KOOB EPUB READER : CONTEXT-SENSITIVE LOOKUP

NLP- WORD SENSE DISAMBIGUATION TECHNICAL REPORT

PROFESSOR ANAGHA KULKARNI
Submitted By: SWATI KOHLI
CSC 820 Natural Language Technologies

INDEX

Introduction

Studies report that at least 32% of the words used in English text are ambiguous and this estimate is considered just a lower bound (Britton, 1978). Words with multiple meanings give English a linguistic richness, but they can also create doubt or confusion especially amongst non-native speakers (Staff, 2013). So, what does one do if one comes across such a situation while reading a book online? In this digitized age of e-reading, although most existing online reading applications provide an in-app feature of looking up the word meaning, they do not provide the context-sensitive sense of the polysemous word (having many senses) that becomes a challenge and dissuasion for a reader with English as a second language (ESL). This complexity is solved by word sense disambiguation (WSD)- a technique in natural language technologies(NLT) to find the exact sense of the ambiguous word in a specific context. WSD is implemented in many areas such as machine translation, speech recognition, information extraction, information retrieval, etc. Therefore, the project aims at providing the best context-based sense of the looked-up word with the aid of disambiguation by refining and structuring the information (meaning) displayed in ranked order of relevance for the reader. This is a case of information retrieval and is attributed to Koob ePub Reader- an online reading application intending to incorporate the solution for the end-users. Figure 1. demonstrates that with WSD techniques, the target word kiwi is a fruit and not a flightless New Zealand bird.

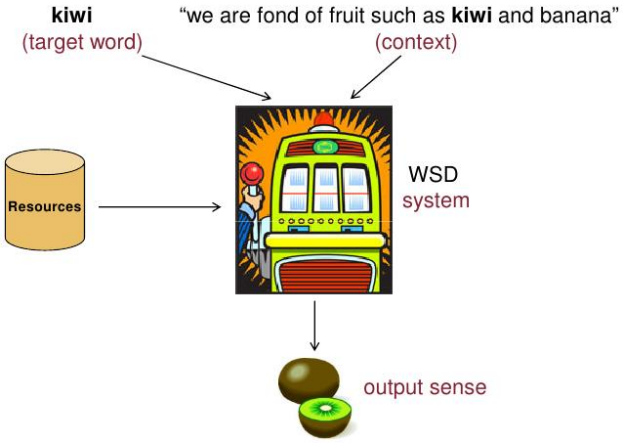

Figure 1. WSD Example

This paper aims to propose and develop several approaches using machine learning algorithms that may be integrated into the application. A comparison between all the approaches is presented based on two evaluation metrics- accuracy and mean reciprocal rank to make a recommendation. The input data comprises of about 270 (predominantly ambiguous) words along with its context- sentence and paragraph where the word appears, queried from the book- Reading with Patrick. The dictionary source for the meanings is Miriam Webster. The data is labeled based on the Subject Matter Expert (SME)- an English Professor from SFSU, who voted the best sense of the word based on the surrounding context. This label is taken as 'ground truth' for the metric evaluation.

Related Work

Several approaches and tools have been developed as context-sensitive dictionary lookups in the realm of academic or commercial projects, or as joint work. Some of the engines are: GLOSSER-RuG prototype targeting Dutch students (Nerbonne et al., 1999), DEFI Matcher (Michiels, 1998; Michiels, 2000) is a sophisticated program for matching text and dictionary entries, MoBiDic/MoBiMouse (Pro´sze´ky, 1998; Pro´sze´ky 2002), is a commercial tool developed by MorphoLogic, Benedict – The New Intelligent Dictionary (Löfberg et al., 2004),

etc. These tools retrieve the base form of the word complemented with part-of-speech (POS) tagging that disambiguates the lexical category, as a word form can belong to different categories (for instance, the word *bear* can be either a noun or a verb) and thereafter perform a morphological analysis to find a match in the dictionary entry list with certain heuristics and criteria (e.g., the longest-match or the closest-match criterion) for ranking competing candidates.

