A Semantic Distance based Architecture for a Guesser Agent in ESSENCE's Location Taboo Challenge

Adrian, K.; Bilgin, A.; Van Eecke, P.

[Link to publication](#)

DIVERSITY @ ECAI 2016

INTERNATIONAL WORKSHOP
ON DIVERSITY-AWARE ARTIFICIAL INTELLIGENCE

The Hague, Netherlands, 29th August 2016

Workshop Proceedings

A Semantic Distance based Architecture for a Guesser Agent in ESSENCE's Location Taboo Challenge

Kemo Adrian [1], **Aysenur Bilgin** [2] and **Paul Van Eecke** [3]

Abstract. Taboo is a word-guessing game in which one player has to describe a target term to another player by giving hints that are neither the target term nor other terms specified in a predetermined list of taboo words. The Location Taboo (LT) Challenge, which has been proposed by the ESSENCE Marie Curie Initial Training Network, is a version of Taboo that only contains cities as target terms and is intended to be played by artificial guesser agents. The hints are extracted from games played by many different human players, whose associations of cities with specific terms are often based on past experiences and therefore very diverse. Modeling this diversity in word associations is one of the main difficulties in solving the LT Challenge. In this paper, we propose a semantic distance based architecture for a guesser agent for the LT Challenge. The proposed architecture employs a two-step approach that narrows down the geographical area of the guess first to the country and then to the city. For ranking countries and cities, different distance metrics are used. As these techniques can be used on web documents crafted by many different individuals, they are well suited to model the diversity in word associations. The results of our evaluation on the LTC test set show that the proposed guesser agent can guess the target city with up to 23.17% accuracy. For 68% of the correct guesses, the proposed agent guesses the target city faster than its human counterpart.

1 Introduction

Taboo is a word-guessing game in which one player has to describe a target term to another player by giving hints that are neither the target term nor other terms specified in a predetermined list of taboo words. For example, a player might have to describe *water* without using *sea*, *blue* or *beverage*. The Location Taboo Challenge (LTC), which has been proposed by the ESSENCE Marie Curie Initial Training Network [1], is a version of Taboo that only contains cities as target terms and is intended to be played by artificial guesser agents. In the LTC, the hints, which are words associated to the target city, are sequentially provided to the guesser agent, and the goal is to guess the target location as soon as possible.

The hints are extracted from games that were played by various human players having different backgrounds and demographics. The associations that individual players make with cities are often based on their own past experiences, and are therefore very diverse. For example, people that have visited Spain only once in their lives might associate *tapas* with *Madrid*, whereas others may think of *tapas* being typical for Andalusian cities and may not even consider it as a

clue for *Madrid*. Modeling this diversity in word associations is one of the main difficulties in solving the ESSENCE LT Challenge.

In this paper, we propose a semantic distance based architecture for an LTC guesser agent. The proposed architecture employs a two-step approach that narrows down the geographical area of the guess first to the country and then to the city. For scoring the associative relevance of countries and cities with the given hints, the proposed architecture uses different distance measures. As these metrics are based on a large number of web documents crafted by various individuals, they considerably capture the diversity in word associations posed by human players.

The rest of this paper is structured as follows: Section 2 presents the game specification of the LT Challenge in more detail. Section 3 is dedicated to the background and previous work on modeling human behavior for word-guessing games. Section 4 presents the proposed architecture and the algorithms employed by our guesser agent. The experiments and the results are presented in Section 5. We provide a critical discussion in Section 6 followed by open research directions in Section 7. Finally, we draw conclusions in Section 8.

2 Game Specification

In this section, we introduce the most important aspects of the Location Taboo Challenge. A complete specification of the challenge can be found in [1].

An LTC game is played by two agents, the *describer* and the *guesser*. The game starts with the describer, providing a hint about a particular city anywhere in the world. Based on this hint, the guesser tries to guess the city that is being described. There are two possible outcomes after a guess has been made. For the outcome where the guess is correct, the game is considered to be successful. However, for the outcome where the guess is incorrect, the describer provides another hint and the game continues until the describer has consumed all the hints. The LT Challenge consists of implementing a guesser agent that can guess the correct city using the fewest number of guesses possible and before the describer runs out of hints. In the case where the describer runs out of hints and the correct guess has not yet been made, the game is considered to have failed.

