Modelling the Generation and Retrieval of Word Associations with Word Embeddings

A Case Study for a Guesser Agent in the Location Taboo Game

B.Sc. Thesis Abstract

Verna Dankers, Aysenur Bilgin, Raquel Fernández

Institute for Logic, Language and Computation, University of Amsterdam

1 Introduction

When words occur in close proximity frequently, an associative link is formed in the observer (Clark, 1970). Distributional Semantic Models (DSM) can be used for the artificial acquisition of associative links (Heath et al., 2013) and learn word representations from word co-occurrence patterns in a corpus.

In this thesis, a method for modelling word associations is developed by creating an Artificial Guesser Agent (AGA) for the Location Taboo Game (LTG).[1] In this word-guessing game, simple textual clues about a well-known target city are provided, and the AGA should guess the city. The AGA is evaluated through games that were successfully played by humans and should be able to mimic the associations that humans have with cities. The proposed AGA employs a semantic vector space that is created using context-predicting DSMs from a tailored corpus about travel destinations. The training of the vector space uses a novel annotation method to strengthen the associative link between the cities and their descriptions in the corpus. Four different game-playing strategies are implemented to retrieve guesses from the vector space, and several lists of candidate cities are compiled to only consider sufficiently relevant places.

2 Approach

Semantic Vector Space A tailored corpus was constructed from Wikipedia and Wikivoyage pages entitled with a country or city name and Wikivoyage outlinks to other Wikivoyage pages. A targeted annotation method is proposed to make the association between a geographical location and the content of its corresponding pages explicit for the DSM. For the pages entitled with a country or city name, the name is inserted in every sentence, for example: *'pizza traditionally eaten locally pasta* ***Verona*** *dishes feature widely restaurant menus'*.

The context-predicting DSMs employed are Continuous Bag of Words and Skip-Gram, combined with the Hierarchical Softmax and Negative Sampling algorithms (Mikolov et al., 2013a, 2013b). The similarity metrics used are Chebyshev, Correlation, Cosine and Tanimoto.

Candidate Cities Three lists of 500 candidate cities were constructed, according to the occurrence counts from the tailored corpus (list 1), the occurrence counts from a Google News dataset[2] (list 2) and the number of visits as listed on NomadList[3] (list 3). A fourth list was created containing only the 120 cities present in example games to investigate the AGA's best-case performance.

Game-Playing Strategy Four game-playing strategies were implemented: Using the cumulative similarity between vectors of clue words and vectors of cities (strategy 1). Eliminating inconsistent

[1] This study is partially supported by the Marie Curie Initial Training Network (ITN) ESSENCE, grant agreement no. 607062.

The LTG has been designed by ESSENCE: https://www.essence-network.com/challenge/.

[2] The vectors are available at: https://code.google.com/archive/p/word2vec/.

[3] https://nomadlist.com

cities across game iterations (strategy 2). Narrowing down to a set of countries and choosing from the cities of those countries (strategy 3). Summing the vectors of all clue words to guess the nearest city in the vector space (strategy 4).

3 Experiments and Results

We have conducted several experiments to investigate the impact of hyper-parameter configuration for the DSMs and the AGA using 162 example games. Additionally, forty games were set aside for the final evaluation. Strategy 1 was the most appropriate game-playing strategy. The Tanimoto, Correlation and Cosine similarity metrics all resulted in very similar performance. Out of the first three lists of candidate cities, list 3 resulted in the highest performance. The targeted corpus annotation had a significant effect for all models (two-sided relative t-test, $p < 0.001$).

For the performance evaluation of the AGA three evaluation metrics were used: the game score, the accuracy and the faster-guessing performance (FGP). Skip-Gram Hierarchical Softmax (SGHS) was the best performing algorithm, for which the results are shown in Table 1, together with the results of the baseline architecture introduced by Adrian et al. (2016).

Table 1: The performance evaluation for SGHS and the baseline architecture. The game score is the number of submitted guesses, plus a penalty of 5 per unsuccessful game. The FGP represents the percentage of successful games for which the AGA submitted fewer guesses than the human.

Corpus	**Algorithm**	**Game Score**	**Accuracy (%)**	**FGP (%)**
Unannotated	SGHS	242	25.00	**50.00**
Annotated	SGHS	**239**	**27.50**	27.27
-	Baseline Architecture	290	5.00	0.00

4 Conclusion and Future Work

In this thesis, an AGA architecture is presented, which uses context-predicting DSMs to infer word associations from a tailored corpus and employs different game-playing strategies for the LTG. The architecture is an improvement compared to the baseline of Adrian et al. (2016) and can guess target cities with up to 27.50% accuracy. The presented method for the generation and retrieval of word associations could be generalised for different domains or adapted for different tasks. Targeted corpus annotation is proposed to amplify the strength of the associations implicitly present in the corpus and it significantly improves the performance in the LTG.

Regarding future work, multiple types of resources can be combined to improve the corpus content and the list of candidate cities. Furthermore, more complex architectures could be created, such as a multi-agent system or a system that can better interpret multi-word clues.

References

Adrian, K., Bilgin, A. & Van Eecke, P. (2016). A semantic distance based architecture for a guesser agent in ESSENCEs location taboo challenge. *DIVERSITY@ ECAI 2016*, 33–39.

Clark, H. H. (1970). Word associations and linguistic theory. *New horizons in linguistics*, *1*, 271–286.

Heath, D., Norton, D., Ringger, E. & Ventura, D. (2013). Semantic models as a combination of free association norms and corpus-based correlations. In *Seventh international conference on semantic computing* (pp. 48–55). doi: 10.1109/ICSC.2013.18

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013a). Efficient estimation of word representations in vector space. *In Proceedings of Workshop at the ICLR*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).