Katarzyna Marszalek-Kowalewska
Institute of Linguistics
Adam Mickiewicz University, Poznan, Poland

Title: Disambiguation of Persian homographs with word2vec

All natural languages contain words that can mean different things depending on different contexts. Lexical ambiguity - the fact that a word can have more then one meaning - has become one of the main challenges in understanding natural language. The correct sense of an ambiguous word can be determined based on the context where it occurs. While most of the time humans do not even think about the ambiguities of a given language, machines need to process unstructured textual information and transform them into data structures which must be analysed in order to determine the underlying meaning. Lexical ambiguity is inherent to all natural languages and Farsi is no exception here. In fact, Farsi is greatly ambiguous at the levels of both polysemy and homonymy. The latter case, and to be more precise, the issue of homography is the main problem addressed here. The very fact that the Persian writing system often omits diacritics generates lots of ambiguity for computer processing of the Persian language, e.g. the form کرم can mean 'worm', 'cream', 'chromium', 'generosity' and 'creamy colour'.

Assigning the most appropriate meaning to an ambiguous word is known as Word Sense Disambiguation (WSD). WSD is a fundamental task in computational lexical semantics and one of the oldest tasks in Natural Language Processing (NLP) and Artificial Intelligence (AI). There are four main approaches to WSD: supervised approach (e.g. Zhong and Ng 2010, Shen et al. 2013), unsupervised (e.g. Agirre et al. 2006, Di Marco and Navigli 2013), semi-supervised (e.g. Mihalcea and Faruque 2004) and finally knowledge-based approach (e.g. Ponzetto and Navigli 2010, Agirre et al. 2014). Lot of work has been done in the area of WSD for the English language, the Persian language in that respect is unfortunately not so much researched. The main contributions to word sense disambiguation for Persian can be found in the works by Hamidi, Borji and Ghidary (2007), Soltani and Faili (2010), Rekabsaz et al. (2016), Makki and Homayounpour (2008), Sarrafzadeh and Yakovets (2015). The main purpose here is to present word embeddings used as a method of WSD for the homographs of the Persian language.

Word embeddings are low dimensional representations of a natural language words as real-valued vectors. They are able to capture important semantic and syntactic features of words in a compact manner. The model presented here focuses on word vectors as described by Miklov et al. (2013). Vector representations of words have proven useful in NLP tasks due to their ability to efficiently model complex semantic and syntactic word relationships and have therefore been increasingly used by many researchers, e.g. Pennington et al. (2014), Trask et al. (2015).

Here, we'd like to present the possible application of word embeddings in the form of word2vec approach for the disambiguation of Persian homographs. Firstly, the word2vec model and related works will be described. Then, a study of 10 ambiguous Persian homographs is to be presented. The results obtained with this approach will be compared with other approaches to word sense disambiguation for the Persian language.

References

Agirre Eneko, David Martínez, Oier López de Lacalle, and Aitor Soroa. 2006. *Two graph-based algorithms for state-of-the-art WSD*. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06).

Agirre Eneko, Oier López de Lacalle, and Aitor Soroa. 2014. *Random walks for knowledge-based word sense disambiguation*. In Computational Linguistics 40(1).

Di Marco Antonio, Roberto Navigli. 2013. *Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction*. In Computational Linguistics, 39(4).

Hamidi Mandana, Ali Borji and Saeed Shiry Ghidary, 2007. *Persian Word Sense Disambiguation.* In Proceeding of 15th Iranian Conference of Electrical and Electronics Engineers (ICEE 2007).

Makki Raheleh and Mohammad Mehdi Homayounpour. 2008. *Word Sense Disambiguation of Farsi Homographs Using Thesaurus and Corpus*. In Proceedings of the 6th international conference on Advances in Natural Language Processing (GoTAL '08).

Miklov Tomas, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. *Efficient Estimation of Word Representations in Vector Space.*

Mihalcea Rada and Ehsanul Faruque. 2004. *Sense- learner: Minimally supervised word sense disam- biguation for all words in open text*. In Proceedings of ACL/SIGLEX Senseval-3.

Pennington Jeffrey, Richard Socher, and Christopher D. Manning. 2014. *GloVe: Global Vectors for Word Representation*. In Proceedings of the EMNLP 2014.

Ponzetto Simone Paolo and Roberto Navigli. 2010. *Knowledge-rich Word Sense Disambiguation rivaling supervised systems*. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10).

Rekabsaz Navid, Serwah Sabetghadam, Mihai Lupu, Linda Andersson and Allan Hanbury. 2016. *Standard Test Collection for English-Persian Cross-Lingual Word Sense Disambiguation*.

Sarrafzadeh Bahareh and Nikolay Yakovets. 2015. *Two Novel Approaches for Persian Word Sense Disambiguation* (https://wiki.eecs.yorku.ca/course_archive/2014-15/W/6339/_media/clreport_final.pdf).

Shen Hui, Razvan C. Bunescu, Rada Mihalcea. 2013. *Coarse to Fine Grained Sense Disambiguation in Wikipedia*. In *SEM@NAACL-HLT 2013.

Soltani Mahmood and Heshaam Faili. 2010. *A statistical approach on Persian Word Sense disambiguation*. In Proceedings of the 7th International Conferences on Informatics and System.

Trask Andrew, Phil Michalak and John Liu. 2015. *sense2vec - A Fast and Accurate Method for Word Sense Disambiguation In Neural Word Embeddings*.

Zhong Zhi and Hwee Tou Ng. 2010. *It makes sense: a wide-coverage word sense disambiguation system for free text*. In Proceedings of the ACL 2010 System Demonstrations (ACLDemos '10).