Exploring ESA to Improve Word Relatedness

Nitish Aggarwal **Kartik Asooja** **Paul Buitelaar**
Insight Centre for Data Analytics
National University of Ireland
Galway, Ireland
firstname.lastname@deri.org

Abstract

Explicit Semantic Analysis (ESA) is an approach to calculate the semantic relatedness between two words or natural language texts with the help of concepts grounded in human cognition. ESA usage has received much attention in the field of natural language processing, information retrieval and text analysis, however, performance of the approach depends on several parameters that are included in the model, and also on the text data type used for evaluation. In this paper, we investigate the behavior of using different number of Wikipedia articles in building ESA model, for calculating the semantic relatedness for different types of text pairs: word-word, phrase-phrase and document-document. With our findings, we further propose an approach to improve the ESA semantic relatedness scores for words by enriching the words with their explicit context such as synonyms, glosses and Wikipedia definitions.

1 Introduction

Explicit Semantic Analysis (ESA) is a distributional semantic model (Harris, 1954) that computes the relatedness scores between natural language texts by using high dimensional vectors. ESA builds the high dimensional vectors by using the explicit concepts defined in human cognition. Gabrilovich and Markovitch (2007) introduced the ESA model in which Wikipedia and Open Directory Project[1] was used to obtain the explicit concepts. ESA considers every Wikipedia article as a unique explicit topic. It also assumes that the articles are topically orthogonal. However, recent work (Gottron et al., 2011) has shown that by using the documents from Reuters corpus instead of Wikipedia articles can also achieve comparable results. ESA model includes various parameters (Sorg and Cimiano, 2010) that play important roles on its performance. Therefore, the model requires further investigation in order to better tune the parameters.

ESA model has been adapted very quickly in different fields related to text analysis, due to the simplicity of its implementation and the availability of Wikipedia corpus. Gabrilovich and Markovitch (2007) evaluated the ESA against word relatedness dataset WN353 (Finkelstein et al., 2001) and document relatedness dataset Lee50 (Lee et al., 2005) by using all the articles from Wikipedia snapshot of 11 Nov, 2005. However, the results reported using different implementations (Polajnar et al., 2013) (Hassan and Mihalcea, 2011) of ESA on same datasets (WN353 and Lee50) vary a lot, due the specificity of ESA implementation. For instance, Hassan and Mihalcea (2011) found a significant difference between the scores obtained from their own implementation and the scores reported in the original article (Gabrilovich and Markovitch, 2007).

In this paper, first, we investigate the behavior of ESA model in calculating the semantic relatedness for different types of text pairs: word-word, phrase-phrase and document-document by using different number of Wikipedia articles for building the model. Second, we propose an approach

[1] http://www.dmoz.org

*Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 51–56,
Dublin, Ireland, August 23-24 2014.

for context enrichment of words to improve the performance of ESA on word relatedness task.

2 Background

The ESA model can be described as a method of obtaining the relatedness score between two texts by quantifying the distance between two high dimensional vectors. Every explicit concept represents a dimension of the ESA vector, and the associativity weight of a given word with the explicit concept can be taken as magnitude of the corresponding dimension. For instance, there is a word *t*, ESA builds a vector *v*, where $v = \sum_{i=0}^{N} a_i * c_i$ and c_i is i^{th} concept from the explicit concept space, and a_i is the associativity weight of word t with the concept c_i. Here, *N* represents the total number of concepts. In our implementation, we build ESA model by using Wikipedia articles as explicit concepts, and take the TFIDF weights as associativity strength. Similarly, ESA builds the vector for natural language text by considering it as a bag of words. Let $T = \{t_1, t_2, t_3...t_n\}$, where *T* is a natural language text that has n words. ESA generates the vector *V*, where $V = \sum_{t_k \epsilon T} v_k$ and $v = \sum_{i=0}^{N} a_i * c_i$. v_k represents the ESA vector of a individual words as explained above. The relatedness score between two natural language texts is calculated by computing cosine product of their corresponding ESA vectors.