Some academic journals also provide techniques and insight on the WSD. For example, authors Sharma, P., & Joshi, N. (2019, January) utilize Lesk algorithm for word sense disambiguation in Hindi language using word knowledge (WordNet) as a source of glosses for the approach. They take a corpus containing 3000 ambiguous sentences and correctly identify 2143 of them providing an overall accuracy of 71.43% and error 28.57% for the system. Further, Basile, P., Caputo, A., & Semeraro, G. enhance the approach by building distributional semantic model that computes the gloss-context overlap by building the vector representation for each gloss associated with the senses of the word and the context.

Other approaches are also employed. For example, Ojha, N. (2019, February) use Naïve Bayes algorithm based on the supervised approach. Sasaki, M., Komiya, K., & Shinnou, H. (2014) use collocation dictionary with SVM algorithm approach. They identify the sense of idiom or common phrase containing a target word before the statistical WSD method is applied by capturing the context information. Experiments compare with baseline results and show that efficiency with the proposed method improved. Moverover, Orkphol, K., & Yang, W. (2019) use word representations through pre-trained word embedding approach (Word2Vec) towards solving WSD on cosine similarity metrics that provides good results.

Project Approaches

The project attempts to leverage three approaches (rather than one) studied in literature research with different similarity/distance measures criteria to thereafter perform a comparative analysis towards approach efficacy based on the evaluation metrics (see Figure 2).The overall methodology is illustrated in the workflow diagram (see Figure 3). Since the sentence as the context proved to be insufficient for the task as concluded in the initial analysis, the paragraph is taken as the context with the target word. **Pre-processing** steps are applied for all the approaches (see Figure 3) like tokenization, stop word removal, conversion to lower case, lemmatization and stemming on both the context and the meanings. Thereafter, the traditional dictionary-based method- **Lesk** is taken as a baseline approach that computes similarity by counting overlapping words between the context and sense definitions that must match exactly. Additionally, POS (Part of speech) tagging is added in both the context and the meanings to add more weight or preference towards the correct meaning through overlap.

Figure 2. Project Approaches

The second approach is developed by taking **context similarity with the Tf-Idf technique**. This technique scores the similarity between the context and multiple word meanings based on the importance of words they commonly share. In simple terms, this method is based on the frequency of occurrence of shared words. It is observed that appropriate meaning will have the same words appearing in the surroundings as in the context. The context with the target word and the meanings are mapped as a point in a multi-dimensional space where the number of dimensions is equal to the number of unique words. The cosine angle between these points (denoted by vectors) is used as a metric of similarity (cosine similarity) where the dot product of the two vectors is the cosine angle between the points denoted by vectors. Two more criteria are taken to compare the model performance – Euclidean distance and Manhattan distance. This method is simple, fast and effective with both data requiring to have exact word matches.

Finally, another perspective to the approach is that similarity should be computed based on **semantic similarity**, i.e. how words are related (rather than overlapping) by representing the context and sense definitions on a vector space model and analyzing distributional semantic relationships among them. Therefore, pre-trained word embeddings for the input are employed as the third approach using three popular pre-trained word embeddings-Word2Vec, FastText, and GloVe. These models represent words on a fixed-size vector space model through techniques like the skip-gram, continuous bag-of-words (CBOW) models etc. and further their performance is compared by employing the relevant criterion and metrics. This method effectively captures semantic and syntactic word similarities from a huge corpus of text (forming vocabulary). Through this approach context sense vector and sense definition vectors are constructed (averaging all the word representations separately in both inputs) and then, given each word sense a score using cosine similarity, computes the similarity between those sentences. The pre-trained word embeddings are based on a corpus of vocabulary that require to be downloaded that takes some time to load first time, however, thereafter the operations are faster. This approach is good with low word overlap because it does not require an exact match (e.g. mortgage with bank are placed closer in the space).