For the LTC, the describer agent is provided by the authors of the challenge and the hints are crowd-sourced from real games played by human players. Therefore, the length of a game - i.e. the number of hints - is not fixed, but determined by the individual players. Also, it should be noted that the real-world dataset, which is provided by ESSENCE Network, consists of only successfully finished games. After each guess, the describer provides not only a new hint, but also the city that the human player (wrongfully) guessed. This information may be useful, or even necessary, in order to interpret

---

[1] IIIA-CSIC, email: kemo.adrian@iiia.csic.es
[2] Institute for Logic, Language and Computation, University of Amsterdam, Netherlands, email: a.bilgin@uva.nl
[3] Sony Computer Science Laboratory Paris, email: vaneecke@csl.sony.fr

the next hints, as these might be relative to the guesser's previous guesses (e.g. 'north' or 'close'). Hints are usually single words, but can occasionally be multi-word expressions. According to the rules of the LTC, the hints do not include proper names. An example game, adopted from [1] is shown in Figure 1.

Target:	Venice
D:	sea
G:	Sydney
D:	festival
G:	Rio de Janeiro
D:	river
G:	Rome
D:	art
G:	**Venice**

Figure 1. Location Taboo Challenge example game, adopted from [1], where D = Describer agent, G = Guesser agent

3 Background and Previous Work

What makes the LT Challenge so interesting and difficult is that the game is not about finding a correct or objectively verifiable answer to a specific question. Instead, it is about mimicking those associations that the human players have made, for whatever possible reason. The hints provided by the describer may not be necessarily true for the target city; yet, they are the depiction of an association that a human player made with this city. Therefore, an ideal implementation of the guesser agent should not only model common sense, but also simulate human beings' associative capabilities and collaborative game-playing behavior.

There is an impressive body of previous work on modeling common sense and human behavior for game playing. Heith et al. [9] present a range of techniques for understanding and conveying concepts based on word associations. These methods utilize human word association resources such as associative thesauri on the one hand; and corpus-based approaches, in particular Latent Semantic Analysis [6], Hyperspace Analog to Language [11] and Direct Co-occurrence Counts on the other hand. The models are evaluated both in a describer and a guesser role on Wordlery, a word-guessing game that is relatively similar to Taboo. The authors find that the models based on human word association resources are superior to the ones using corpus-based approaches.

A second, more famous, relevant research project is IBM's Watson, competing in the clue-guessing game Jeopardy! [4]. Watson uses IBM's massively parallel DeepQA architecture, combining hundreds of techniques and approaches in real time [7, 8]. The main difference between LT and Jeopardy! is that LT is a collaborative game, in which the describer tries to make the clues as easy and relevant as possible, whereas in Jeopardy!, the clues are made difficult on purpose. Furthermore, the clues in Jeopardy! are crafted by a team of people having all information available and are therefore always relevant and true in some way, whereas in LT, they have to be invented on the spot by a human player.

Finally, Pincus et al. [13] present a WordNet-based describer agent that generates clues for clue-guessing games, a project complementary to the implementation of a guesser agent in the LT Challenge.

---

[4] Jeopardy! is an American television game show created by Merv Griffin.

4 Guesser Agent Architecture

In this section, we present the proposed architecture for our guesser agent, as well as the different techniques and experimental configurations that will be used in the results section.

4.1 Basic Architecture

The basic architecture of our guesser agent can be described as follows. For the first incoming hint, the agent calculates the semantic distance between each country in the world and the given hint, using one of the metrics discussed in Section 4.2. Then, the guesser agent selects the top N countries, which were closest to the provided hint, and calculates the distances between the hint and each city in these countries. The idea is to provide the city with the highest score as a guess. If the guess is correct, the game finishes successfully. If the guess is incorrect and a new hint is provided, the distance between this new hint and each country in the world is calculated and added to the score of the previous hints. Unsuccessfully guessed cities are removed from the list of cities, such that they are never guessed twice. The process continues until the guess is correct or the describer runs out of hints. The algorithm is shown in Algorithm 1.

