The Taboo Challenge Competition

Michael Rovatsos, Dagmar Gromann, Gábor Bella

■ *Games have always been a popular domain of AI research, and they have been used for many recent competitions. Reaching human-level performance, however, often either focuses on comprehensive world knowledge or solving decision-making problems with unmanageable solution spaces. Building on the popular Taboo board game, the Taboo Challenge Competition addresses a different problem — that of bridging the gap between the domain knowledge of heterogeneous agents trying to jointly identify a concept without making reference to its most salient features. The competition, which was run for the first time at the 2017 IJCAI conference, aims to provide a simple testbed for diversity-aware AI where the focus is on integrating independently engineered AI components, while offering a scenario that is challenging enough to test the concept, yet simple enough not to require mastering general commonsense knowledge or natural language understanding. We describe the design of and preparation for the competition, and discuss the results and lessons learned.*

Successful approaches at solving games, such as Google's AlphaGo (Silver et al. 2016) or IBM's Watson playing Jeopardy (Ferrucci et al. 2010), have attracted broad interest from researchers and the general public. However, such approaches rely on large amounts of data, substantial computing resources, and the participants' ability to combine a host of different methods. In an effort to produce a scenario that stimulates research on challenging AI problems but is accessible to a broad range of participants, not just AI experts, we conceived of the Taboo Challenge. Based on the fun and popular game, the challenge is nontrivial, yet generally solvable for humans. In the Taboo board game, one agent guesses a concept that another agent describes without the use of taboo words that would make the concept too easy to guess. Teams consisting of a *Guesser* and a *Describer* achieve a higher score if they can detect the concept in question faster than their opponent teams.

Achieving human-level performance at Taboo requires significant commonsense reasoning capabilities, but is limited to guessing or describing a target concept. Thus, it does not require a comprehensive knowledge of the world or a deep understanding of natural language, as, for example, the Winograd Schema Challenge does (Levesque 2011). Additionally, the game is interactive, which means that it requires agents to respond based on previous steps in the dialogue, rather than just identifying a correct solution from among several choices, as in Jeopardy, the Winograd Schema Challenge, or standardized academic tests (Clark and Etzioni 2016). This aspect of the game offers opportunities to develop diversity-aware AI methods, as participants submitting agent implementations to the competition have to face teammates who have been independently developed and who thus will have internal semantic processing and interactive decision-making strategies unknown to the agent.

In a stylized, simplified scenario, this approach addresses a fundamental AI challenge that has been overlooked by other competitions — overcoming the diversity between different AI systems that need to be integrated. The obstruction introduced by the prohibited Taboo words makes it impossible to rank possible hypotheses according to the most salient features of the concept (which can be detected, for example, by using Web search engines and knowledge bases). Hence, the game forces agents to speculate about their partner's understanding of the domain, rather than just performing inference on their own knowledge.

A final attractive aspect of the problem is that it can potentially be solved through a wide range of AI approaches alone — logical inference, distributional semantics, graph-based algorithms, machine learning methods, and so on — or by various approaches in combination. Thus, it allows for the comparison both between different AI approaches and between AI solutions and human performance.

The Competition

In the first Taboo Challenge Competition,[1] held in June 2017 and presented as a workshop at IJCAI 2017 in Melbourne, we restricted the challenge to developing Guesser agents, and also limited the domain of concepts to the names of popular cities. Additionally, to reduce the complexity of natural language understanding tasks like parsing and semantic analysis, we restricted hints to simple noun phrases (nouns plus adjectives and/or adverbs). Figure 1 provides an overview of the challenge, including examples of the games played and also the iterative process we aim to establish for future editions.

The descriptions provided by our Describer agent were replayed hints from human games, and entries to the competition had to guess each city by interacting with a simple REST API. Taboo words for each city in the human games were crowdsourced on CrowdFlower[2] for an initial set of 300 large cities. We elicited popular terms for each city from 82 participants (mostly from the US and the UK), which, after we eliminated those for which we'd obtained fewer than four Taboo words, resulted in a final set of 226 cities. With this approach, eight to twelve Taboo words were generated for every target concept by more than one worker.

