Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 5 - 38123 Povo - Trento (Italy)
http://disi.unitn.it

MANAGING LANGUAGE DIVERSITY ACROSS CULTURES: THE ENGLISH-MONGOLIAN CASE STUDY[1]

Amarsanaa Ganbold, Feroz Farazi,
Fausto Giunchiglia

October 2013

Technical Report # DISI-13-036

[1] This is a short version of the paper with title "An Experiment in Managing Language Diversity Across Cultures", also published as DISI Technical report DISI-13-032, and presented at the conference ICKM 2013.

Managing Language Diversity Across Cultures: the English-Mongolian Case Study

Amarsanaa Ganbold[†], Feroz Farazi and Fausto Giunchiglia

Department of Information Engineering and Computer Science,
University of Trento, via Sommarive 14 I-38123, Povo, Trento, Italy
{amarsanaag, farazi, fausto}@disi.unitn.it

Abstract. Developing ontologies from scratch appears to be very expensive in terms of cost and time required and often such efforts remain unfinished for decades. Ontology localization through translation seems to be a promising approach towards addressing this issue as it enables the greater reuse of the ontological (backbone) structure. However, managing language diversity across cultures remains as a challenge that has to be taken into account and dealt with right level of attention and expertise. In this paper we report the result of our experiment that was performed with approximately 1000 concepts, taken from the space ontology originally developed in English, by providing their translation into Mongolian.

Keywords: Ontology localization, space ontology, space domain, ontology, Semantic Web

1 Introduction

Building a true, flourishing and successful Semantic Web (Berners-Lee et al., 2001) should involve the participation from all cultures and languages across the world. In the development of the traditional Web this participation was spontaneous and could be made possible as the necessary tools and resources were available. With the Semantic Web one of the crucial lacks is the capacity to assign precise meaning to words that requires NLP tools that use Knowledge Bases (KB) – consisting of a TBox or ontological part (concepts and relations) and an ABox (entities and relations) – providing the necessary background knowledge. For enabling KB-human interaction and to process lexical content, concepts are represented as synsets in natural languages (e.g., English and Italian). WordNet (http://wordnet.princeton.edu/) is such a synset base developed in English at Princeton. Still for many languages such resources are not developed at all and for some others what is out there cannot be used effectively as they could not achieve critical mass. Yet its coverage is often unsatisfactory while dealing with domain specific tasks (Giunchiglia et al., 2010).

Towards solving the issue of the lack of coverage and to gain a critical mass of the concepts, some domain ontologies have already been developed. The *space ontology*

[†] On leave from the School of Information Technology, National University of Mongolia

(Giunchiglia[b] et al., 2012) is an example of such an ontology developed in English with comparatively a very large coverage of geo-spatial features and entities around the globe. Domain ontologies can also deal with the specificity of an area of knowledge, for example, relations and attributes specific to the domain. By reducing polysemy (the amount of words with same meaning), they can enable better semantic interoperability.

In this paper we describe the development of the space ontology in Mongolian starting from its English counterpart taken from the UKC that is an ontology with translation in multiple languages developed at the University of Trento. Building an ontology without human level accuracy is a potential obstacle in developing applications (e.g., word sense disambiguation and document classification). Synset base resources (linguistic representation of ontologies) such as WordNet and FinnWordNet (Lindén et al., 2010) are built manually to obtain better quality. Being concerned about the quality and giving utmost importance to it, we followed a manual approach. The contributions of our paper include:

i) The development of an ontology localization methodology that is domain and language independent and seems to achieve very high quality

ii) The development of a methodology for dealing with diversity (e.g., lexical gaps) across cultures and languages

iii) Lessons learned from the execution of the whole process in the generation of the space ontology in Mongolian

The paper is organized as follows. In Section 2 we provide the detailed description of the UKC. Section 3 gives an overview of the space ontology. In Section 4 we describe the macro-steps of the translation process. In Section 5 we describe the diversity across English and Mongolian cultures in terms of space related features. Section 6 reports the results, Section 7 discusses the lessons learned and Section 8 describes the related work. In Section 9, we provide the concluding remarks.