Algorithm 1: Guesser agent main algorithm

input: CountryList

while Success = *false **and** new hints exist* **do**
  Hint ← `GetNewHint()`;
  foreach Country *in* CountryList **do**
    Country.Hint ← `CalcDistance(`Hint,Country`)`;
    Country.Score ← `AggregateDist(`Country`)`;
  end
  BestCountries ← `SortOnDistance(`CountryList,n`)`
  foreach Country *in* BestCountries **do**
    Country.Cities ← `GetCities(`Country`)`;
    foreach City *in* Country.Cities **do**
      City.Hint ← `CalcDistance(`Hint,City`)`;
    end
  end
  BestGuess ← `GetClosestCity(`BestCountries`)`
  Success ← `GuessCity(`BestGuess`)`
end

We have adopted this two-level approach, first pinpointing the countries and then the cities of the highest-ranked countries, for two main reasons. The first reason is that we observed that when humans play this game, many hints are as relevant for the country as for the city itself, with some hints even being more relevant for the country than for the city (such as *tapas* being more relevant for *Spain* than for *Madrid*). The second reason is related to efficiency. Calculating the distance for each hint in combination with all countries in the world requires a much lower number of queries than calculating this for all cities in the world.

4.2 Corpora and Distance Metrics

For calculating the distance between the geographical locations and the hints, we have used two different types of resources with their associated distance metrics. The following subsections will detail the types of resources, which are WordNet and Wikipedia, together with the distance measures.

4.2.1 WordNet

The first resource is WordNet [12], a lexical database linking English nouns, verbs, adjectives and adverbs by their semantic relations, including synonymy, hyperonymy, hyponymy and meronymy. The basic idea here is to exploit these hierarchical relations for measuring the semantic distance between the geographical locations and hints. The specific metric that we use is known as the Jiang-Conrath distance [10], which was found to perform very well when applied to WordNet [3]. The Jiang-Conrath (JC) distance subtracts the sum of the conditional log probabilities (reflecting *information content*) of the two terms from the conditional log probability of their lowest super-ordinate. The lower this number is, the closer the distance between the two terms. The formula of JC distance is presented in Equation (1) where $t1$ and $t2$ represent the two terms and lso stands for their lowest super-ordinate in the database. For words to which multiple synsets are associated, all synsets are tried and the best result is taken.

$$dist_{JC}(t_1, t_2) = \\ 2log(p(lso(t_1, t_2))) - (log(p(t_1)) + log(p(t_2)))) \quad (1)$$

4.2.2 Wikipedia

The second resource that we used consists of all pages of English Wikipedia, as consulted on June 16, 2016. Using the Wikipedia API [5], the guesser agent queries the number of hits in the Wikipedia pages for a hint, a geographical location, and the hint and the geographical location combined. Then, using these hit counts, it employs three different metrics to score the association between the hint and the geographical location.

The first metric, which we call *Normalized Wiki Distance (NWD)*, is based on the Normalized Google Distance [5], but applied to the Wikipedia corpus. The formula is presented in Equation (2). t_1 and t_2 represent the two terms, $c(t)$ stands for the page counts of term t on Wikipedia and N stands for the total number of pages in Wikipedia. A lower NWD indicates a closer association between two terms.

$$NWD(t_1, t_2) = \frac{max(log(c(t_1)), log(c(t_2))) - log(c(t_1, t_2))}{log(N) - min(log(c(t_1)), log(c(t_2)))} \quad (2)$$

The second metric, which we call *Probabilistic Distance* (PD) is based on the ratio between the documents in which both terms occur and the documents in which the most frequent term occurs. When subtracted from 1, the closer this number is to 0, the higher the association between the two terms. The formula of PD is shown in Equation (3).

$$PD(t_1, t_2) = 1 - \frac{log(c(t_1, t_2))}{log(max(c(t_1), c(t_2)))} \quad (3)$$