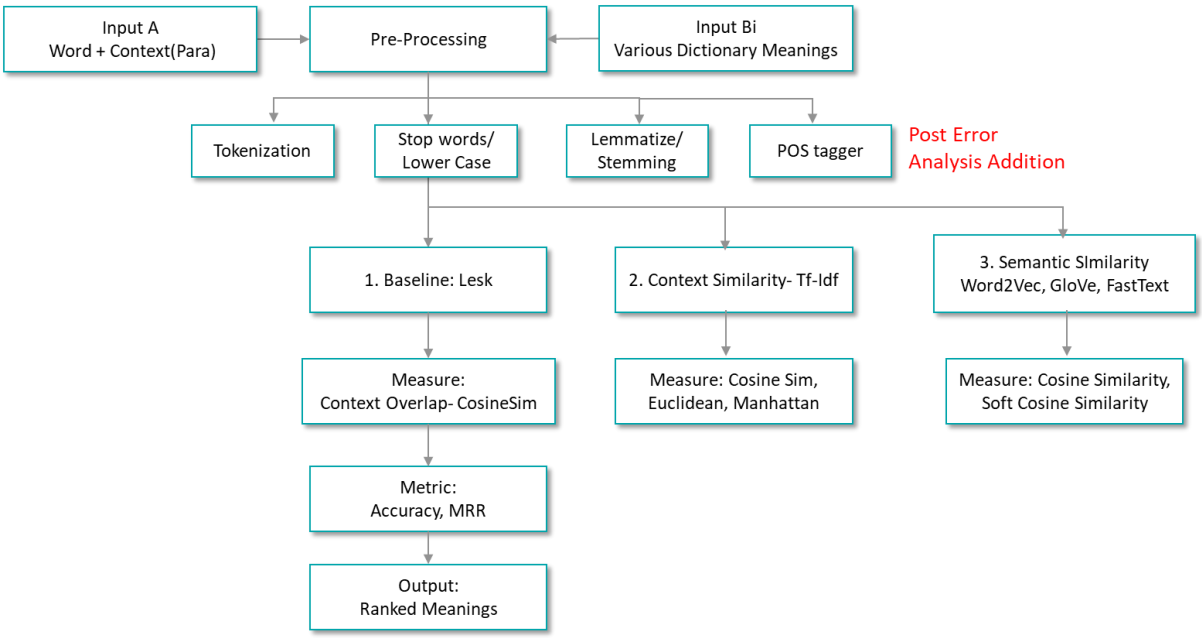

Figure 3. Project Methodology

Figure 3 shows two inputs:
1. Input A = Target Word + Paragraph
2. Input Bi = POS & Meanings from dictionary, where i is each different meaning for the target word

Results and Analysis

The models are measured on several criterias for a deeper analysis. The same are discussed point-wise for each of the approaches. Further, all the approaches are measured on two metrics- Accuracy and Mean Reciprocal Rank (MRR) to gauge the capability and performance of the methods. Accuracy is a statistical measure of how well a binary classification test correctly identifies a condition. In other words, accuracy is the proportion of correct predictions among the total number of cases examined (Wikipedia). It, however, takes the hard right or wrong and disregards the ranking order. To overcome this limitation and analyse how well the model is performing, another metric is taken for the project, that is, Mean Reciprocal Rank. MRR is valuable and suitable for the project because it takes into account the order of correctness as well. In other words, it calculates how far the predicted result is from the truth. Following is the formula for it; where N is the total data points and i is the rank of the predicted response compared to the true response.

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i}$$

1. Lesk is computed on cosine similarity measure along with POS tagged word added to revise and improve the initial results. The results are analysed and compared later.

2. For Tf-Idf approach, the results show that Euclidean distance performs better overall, even though MRR for Cosine Similarity is high. This is because CS gives a zero score if there is no overlap and if more than one score is zero then it returns the best meaning as smallest 'Meaning ID' assigned that also tends to be the teacher voted meaning usually. As illustrated in Figure 4, all the meanings are assigned a score of 0 and smallest meaning id 3878 is returned as best which is misleading and therefore this condition is taken care of with an if statement that results in a very low accuracy. However, since distance matrix does not give a score of zero, they give a better picture of accuracy.