Using these city names and Taboo words, 30 English first language players generated 174 games using a web application, and another 109 games were generated using the mobile Android version of the web app called GUESSence.[3] Both apps were developed specifically for the competition. In this process, we collected only games that were successfully solved by human players in order to avoid games with low-quality hints that artificial Guessers would be unlikely to solve.

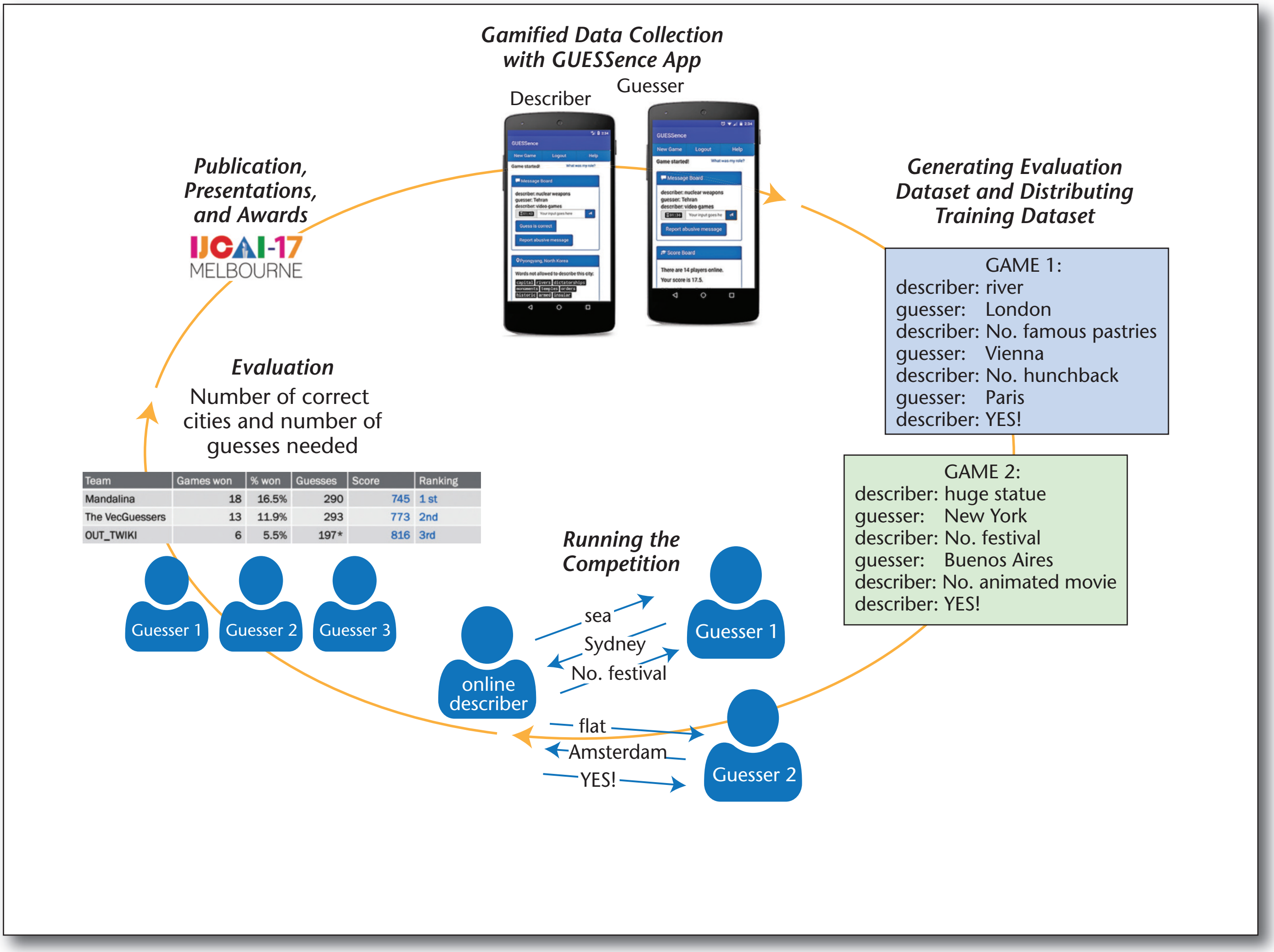
Interestingly, the number of hints needed to guess a city varied more than we expected. More than 25 percent of games were solved after just one hint, and more than another 50 percent could be solved using two to four hints, while a small number of games required up to ten hints. This might be taken as an indication that human-level performance is still far superior to that of (at least reasonably elaborate) artificial guessers trying to play the game.