Finally, we also used the *Pointwise Mutual Information (PMI)* measure [4], a word association metric that is commonly used in the field of computational linguistics for collocation extraction [2]. The formula is given in Equation (4). A higher PMI indicates a higher association of the two terms.

$$PMI(t_1, t_2) = log\frac{c(t_1, t_2)}{c(t_1)c(t_2)} \quad (4)$$

---

[5] https://www.mediawiki.org/wiki/API%3AQuery

4.3 M Most Salient (Famous) Countries

Algorithm 1 takes a list of countries as input. Only the countries in this list will be used in the computations and therefore, only the cities in these countries may be considered as a guess. The most salient (famous) countries are extracted from a ranked list of the countries with the corresponding number of hit counts in Wikipedia. We vary the number of most salient countries throughout the different experiments using a parameter *M*. Choosing a smaller *M* bears the risk of not considering the country of the target city, which will lead to a lost game. When considering countries with too few hit counts (larger *M*) on the other hand, the distance metrics described in the previous subsections may yield unexpected results due to data sparseness.

4.4 N Top Scoring (Best) Countries

In our guesser agent algorithm (see Algorithm 1), we first calculate the distance between the hints and the different countries from the provided country list. Then, for the *N* top scoring countries (i.e. having the closest semantic distances), we calculate the distances between their cities and the hints. So, only cities of the *N* best countries are considered as guesses. This parameter *N* regulates how much weight is given to the association between countries and the hints (instead of the cities).

5 Experiments and Results

We have evaluated our guesser agent on a set of 82 real-world games provided by ESSENCE. This section presents the cross categorical experiments and their results.

5.1 Experimental Setup

We have run several experiments varying the parameters *M* and *N* as discussed in the previous section. In the experiments, *M* takes the values 0, 10, 20, 30, 40, 50 and 60. The 0 value means that the country (salience) restriction is not active and that all countries in the world are considered. The parameter *N* takes the values 1, 2, 3, 4, 5, 10, 15, 20, 25, 50, 100 and ALL. In the case of ALL, all of the cities in all *M* countries are considered. The naming of the experiments follows the same abbreviation, which can be formalized as *FMBN*. In this abbreviation, *F* refers to Famous countries as mentioned in Section 4.3 and *B* refers to Best scoring countries as mentioned in Section 4.4. The parameters *M* and *N* in the *FMBN* abbreviation take the aforementioned values and hence we have 84 experiments for each metric. It should be noted that when *M=0*, the abbreviation is represented as *BN*, rather than *F0BN*.

5.2 Results of the experiments using WordNet

5.2.1 Jiang-Conrath Distance (JCD)

In this set of experiments, we have used the Jiang-Conrath Distance on WordNet to calculate the semantic distance between the hints and the geographical locations. The results of the 84 experiments suggest that the use of the 50 most salient (famous) countries in combination with a small selection (3-5) of best scoring countries yields the best results. Table 1 displays the top 5 configurations in terms of accuracy and in terms of successful games that were solved by the guesser agent using fewer number of guesses than the human counterpart. The top configuration for this set of experiments is F50B3 with an accuracy of 6,09% and a *faster guessing performance* of 80%.

Table 1. Results of experiments using Jiang-Conrath Distance on WordNet

Experiment Type	Successful Guesses	Faster Guesses	Accuracy (%)	Relative Faster Guessing Performance (%)
F50B3	5	4	6.09	80
F50B5	5	3	6.09	60
F50B4	4	3	4.87	75
F10B15	4	2	4.87	50
F50B15	3	3	3.65	100

Table 2. Top 5 results of experiments using Normalized Wiki Distance on Wikipedia

Experiment Type	Successful Guesses	Faster Guesses	Accuracy (%)	Relative Faster Guessing Performance (%)
F30B10	16	9	19.51	56.25
F30B15	15	9	18.29	60
F20B15	15	8	18.29	53.33
F30B5	15	6	18.29	40
F60B10	15	4	18.29	26.66