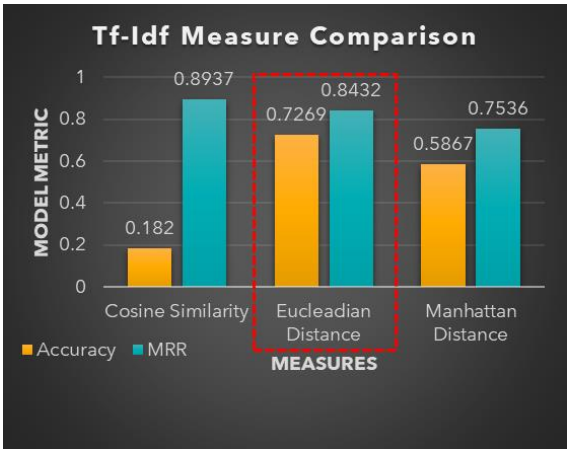

Figure 3. Tf-Idf results on metrics

Result Analysis: Add another condition

Lookup word: meticulousness

	id	meaning	score	isTeacher
0	3878	very careful about doing something in an extre...	0.0	1.0
1	6366	marked by extreme or excessive care in the con...	0.0	0.0

Figure 4. Analysis of Data based on Results

3. Semantic similarity approach is taken for multiple word embeddings across two criterias- cosine similarity and soft cosine similarity. The difference is that cosine similarity considers the vector space model (VSM) features as independent or completely different, while the soft cosine similarity measure considers the similarity of features in VSM. For example, onion and potato will not have a zero score because of no overlap, rather have some score based on the fact that both are root vegetables and occur closely in the multi-dimensional vector space. Figure 5. shows that semantic similarities measured with GloVe embeddings yield best results out of the three pre-trained word embedding models and it also generally performs best on cosine similarity.

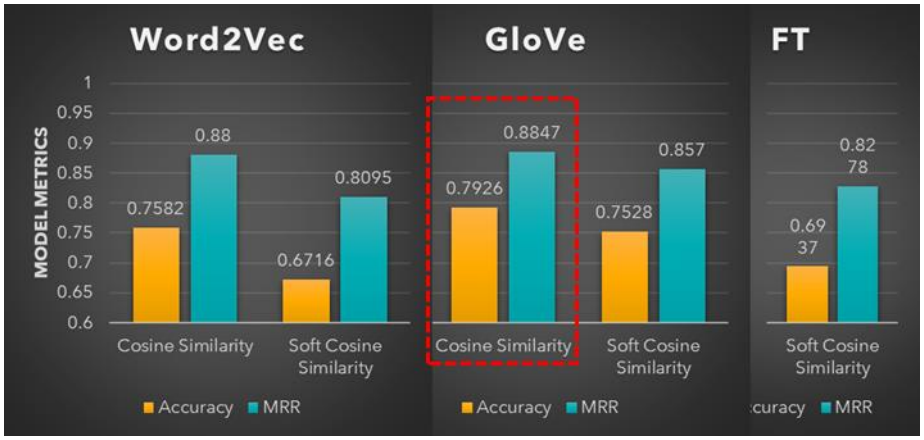

Figure 5. Results on pre-trained word-embedding approach

Upon assessment, it is seen that the results are good and more or less comparable across all approaches however, semantic similarity approach gives the best result. Overall, GloVe word embeddings provide best performance on both metrics, accuracy and MRR, as 75.28% and 85.7% respectively. Also, the Tf-idf approach, comes a close second. Therefore, these both are recommended for implementation based on their complexity and deployability.

