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. However, reaching human-level performance 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 IJCAI 2017, 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 yet simple enough not to require mastering general commonsense knowledge or natural language understanding. We describe the design and preparation of the competition, discuss results, and lessons learnt.

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 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 as a scenario that is based on a popular, fun game, and is non-trivial, yet generally solvable for humans. In the Taboo board game, one agent guesses a concept another agent describes without the use of "Taboo words" that would make the concept too easy to guess. Teams of "Guesser" and "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 just guessing or describing a target concept. Thus, it does not require having a comprehensive knowledge of the world or a deep understanding of natural language, as, e.g., the Winograd Schema Challenge (Levesque et al 2011). Additionally, the game is interactive, which means that it requires agents to respond based on previous steps in the dialogue, rather than just identify a correct solution among several choices, as in Jeopardy, the Winograd Schema Challenge, or standardized academic tests (Clark and Etzioni 2016). This offers opportunities to develop *diversity-aware* AI methods, as participants in the competition submitting agent implementations have to face teammates that have been independently developed, and will have internal semantic processing and interactive decision-making strategies unknown to each other.

In a stylized, simplified scenario, this addresses a fundamental AI challenge that has been overlooked by other competitions—overcoming the diversity between different AI systems that need to be integrated. By introducing the "obstruction" of prohibited Taboo words, ranking possible hypotheses according to the most salient features of the concept (that can be detected, for example, by using Web search engines and knowledge bases) becomes impossible. Hence, the game *forces* agents to speculate about each other'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 very wide range of AI approaches alone or in combination, such as logical inference, distributional semantics, graph-based algorithms, and machine learning methods. Thus, it allows for the comparison between different AI approaches and between AI solutions and human performance.

The Competition

In the first edition of the 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) —see Figure 1 for examples of games and an overview of the whole process.

The descriptions provided by our Describer agent were "re-played" hints from human games, and entries to the competition had to guess each city by interacting with a very 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, eliciting popular terms for each city from 82 participants (mostly from the USA and the UK), which resulted in a final set of 226 cities after eliminating those for which we obtained fewer than four Taboo words, so that eight to twelve Taboo words provided by more than one worker were generated for every target concept.

Using these city names and Taboo words, 30 English first language players generated 174 games in 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 unlikely be able to solve.

Interestingly, the number of hints needed to guess a city varied more than we expected. More than 25% of games were solved with just one hint, and more than another 50% could be solved with 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, ten teams registered to participate, and three teams submitted a Guesser agent to the competition: The VecGuessers (University of Amsterdam), whose agent used a distributional semantics approach to match hints against cities, Mandalina (Boğaziçi University, Turkey), an agent that also 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 using the mobile app, with penalties for games not solved (i.e., those were the online Describer ran out of human-sourced hints) and for the number of hints required in a game until the solution was identified. A specific advantage of the design of the competition was that evaluation is fully automated, and did not involve human judgment. We also provided participants with a baseline Guesser agent (Adrian et al. 2016) that was based on attempting to geographically "hone in" on the region for which the hints seemed most relevant.

The Mandalina team emerged as the winner of the competition, with 16.5% of all games solved and 290 hints required overall. The VecGuessers came in second place with 11.9% games solved and 293 hints required. Finally, OUT_TWIKI solved 5.9% of the games consuming 197 hints, though this was due to it timing out frequently due to overly long response times caused by its complex reasoning engine.

Interestingly, out of the 30 cities correctly guessed by any of the submitted Guesser agents overall, 24 were only correctly guessed by a single competition entry, and only one city (Paris) appears in the list of the twelve cities that each of the three Guesser agents most frequently generated as a guess. This suggests that there is not only a high degree of diversity among the human games in our evaluation data, but also among the behaviour produced by 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 29th August 2017 in Melbourne as part of the IJCAI 2017 programme, where the participants also had an opportunity to present the papers they had submitted alongside their implementations.

Lessons learnt

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

First, many human players solved the game often after just one or two hints, and these were often highly contextual (e.g. “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 do not take previous guesses into account, whereas 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 their 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 humans find it challenging enough to get enjoyment and suspense out of playing it.

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 IJCAI-ECAI-2018 in Stockholm 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, although our experience regarding the difficulty of the task at hand even when only using successful human games suggests this may only become relevant once submitted solutions achieve high performance on the current, simpler task.

Figure 1: Taboo Challenge Competition data collection and evaluation process, showing also the iterative process we aim to establish for future editions

Acknowledgments

This research has been funded by the European Community's Seventh Framework Programme (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 https://www.essence-network.com/challenge for further details.

2. https://www.crowdflower.com

3. The app is available from the Google Play Store at https://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. In *Proceedings of International Workshop on Diversity-Aware Artificial Intelligence (DIVERSITY 2016)*, The Hague, NL, 2016.

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, 2016.

Ferrucci, D.; Brown, E.; Chu-Carroll, J.; Fan, J.; Gondek, D.; Kalyanpur, A.A.; Lally, A.; William Murdock, J.; Nyberg, E.; Prager, J.; Schlaefer, N.; and Welty C. 2010. Building Watson: An Overview of the DeepQA Project. *AI Magazine*, Fall 2010.

Levesque, H.J.; Davis, E.; and Morgenstern, L. 2011. The Winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning,* vol. 46.

Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484--489, 2016.

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 is the overall Coordinator for the €4-million ESSENCE Marie Curie Initial Training network which conceived of and organised the Taboo Challenge Competition.

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 (e.g., data integration, ontology matching) over structured data sets.

Dagmar Gromann is a post-doc 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.