2 The Universal Knowledge Core

The Universal Knowledge Core (UKC) is a large-scale ontology which includes hundreds of thousands of concepts (e.g., lake, mountain chain) of real world entities (e.g., Lake Garda, Alps). It consists of three main components: *domain core*, *concept core* and *natural language core*.

As described in (Giunchiglia[b] et al., 2012), the domain core consists of various **domains,** where each of them represents an area of knowledge or field of study that we are interested in or that we are communicating about (Giunchiglia[a] et al., 2012). In other words, a domain can be a conventional subject of study (e.g., mathematics, physics), an application of pure disciplines (e.g., engineering, mining), the aggregation of such fields (e.g., physical science, social science) or a daily life topic (also called Internet domains, e.g., sport, music). Each domain is organized in

facets, where a facet can be defined as a hierarchy of homogeneous concepts describing the different aspects of meaning (Giunchiglia et al., 2009). According to our methodology (Giunchiglia et al., 2013), called DERA, where D stands for Domain, facets are classified into three categories: Entity class (E), Relation (R) and Attribute (A). For example, in the space ontology, country and continent are **entity classes**. **Relations** describe relations between entities; examples of spatial relations are near, above, far, etc. An **attribute** is a property of an entity, e.g., depth of a lake.

The concept core consists of concepts or classes and semantic relations between them. The concepts in the concept core form a directed acyclic graph which provides the terms and the structure from which facets are defined. A **concept** is a language independent representation of a set of synonymous words (synset) in natural language. For example, country, city, etc. The concept *city* can be represented as *city* in English, *città* (chit'a) in Italian, *xom* (khot) in Mongolian.

The natural language core is built with the complete integration of hierarchically organized synset bases, for instance WordNet and the Italian part of MultiWordNet (http://multiwordnet.fbk.eu). This component consists of words, senses, synsets and exceptional forms. A **word** is the basic lexical unit of the natural language core represented as a lemma. It can be multiword, phrase, collocation, etc. The words in the natural language core provide, for any given language, the translation of the concepts stored in the concept core.

Word senses are organized into four part-of-speeches -- noun, verb, adjective and adverb, one word may have more than one part-of-speech and synonym word senses with the same part-of-speech are grouped into synset. A **sense** is the meaning of a word. A word can have one or more senses each having a part-of-speech tag and belongs to only one synset. All senses of a given word are ranked according to most preferred usage. A **synset** is a set of words which share 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.

3 The Space domain

The space domain (Giunchiglia[b] et al.,) is a large-scale geospatial ontology built using the faceted approach. It was developed as the result of the complete integration of GeoNames (http://www.geonames.org) and WordNet. It is also known as space ontology and in this paper we refer to it with any of these names. It currently consists of nearly 17 facets, around 980 concepts and 8.5 million entities. The ontology (excluding entities) is integrated into the UKC. Some examples of facet are *geological formation* (e.g., mountain, hill), *body of water* (e.g., sea, lake), *administration division* (e.g., state, province) and *facility* (e.g., university, industry).

In Figure 1 we provide a partial bird's eye view of the whole set of facets. Note that facets are not connected to each other and they do not have concept overlap across or within them.