Table 3. Results of experiments using Probabilistic Distance on Wikipedia

Experiment Type	Successful Guesses	Faster Guesses	Accuracy (%)	Relative Faster Guessing Performance (%)
F50B2	18	10	21.95	55.55
F50B15	18	8	21.95	44.44
B2	17	9	20.73	52.94
B3	17	9	20.73	52.94
F60B2	17	9	20.73	52.94

Table 4. Results of experiments using Pointwise Mutual Information Measure on Wikipedia

Experiment Type	Successful Guesses	Faster Guesses	Accuracy (%)	Relative Faster Guessing Performance (%)
F20B10	19	13	23.17	68.42
F30B25	17	10	20.73	58.82
F20B15	17	9	20.73	52.94
F30B30	16	11	19.51	68.75
F30B30	16	11	19.51	68.75

5.3 Results of the experiments using Wikipedia

In the following 3 sets of experiments, we have used the English Wikipedia as a corpus for calculating the semantic distance between the hints and the geographical locations.

5.3.1 Normalized Wiki Distance (NWD)

In this set of experiments, we have used the Normalized Wiki Distance as formulated in Equation (2). The results of the 84 experiments show that the use of the 30 most salient (famous) countries in combination with a medium selection (5-15) of best scoring countries yields the best results. The highest result, yielded by the F30B10 experiment, shows an accuracy of 19.51% and a relative faster guessing performance of 56.25%. Table 2 displays the results of the 5 most accurate experiments in this series.

5.3.2 Probabilistic Distance (PD)

For this series of experiments, we have used the Probabilistic Distance metric as formulated in Equation (3). Similar to the results of the experiments using WordNet, the use of the 50 most salient (famous) countries in combination with a small selection (2-15) of best scoring countries gives the best results, with F50B2 topping the list with an accuracy 21.95% and a faster guessing performance of 55,55%. Table 3 displays the 5 best-scoring configurations.

5.3.3 PMI Distance

In this set of experiments, we have used the Pointwise Mutual Information measure as formulated in Equation (4). The results are in agreement with the majority of the previously recorded results and they show that the use of 20 most salient (famous) countries in combination with a medium selection (10-30) of best scoring countries gives the best success accuracy. The best-scoring configuration here is F20B10 with an accuracy of 23.17% and a faster guessing performance of 68.42%. The 5 best-scoring configurations are shown in Table 4.

5.4 Summary of Results

Overall, we have performed 84 experiments for each resource (i.e. WordNet and Wikipedia) and the associated distance measures. In total, this makes 336 different experiments (i.e. configurations using the M and N parameters). Table 5 summarizes the success rates of both WordNet and Wikipedia and all associated distance measures. According to the results, the maximum accuracy (23.17%) was reached using the PMI distance measure on the Wikipedia corpus. On the other hand, the highest mean of the accuracy throughout the different configurations was recorded for the PD measure, on the Wikipedia corpus as well.

In this section, we have only presented the best scoring configurations, but for the sake of completeness, the results of all experiments and configurations are visualized in Figure 2. This figure clearly visualizes which configurations (M and N values) are optimal for the different metrics.

6 Discussion

The results of hundreds of experiments demonstrate that using the Wikipedia corpus yields substantially better results than using WordNet as a resource for semantic distance calculation in our guesser agent. This might be due to the very nature of the word associations that the Taboo game requires. The format of the game already rules out the best clues, i.e. the most closely associated words, from the set of hints. This means that there is always a considerable distance between the two terms. WordNet has difficulties with this, as the annotated hierarchical relations are only made between terms that are semantically very closely associated, and paths that link hints to locations might not exist, or might not be very meaningful due to their length (of the link chain). The Wikipedia approach seems to be much more robust against this. Even if the hints are not that closely related to each other, there almost always exists documents on which hint and geographical location occur together. For this task, the size of Wikipedia has the upper hand over the precision annotation of WordNet.