Results

Overall, 10 teams registered to participate, three of which submitted a Guesser agent to the competition: The VecGuessers (University of Amsterdam), whose agent used a distributional semantics approach to match hints to cities; Mandalina (Bogazici University, Turkey), whose agent used a distributional semantics approach enhanced with geographical information; and OUT TWIKI (Open University of Cyprus), a system that used a combination of supervised learning and logic-based reasoning.

Participants were given a number of test games to use for training, and their Guesser agents were evaluated using the 109 games crowdsourced through the mobile app, with penalties for games not solved (that is, those where the online Describer ran out of human-sourced hints) and scoring based on the number of hints required to arrive at the solution. A specific advantage of the design of the competition was that the evaluation could be fully automated, making it independent of human judgment. We also provided participants with a baseline Guesser agent (Adrian et al. 2016) that attempted to geographically home in on the region for which the hints seemed most relevant.

The Mandalina team emerged as the winner of the competition, with 16.5 percent of all games solved and 290 hints required overall. The VecGuessers


Gamified Data Collection with GUESSence App
Describer
Guesser
Generating Evaluation Dataset and Distributing Training Dataset
Publication, Presentations, and Awards
IJCAI-17 MELBOURNE
Evaluation
Number of correct cities and number of guesses needed
Running the Competition
Guesser 1
Guesser 2
Guesser 3
online describer
Guesser 1
Guesser 2
sea
Sydney
No. festival
flat
Amsterdam
YES!

GAME 1:
describer: river
guesser: London
describer: No. famous pastries
guesser: Vienna
describer: No. hunchback
guesser: Paris
describer: YES!

GAME 2:
describer: huge statue
guesser: New York
describer: No. festival
guesser: Buenos Aires
describer: No. animated movie
describer: YES!

Team	Games won	% won	Guesses	Score	Ranking
Mandalina	18	16.5%	290	745	1 st
The VecGuessers	13	11.9%	293	773	2nd
OUT_TWIKI	6	5.5%	197*	816	3rd

Figure 1. Taboo Challenge Competition Data Collection and Evaluation Process.

came in second, with 11.9 percent games solved and 293 hints required. Finally, OUT TWIKI solved only 5.9 percent of the games, consuming 197 hints, though this result followed from the overly long response times of its complex reasoning engine, which caused the system to time out frequently.

Interestingly, of the 30 cities correctly guessed by any of the submitted Guesser agents overall, 24 were correctly guessed by a single competition entry, and only one city (Paris) appears in the list of 12 cities that each of the three Guesser agents most frequently generated as a guess. This result suggests that there is a high degree of diversity not only in the human games in our evaluation data, but also in the behaviors of the submitted Guesser agents, which reinforces our confidence that the scenario is indeed one where diversity awareness is key.

Awards for the winners were presented at The Taboo Challenge Competition Workshop that took place on August 29, 2017 in Melbourne as part of the IJCAI 2017 program, where the participants also had an opportunity to present the papers they had submitted alongside their implementations.

Lessons Learned

Despite our best attempts to simplify some of the elements of the competition, the task turned out to be much harder than expected. We attribute this to two factors.

First, many human players often solved the game after just one or two hints, and such hints were often highly contextual (for example, *terrorist attack* would immediately suggest a city where such an attack had taken place most recently). It is easy to see why an artificial Guesser developed to achieve good performance over a broad range of games would be unable to match human performance in these instances, but it should be possible to solve this problem by gathering more game data so that only more "solvable" instances are used for evaluation.