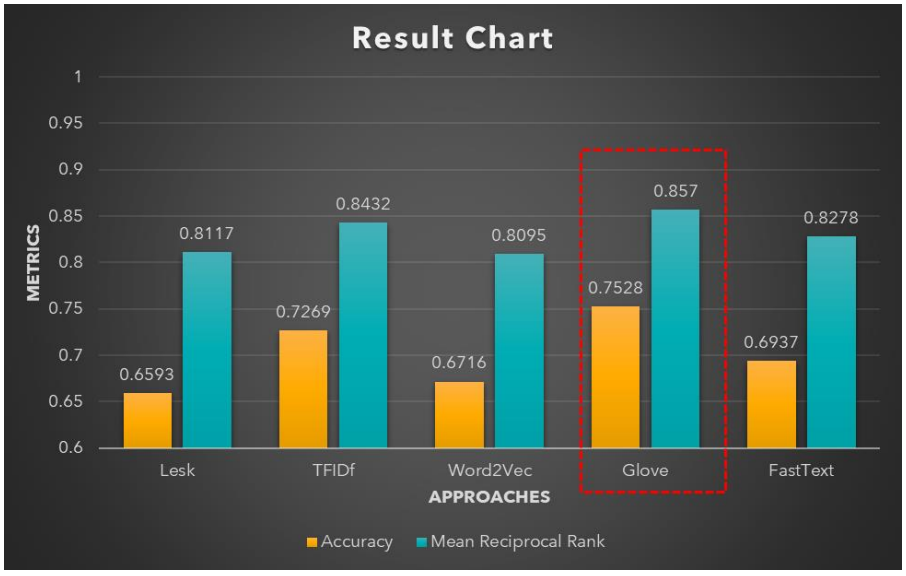

Figure 6. Comparative Results of all the Project approaches

Contribution

This project contributes to the realm of digitized reading and learning by expanding the knowledge of the ESL readers providing them correct context-sensitive meanings of ambiguous words they come across while reading a book through NLP and machine learning techniques: a feature that is currently not available in the industry. Further, the project aims to make the end-user aware of not only the best but also other possible meanings through the hierarchy of relevance (ranked order of best senses) to deepen their knowledge. On the oter hand, the project contributes towards the app development as it illustrates and proves the potential of WSD systems by discussing the technical aspects of how various approaches can be integrated into the eBook app. Further, it provides options and flexibility to the app developer as the project does not rely on a single approach, rather explores different deployable approaches eliciting their requirements, pros and cons. Further, the approaches are measured on different similarity and distance measure criterias to deeply analyse how different parameters affect the performance. Finally, since the performance is judged on two metrics it gives a good grasp of the model capabilities in a holistic way. Figure 7a & 7b illustrate an example of the input query and the output generated for the ambiguous word.

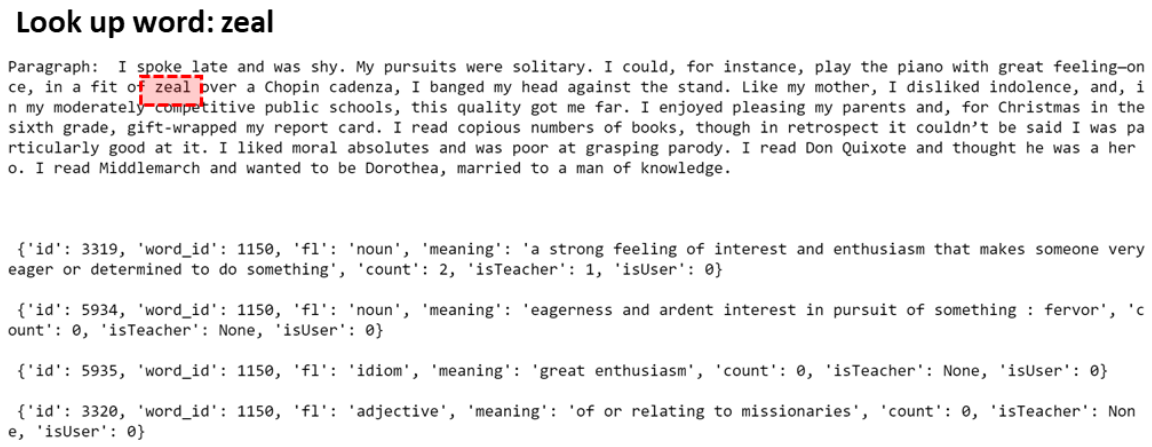

Figure 7a. Example of word look-up 'zeal' in the context with dictionary meanings

	id	pos	meaning	score	isTeacher	pred_rank
0	3319	noun	a strong feeling of interest and enthusiasm th...	0.893703	1.0	1.0
1	5934	noun	eagerness and ardent interest in pursuit of so...	0.717988	0.0	2.0
2	5935	idiom	great enthusiasm	0.704320	0.0	3.0
3	3320	adjective	of or relating to missionaries	0.442030	0.0	4.0

Figure 7b. Example output generated for the word look-up 'zeal' in the context as given in Figure 7a.