Throughout the different configurations in our experiments, we observed that limiting the number of countries in the country list can improve the performance. As we mentioned earlier, this has the risk that some of the games will fail because their target location falls outside the list. On the other hand, it has the advantage that countries for which the hit counts are sparser do not influence the results too much. The results show that the NWD and PMI metrics benefit from limiting the number of countries to 20 or 30, whereas PD seems to be less disturbed by the sparseness effect. Indeed, PD benefits from configurations having higher numbers such as 50, 60 and ALL.

Once the countries have been ranked based on the metric, we also limited the number of countries for which the cities were considered (the parameter N). This also influences the performance differently from one distance measure to another. The PMI and NWD metrics score the best with higher N values (10-30), whereas the PD metric scores equally well with high (15) and low (2-3) N values. This indicates that the PD measure performs better at ranking the countries based on the hints.

7 Future Work

The research described in this paper is only a first step towards solving the ESSENCE LT Challenge. Using well-established word association techniques and freely available corpora, we aimed to establish a baseline to which future approaches can be compared. A first, promising extension of our guesser agent would be to equip it with machinery for resolving hints that are relative to the previous answer (e.g. *close*, or *north*). Another extension, which is closely related to the diverse nature of the real-world dataset, would be to model the associative behavior of the individual describers. This is possible, as with each game in the challenge, the ID of the human describer is provided. This way, the diversity in associations and game-playing behavior of the different players could be taken into account in order to improve the number of correct guesses. Further improvements could include investigating how lemmatization of the hints influences the accuracy of the guesser agent, as well as to explore ways to fuse the different metrics that were described in this paper.

8 Conclusion

We have proposed a semantic distance based architecture for a guesser agent for the Essence Location Taboo Challenge. The proposed architecture employs a two-step approach, narrowing down

Table 5. Summary of accuracy results for each resource and distance measure

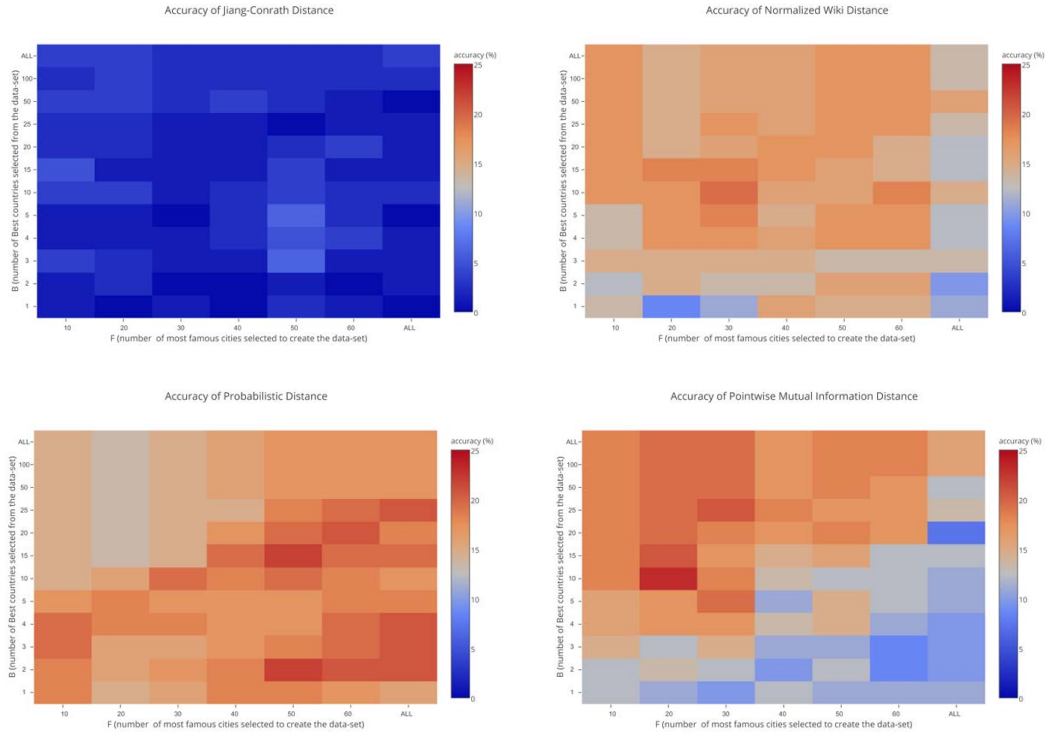| Corpora	Distance Metric	Accuracy (%)			
		Maximum	Minimum	Mean	Standard Deviation
WordNet	Jiang-Conrath	6.09	0	2.06	1.35
	NWD	19.51	8.53	15.36	2.05
Wikipedia	Probabilistic	21.95	13.41	17.16	2.18
	PMI	**23.17**	7.31	15.15	3.62