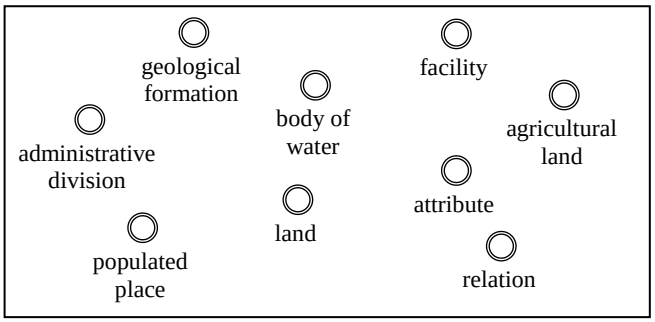

Figure 1. A subset of the facets of the *Space* domain

4 Translation approach

The main idea of the translation process is to take the objects of a domain of interest from the source language, in this case English, and to produce the corresponding representation in a target language, e.g., Mongolian in order to update the UKC with translations. The process includes the translation of the synset words and glosses. A direct translation of the English synsets and glosses is provided whenever possible. However, the world is full of diversity and people of a particular culture might not be aware of some concepts. For instance, Mongolia is a landlocked country, thus some terms (e.g., *dry dock*, *quay*, *pier*, etc.) related to seaport are not known to the community or are rarely used.

In order to provide the most suitable translation for a synset, we follow the macro-steps described below.

1. A **language translator** takes a synset provided in the source language and gets a clear understanding of its meaning. In case of difficulty, he/she finds the corresponding images or videos of the synset word(s) on the Web to perceive the concept through visualization.
2. The **language translator** provides a suitable translation of the word(s) in the target language. With suitable we mean word, multiword, co-occurrence and phrasal representation as we do not allow free combination of words as translation of a word. In case of unavailability of the word(s) for the given meaning, the translator can mark it as a lexical gap. However, the translator always provides the translation of the gloss.
3. A **language validator** evaluates the translation of the word(s) and gloss of the synset. In case the concept is marked as a gap, the validator either confirms the gap or translates the word(s).
4. Upon receiving feedback on the synset, the **language translator** goes through the comments and updates the translation where necessary. In case of disagreement, the language translator provides comments including mostly the rationale about the disagreement.
5. The **language validator** reevaluates the updated translation. In case of disagreement, the validator generates further feedback and sends it back to the

language translator (step 4). Even if after a few iterations a disagreement is not resolved, a second language validator is consulted. If agreed upon, the validation for the given synset is over.

6. A **UKC language validator** takes the validated translation to evaluate its correctness from both the language and UKC perspectives. The validator corrects the mistakes and resolve the issues (if any) communicating with the language validator (if necessary), possibly in a few iterations. Finally, he/she asks a UKC validator for importing the translation to the UKC.
7. The **UKC validator** runs an automatic validation tool to evaluate if the provided input is compliant with the UKC. In case of finding errors in the data, it is corrected with the help of the UKC language validator (if needed) possibly iterating a few times. Once all the issues are resolved, the UKC validator imports the translation to the UKC.

Following these steps we translated the *space ontology* into Mongolian end-to-end, evaluated and finally imported the translations to the UKC. To achieve optimal quality while executing the whole process, we set the criteria that translators and various validators possess competences necessary for the task. The language translator should be a native speaker in the country of origin of the target language with a good command of the source language. The language validator should be a linguist possessing the necessary language competences. The UKC language validator is a native speaker of the target language with knowledge of the UKC. The UKC validator is an expert on the UKC with no specific competence on the language.

From a geographical point of view we expect that, in most cases, the language core will be developed in the countries where that language is spoken, while the UKC is and will be developed centrally. The UKC language validator whenever possible should operate centrally where the UKC validator is. This spatial distribution of operations and operators has been designed as an attempt to preserve local diversity and, at the same time, to deal with the need of central coordination required because of existence of a unique, single UKC. The underlying model is that there is a single world, represented by the UKC, and many different views of the world, each represented by a different natural language. The diversity of the world is therefore captured, as it will be described in detail in the next section, in the mapping from the informal natural languages and the unique UKC formal language.

5 Types of diversity

The translation or localization is the adaptation of a piece of knowledge to a particular language and culture (Suárez et al., 2008). This is nontrivial and linguistic experts might help in this task. Moreover, the localization should be based on the perception of the concepts and entities in the real world within the local communities and not on the literal translation.