Second, while replaying hints from human games offers the great benefit of fully automated evaluation, successive hints in the replay do not take previous guesses into account, as they did when the original game was played by human players.

Unless competent Describer agents are implemented (which would, however, make the performance of a given Guesser dependent on the quality of the Describer), the only solution to this problem is to introduce human-based evaluation, where a Guesser plays against a human Describer who can take past guesses into account. Undoubtedly, this would also encourage implementations of more interesting dialogue strategies in Guesser agents, which is something we would like to see.

We believe that, even if these two problems were solved, the scenario would remain challenging in the future — after all, the success of the commercial board game version suggests that humans find it challenging enough to get enjoyment and suspense out of playing it repeatedly.

Future Plans

We aim to continue running the competition, and expect that its next installment will take place in spring 2018, with results presented at the joint IJCAI ECAI conferences in Stockholm, Sweden in July 2018. We are currently planning to add a Describer track to the current Guesser track, and to explore human-based evaluation as an additional way of assessing entries.

Using data from unsuccessful games is an avenue we wish to explore further. Our experience regarding the difficulty of the task even when using only successful human games, however, suggests that this direction may only become relevant once the submitted solutions achieve a higher performance on the current, simpler task.

Acknowledgements

This research has been funded by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement no. 607062, ESSENCE: Evolution of Shared Semantics in Computational Environments. We wish to thank the participants in the competition, and the volunteers who helped with the process of crowdsourcing Taboo words and human games.

Notes

[1] See www.essence-network.com/challenge for further details.

[2] www.crowdflower.com.

[3] The app is available from the Google Play Store at play.google.com/store/apps/details?id=com.guessence.iiia.essence.

References

Adrian, K.; Bilgin, A.; and Van Eecke, P. 2016. A Semantic Distance–Based Architecture for a Guesser Agent in ESSENCE's Location Taboo Challenge. Paper presented at the International Workshop on Diversity-Aware Artificial Intelligence (DIVERSITY 2016), The Hague, the Netherlands, July 26.

Clark, P., and Etzioni, O. 2016. My Computer Is an Honor Student — but How Intelligent Is It? Standardized Tests as a Measure of AI. *AI Magazine* 37(1): 5–12. doi.org/10.1609/aimag.v37i1.2636

Ferrucci, D.; Brown, E.; Chu-Carroll, J.; Fan, J.; Gondek, D.; Kalyanpur, A. A.; Lally, A.; Murdock, J. W,; Nyberg, E.; Prager, J.; Schlaefer, N.; and Welty, C. 2010. Building Watson: An Overview of the DeepQA Project. *AI Magazine* 31(3): 59–79. doi.org/10.1609/aimag.v31i3.2303

Levesque, H. J. 2011. The Winograd Schema Challenge. In *Logical Formalizations of Commonsense Reasoning — Papers from the AAAI 2011 Spring Symposium*. Technical Report SS-11-06.

Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; and Hassabis D. 2016. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* 529(7587): 484–489. doi.org/10.1038/nature16961

Michael Rovatsos is a reader at the School of Informatics of the University of Edinburgh, where he has led the Agents Group since 2004. He has published over 90 papers in multiagent systems on topics related to agent communication, multiagent planning, multiagent learning, and argumentation, and he is the overall coordinator for the 4-million euro ESSENCE Marie Curie Initial Training Network, which conceived of and organized the Taboo Challenge Competition.

Dagmar Gromann is a postdoc researcher at the Artificial Intelligence Research Institute (IIIA) in Spain and an experienced researcher in the ESSENCE Network. Her research focuses on learning cognitive schemas and knowledge representations from multilingual texts using machine learning and distributional semantics approaches, as well as aligning domain-specific resources.

Gábor Bella is a research associate at the University of Edinburgh and at the University of Trento. He is a senior member of the ESSENCE Network. His main area of study is multilingualism in computer systems, with a current focus on cross-lingual and domain-aware semantic interoperability (for example, data integration, ontology matching) over structured data sets.