Figure 2. Results of all experiments. The X axis represents the M parameter (most salient countries) and the Y axis represents the N parameter (cities of N best countries considered). The red-blue scale indicates the accuracy of the experiment.

the geographical area of the guess first to the country and then to the city. We have explored different resources and metrics for measuring the diverse associations between the hints and the geographical locations that were made by human players with different backgrounds. The highest score with 23.17% accuracy and 68.42% of faster guessing performance was achieved with the PMI measure applied to the Wikipedia corpus. Although this research is only a first step to model the diversity in word associations that individual humans exhibit, it can serve as a strong baseline to which future attempts to solve the ESSENCE LT Challenge can be compared.

ACKNOWLEDGEMENTS

REFERENCES

[1] Kemo Adrian, Khuyagbaatar Batsuren, Nicola Bova, Thomas Brochhagen, Paula chocron, Paul Van Eecke, Mercedes Huertas-Miguelanez, Mladjan Jovanovic, Tania Marques, Julian Schloder, and Aimilios Vourliotakis, 'Taboo challenge', Technical report, Essence Marie Curie Initial Training Network, (06 2015).

[2] Gerlof Bouma, 'Normalized (pointwise) mutual information in collocation extraction', *Proceedings of GSCL*, 31–40, (2009).

[3] Alexander Budanitsky and Graeme Hirst, 'Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures', in *Workshop on WordNet and Other Lexical Resources*, volume 2, pp. 2–2, (2001).

[4] Kenneth Ward Church and Patrick Hanks, 'Word association norms, mutual information, and lexicography', *Comput. Linguist.*, **16**(1), 22–29, (March 1990).

[5] Rudi L Cilibrasi and Paul Vitanyi, 'The google similarity distance', *Knowledge and Data Engineering, IEEE Transactions on*, **19**(3), 370–383, (2007).

[6] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman, 'Indexing by latent semantic analysis', *Journal of the American society for information science*, **41**(6), 391, (1990).

[7] David Ferrucci, 'Build watson: an overview of deepqa for the jeopardy!

challenge', in *Proceedings of the 19th international conference on Parallel architectures and compilation techniques*, pp. 1–2. ACM, (2010).

[8] David Ferrucci, 'Build watson: an overview of deepqa for the jeopardy! challenge', in *Proceedings of the 19th international conference on Parallel architectures and compilation techniques*, pp. 1–2. ACM, (2010).

[9] Don Heath, David Norton, Eric Ringger, and Daniela Ventura, 'Semantic models as a combination of free association norms and corpus-based correlations', in *Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on*, pp. 48–55. IEEE, (2013).

[10] Jay J. Jiang and David W. Conrath, 'Semantic similarity based on corpus statistics and lexical taxonomy', in *Proc of 10th International Conference on Research in Computational Linguistics, ROCLING97*, (1997).

[11] Kevin Lund and Curt Burgess, 'Producing high-dimensional semantic spaces from lexical co-occurrence', *Behavior Research Methods, Instruments, & Computers*, **28**(2), 203–208, (1996).

[12] George A Miller, 'Wordnet: a lexical database for english', *Communications of the ACM*, **38**(11), 39–41, (1995).

[13] Eli Pincus, David DeVault, and David Traum, 'Mr. clue-a virtual agent that can play wordguessing games', in *Proc. of the 3rd Workshop on Games and NLP (GAMNLP), Raleigh, North Carolina, USA*, (2014).