All natural languages contain words that can mean different things depending on different contexts. Lexical ambiguity - the fact that a word can have more than 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 كر م 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