5.1 Concepts

We assume concepts to be universal; however, their representation in natural languages varies. Within the same language a concept might be referred with multiple terms (known as synonymy) and multiple concepts might be referred with the same term (known as polysemy).

The concepts *valley*, *dale* and *hollow* are represented with the same term in Mongolian. Moreover, in the UKC dale and hollow are subordinate concepts of the valley. In this case translating them in the target language increases the polysemy. However, we translate them because within the Mongolian culture people can classify their (real world) entities under the specific concept.

Lexical gaps are those concepts that do not have a succinct representation in a given language. However, they can be expressed as a free combination of words (Bentivogli et al., 2000). For example, the concept *parish – (the local subdivision of a diocese committed to one pastor)* is a lexical gap in Mongolian. The variation in the concept lexicalization from the source language (S) to the target language (T) is depicted in Figure 2(a).

As the lexical gap is a feature of the languages, it does happen with all of them. There can be a gap also from the target to source language. For instance, the Mongolian words **буу**ц (buuts) and **буй**р (buir) are gaps in English. The word *buuts* can be represented in English as *an area of dried and accumulated manure where a nomadic family was living* and the word *buir* can be represented in English as *a round shaped spot where a nomadic yurt was built*. Note that these words lack a succinct representation in English. Therefore we consider them as gaps. This phenomenon is drawn in Figure 2(b).

The nomadic lifestyle of Mongolians is the source of these concepts that are not used in the English speaking cultures across the globe.

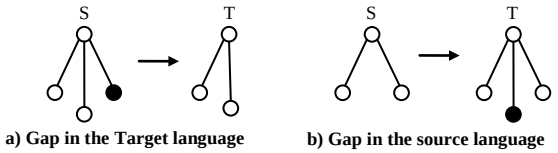

Figure 2. Variations of concept localization

5.2 Senses

In space ontology some words have multiple senses that have subtle difference in meaning. The word *fissure* has two such senses as provided below:

[S1]: *crack, cleft, crevice, **fissure**, scissure – (a long narrow opening)*
[S2]: ***fissure** – (a crack associated with volcanism)*

The two concepts associated with the given word are hyponym of *continental depression* and they can be represented with the same word(s) in the target language. This phenomenon is shown in Figure 3(a).

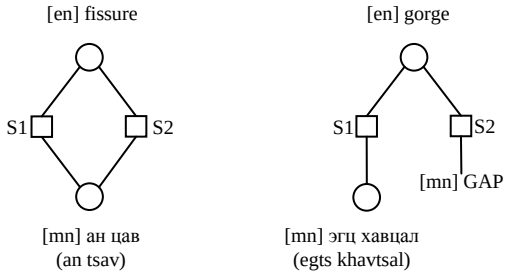

Figure 3. Word sense diversity

Polysemous words in the source language might correspond to lexical gaps for a subset of senses. For instance, *gorge* has two senses within Space ontology and one of them is a gap as depicted in Figure 3(b), where 'mn' and 'en' denote Mongolian and English, accordingly.

5.3 Synsets

Words in a synset can be directly translated into the target language. However, for some of them there might be a lack of translation. For example, the synset *mountain peak (the top point of a mountain or hill)* has 6 words of which 3 of them lack translation into Mongolian.

In gloss paraphrasing some parts of the glosses sometimes are obtained using words with a very close or similar meaning instead of exact translation. Though our first preference is to provide the exact translation, in many cases this could not be achieved. The following example shows a paraphrased translation where the phrase "near a shore" is eliminated from Mongolian version. In this situation, there is no difference between bank and shore in Mongolian language.

> [in English] ***oceanic sandbank*** *– a submerged bank of sand near a shore, can be exposed at low tide*
> [in Mongolian] ***далайн элсэн эрэг*** (gl. oceanic bank of sand) *– шунгаж орсон далайн элсэн эрэг, далайн давалгааны намхан хаялганд үзэгддэг (gl. a submerged sea bank of sand, visible at low tide)*

Example sentences in glosses were also paraphrased or added newly in order to provide a better explanation.

