
id:83    domain:Veterinary    :::    number:1
id:84    domain:Psychiatry    :::    number:104
id:85    domain:Astronomy    :::    number:223
id:86    domain:Geography    :::    number:977
id:87    domain:Law    :::    number:1224
id:88    domain:Hydraulics    :::    number:76
id:89    domain:Electronics    :::    number:219
id:90    domain:Tax    :::    number:79
id:91    domain:Biochemistry    :::    number:9
```

```
id:162    domain:Plastic Arts    :::    number:11
id:163    domain:Ecology    :::    number:1
id:164    domain:Football    :::    number:67
id:165    domain:Painting    :::    number:106
id:166    domain:Heraldry    :::    number:116
id:167    domain:Animals    :::    number:6737
id:168    domain:Animal Husbandry    :::    number:55
id:169    domain:Politics    :::    number:806
id:170    domain:Numismatics    :::    number:43
```