Learnings

The project expanded my knowledge and understanding of concepts and implementation both broadly and deeply. Firstly, through the project, it is seen that Python programming is much useful and facilitated in implementing the different powerful libraries like NLTK, Gensim, Sk-

Learn, etc. for modeling the project tasks. Further, I learnt various approaches through research, analysis, and brainstorming sessions with the instructor on how to implement them.

Through the multiple approaches it is found that the traditional Lesk approach is decent performing but lags at various levels due to counting just word overlap. However, the Tf-Idf approach assigns weight to a word based on its frequency of occurrence in the context and also taking into account the frequency of the word in all the meanings (word to word co-occurrence matrix). This approach is better than the baseline Lesk because it lowers the weight of the words that occur a lot in all documents and increases the weight of the words that could be more important in the document. However, it creates long and sparse vectors. To overcome this limitation in the third approach, it is seen through the results that the short and dense vectors of word representation perform better than sparse ones wherein it predicts the context given a word rather than count. This is because the dense word vectors or 'word embeddings' encode semantic properties of words and this provides useful learning, that is, to take advantage of the potential of the distributional models proving that the advanced approach outperforms all approaches. Further, it is learned by implementing various existing approaches to create multiple word embeddings (Word2Vec, FastText, and GloVe) to not rely on a single word-embedding model and that different word embeddings yield different result (see Figure 5).

Conclusion and Future Work

This paper explores and implements several approaches towards finding appropriate sense with correct meanings (in a ranked order) for ambiguous words in a given context (paragraph) through both traditional and advanced word sense disambiguation techniques. The comparison results of techniques illustrate that semantics similarity approach that makes use of pre-trained word embeddings, if designed properly, can provide significant performance improvement and demonstrates best results. Moreover, the Tf-Idf approach comes a close second. These two approaches are suitable for the app considering ease of deployability. However, they could be refined further to achieve better performance such as applying more pre-processing steps to the input data. Further, diving into the data could be helpful, such as, study the discrepancies in the data. For example, there are same/similar meanings in the dictionary for a word that are counted as two meanings whereas only one of them is labeled as correct. Also, filtering out only ambiguous words for the analysis to diagnose and justify approach formulation is another step that would be useful. Additional approaches like BERT and LSTM could also be explored considering the time and computational complexity of the models with respect to integration of the same within the application.

References

Basile, P., Caputo, A., & Semeraro, G. (n.d.). An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model. Www.Aclweb.Org. https://www.aclweb.org/anthology/C14-1151.pdf

Britton, B. K. (1978b, January 1). Lexical ambiguity of words used in english text. Behavior Research Methods. https://link.springer.com/article/10.3758/BF03205079?error=cookies_not_supported&code=bc85b40a-74fc-4c9f-978c-585201ac3847

Ojha, N. (2019, February). Approach to Correctly Distinguish the Meaning of Word in a Context. Https://Www.Ijeat.Org/. https://www.ijeat.org/wp-content/uploads/papers/v8i3S/C11820283S19.pdf

Orkphol, K., & Yang, W. (2019). Word Sense Disambiguation Using Cosine Similarity Collaborates with Word2vec and WordNet. Future Internet, 11(5), 114. https://doi.org/10.3390/fi11050114

Sasaki, M., Komiya, K., & Shinnou, H. (2014). Word Sense Disambiguation Based on Semi-automatically Constructed Collocation Dictionary. Www.Thinkmind.Org. https://www.thinkmind.org

Seretan, V., & Wehrli, E. (n.d.). Dictionaries. An International Encyclopedia of Lexicography [E-book]. https://www.researchgate.net/publication/262152158_Context-sensitive_look-up_in_electronic_dictionaries

Sharma, P., & Joshi, N. (2019, January). Design and Development of a Knowledge-Based Approach for Word Sense Disambiguation by using WordNet for Hindi. Https://Www.Ijitee.Org/. https://www.ijitee.org/wp-content/uploads/papers/v8i3/C2580018319.pdf

Staff, S. X. (2013, May 31). Our ambiguous world of words. Phys.Org. https://phys.org/news/2013-05-ambiguous-world-words.html