6 Results

We could translate 91.88% of the concepts of the space ontology into Mongolian and the remaining 8.12% were identified as lexical gaps. In Table 1 we report the detailed statistics of the translation task and the obtained results.

Facets	Concepts	Translated	Disagreed words	Disagreed glosses	Translator Identified Gaps	Finally accepted Gaps	Finally Localized Concepts
administrative division	18	18	2	4	0	0	18
agricultural land	19	19	2	1	0	0	19
attribute	85	73	1	23	12	10	75
barren land	7	7	1	0	0	0	7
facility	357	357	54	64	0	2	355

forest	5	5	5	4	0	0	5
geological formation	200	150	73	87	50	52	148
land	15	15	2	3	0	2	13
plain	12	12	0	0	0	3	9
rangeland	8	8	1	4	0	0	8
region	46	44	6	0	2	2	44
relation	54	54	8	32	0	0	54
wetland	8	8	3	1	0	0	8
abandoned facility	16	15	4	1	1	1	15
body of water	116	106	24	17	10	3	113
populated place	13	10	2	1	3	2	11
seat of government	6	4	0	1	2	2	4
Total number of objects	**985**	**905**	**188**	**243**	**80**	**79**	**906**

Table 1. Localization result of the Space domain

In Table 1, the number of concepts per facet is shown separately, e.g., administrative division has 18 concepts, agricultural land has 19 concepts and so on. Note that for the sake of space, we group the statistics of all attribute facets as attribute and relational ones under relation.

Language Translators provided Mongolian translation for 905 concepts *Language Validators* provided feedback on each of the produced synset words and glosses separately that help us achieving better quality. The validation procedure identified 188 disagreed words and 243 disagreed glosses. Cases such as disagreements and modifications for improvement were solved in iterations (as many as needed) until the translators and validators reached to an agreement. The highest number of iterations was recorded as 4.

Language Validators' evaluation of the lexical gaps revealed that the translators proposed 10 false positives out of 80. We also identified that the translators produced 9 false positive translations of the concepts whereas they are gaps. In the end, we found that there are in total 79 gaps and 906 concept translations being accepted. The *UKC Language validator* and *UKC validator* reported a few (around 5) conflicts which were then solved with little effort. It is worth mentioning that *Language Translators* proposed to add 7 new concepts to the space ontology. This is only initial work and we expect that a few more concepts will be added with the evolution of the space ontology.

7 Lessons learned

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 label. But it remained obscure until the whole translation task was finished. This ranking could be defined once all the concepts are translated. This is a non-trivial task to accomplish because deciding acceptable ranks might require local community agreement or the consultation of high quality linguistic resources that are often insufficient for domain specific tasks in many languages.

Synonymous words within the synsets were often increased after translations were evaluated by the Language Validators. This was the case since Language Translators concentrate in providing the target language correspondence

representation of the knowledge objects taken from the source language within a reasonable amount of time. This often results in the postponement of the addition of synsets.

Parts of the glosses that follow the same syntactic pattern in the source language can be translated with little effort. For instance, the gloss part *a facility for [verb]+ing [object]* appeared in around one tenth of the concepts. We repeated the same translation for the part that matched completely. Moreover, we used the translation memory technique which provides a translation with recurrent structure in the same way as previous translations.

In order to introduce foreign cultures to the community, we can translate lexical gaps as free combination of words. However, this should not always be the case. A first reason is computational: the explicit marking of the lexical gaps could support the KB-based applications in reducing computation time by avoiding the management of (multi)words which will be very rarely or never used. A second, more important reason, is related to the actual existence of a free combination of words capable of capturing, in the mind of a native speaker with no knowledge of the original concept (as it exists in the foreign culture) what the concept actually means, in the real world.

8 Related work

MultiWordNet consists of several European language WordNets (Bentivogli et al., 2000). While producing Italian version of MultiWordNet, no literal translation was provided. They provided best possible Italian equivalents according to their skills and experiences in knowledge organization and linguistics. However, limited number of glosses has been provided, e.g., around 2k in Italian over 33k.

The ontology localization activity described in (Espinoza et al., 2009) is an attempt to address the localization and diversity issues. They proposed guidelines and methodologies for enriching ontology with multilingual information. However, we differ from them with respect to the target language and the development approach.

FinnWordNet was produced from WordNet with the help of professional translators and the output is monitored by bulk validation (Lindén et al., 2010). While producing the whole WordNet in Finish in 100 days, they traded off the quality for reducing the amount of translation time. Diversity in the languages such as lexical gaps is overlooked in this task.

9 Conclusion

In this paper, we proposed an approach for generating ontologies through translation from one language into another. This approach was developed to be applied independently of domain and language and to deal with the diversity across the languages. While translating the ontologies, we manage diversity with the identification of diversity features and their presence in a given target language by working together with the linguistic experts and/or native speakers living in the

country where it is spoken. We evaluated the effectiveness of the methodology by performing a case study for translating the space ontology into Mongolian. Finally, we achieved a very high quality human crafted space ontology in Mongolian. Our future plan includes the exploitation of this valuable resource to improve the accuracy of NLP tasks (see (Zaihrayeu et al., 2007)) and Concept Search (see (Giunchiglia et al., 2009)) in space domain.

Acknowledgments: The research leading to these results has received funding from the FP7 EU project Smart Society (G.A. n. 600854). We are thankful to Vincenzo Maltese for his valuable feedback.

References

1. Bentivogli, L., & Pianta, E. (2000). Looking for lexical gaps. In *In Proceedings of the Ninth EURALEX International Congress*.
2. Espinoza, M., Montiel-Ponsoda, E., & Gómez-Pérez, A. (2009). Ontology localization. In *K-CAP '09* (p. 33).
3. Giunchiglia, F., Dutta, B., & Maltese, V. (2013). From Knowledge Organization to Knowledge Representation. In *ISKO UK Conference*.
4. Giunchiglia[a], F., Dutta, B., Maltese, V., & Farazi, F. (2012). A facet-based methodology for the construction of a large-scale geospatial ontology. *Journal on Data Semantics*, *1*(1), 57–73.
5. Giunchiglia, F., Kharkevich, U., & Zaihrayeu, I. (2009). Concept search. *The Semantic Web Research and Applications, 5554/2009.*
6. Giunchiglia[b], F., Maltese, V., & Biswanath, D. (2012). Domains and context: first steps towards managing diversity in knowledge. *Web Semantics: Science, Services and Agents on the WWW, 12-13*, 53–63.
7. Giunchiglia, F., Maltese, V., Farazi, F., & Biswanath, D. (2010). GeoWordNet: a resource for geo-spatial applications. In *ESWC'10*.
8. Lindén, K., & Carlson, L. (2010). FinnWordNet – Finnish WordNet by Translation. *Nordic Journal of Lexicography*, *17*, 119–140.
9. Suárez-Figueroa, M. C., & Gómez-Pérez, A. (2008). First Attempt towards a Standard Glossary of Ontology Engineering Terminology. In *TKE08* (pp. 1–15).
10. Zaihrayeu, I., Sun, L., Giunchiglia, F., Pan, W., Ju, Q., Chi, M., & Huang, X. (2007). From Web Directories to Ontologies : Natural Language Processing Challenges. In *ISWC7/ASWC'07*(pp. 623–636).
11. Berners-Lee, T., Hendler, J. and Lassila, O. (2001). The semantic web. Scientific American, (284(5)), 34–43.
12. Giunchiglia, F., Dutta, B., Maltese, V.: Faceted Lightweight Ontologies. In "Conceptual Modeling: Foundations and Applications", Springer (2009).