In recent years, some extensions (Polajnar et al., 2013) (Hassan and Mihalcea, 2011) (Scholl et al., 2010) have been proposed to improve the ESA performance, however, they have not discussed the consistency in the performance of ESA. Polajnar et al. (2013) used only 10,000 Wikipedia articles as the concept space, and got significantly different results on the previously evaluated datasets. Hassan and Mihalcea (2011) have not discussed the ESA implementation in detail but obtained significantly different scores. Although, these proposed extensions got different baseline ESA scores but they improve the relatedness scores with their additions. Polajnar et al. (2013) used the concept-concept correlation to improve the ESA model. Hassan and Mihalcea (2011) proposed a model similar to ESA, which builds the high dimensional vector of salient concepts rather than explicit concepts. Gortton et al. (2011) investigated the ESA performance for document relatedness and showed that ESA scores are not tightly dependent on the explicit concept spaces.

Minimum unique words (K)	Total number of articles (N)
100	438379
300	110900
500	46035
700	23608
900	13718
1100	8322
1300	5241
1500	3329
1700	2126
1900	1368

Table 1: The total number of retrieved articles for different values of K

3 Investigation of ESA model

Although Gortton et al. (2011) has shown that ESA performance on document pairs does not get affected by using different number of Wikipedia articles, we further examine it for word-word and phrase-phrase pairs. We use three different datasets WN353, SemEvalOnWN (Agirre et al., 2012) and Lee50. WN353 contains 353 word pairs, SemEvalOnWN consists of 750 short phrase/sentence pairs, and Lee50 is a collection of 50 document pairs. All these datasets contain relatedness scores given by human annotators. We evaluate ESA model on these three datasets against different number of Wikipedia articles. In order to select different number of Wikipedia articles, we sort them according to the total number of unique words appearing in each article. We select N articles, where N is total number of articles which have at least K unique words. Table 1 shows the total number of retrieved articles for different values of K. We build 20 different ESA models with the different values of N retrieved by varying K from 100 to 2000 with an interval of 100. Figure 1 illustrates Spearman's rank correlation of all the three types of text pairs on Y-axis while X-axis shows the different values of N which are taken to build the model. It shows that ESA model generates very consistent results for phrase pairs similar to the one reported in (Aggarwal et al., 2012), how-

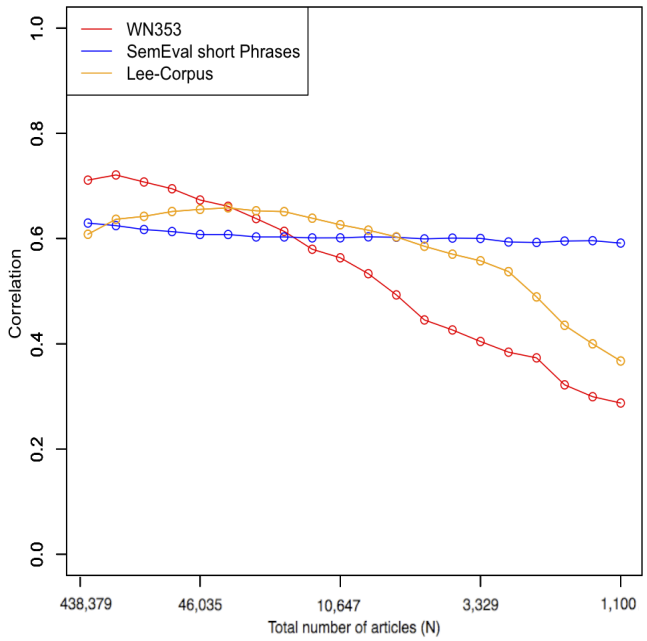

Figure 1: ESA performance on different types of text pairs by varying the total number of articles

ever, the correlation scores decreases monotonously in the case of word pairs as the number of articles goes down. In the case of document pairs, ESA produces similar results until the value of N is chosen according to K = 1000, but after that, it decreases quickly because the number of articles becomes too low for making a good enough ESA model. All this indicates that word-word relatedness scores have a strong impact of changing the N in comparison of document-document or phrase-phrase text pairs. An explanation to this is that the size of the ESA vector for a word solely depends upon the popularity of the given word, however, in the case of text, the vector size depends on the popularity summation of all the words appearing in the given text. It suggests that the word relatedness problem can be reduced to short text relatedness by adding some related context with the given word. Therefore, to improve the ESA performance for word relatedness, we propose an approach for context enrichment of words. We perform context enrichment by concatenating related context with the given word and use this context to build the ESA vector, which transforms the word relatedness problem to phrase relatedness.

4 Context Enrichment

Context enrichment is performed by concatenating the context defining text to the given word before building the ESA vector. Therefore, instead of building the ESA vector of a word, the vector is built for the short text that is obtained after concatenating the related context. This is similar to classical query expansion task (Aggarwal and Buitelaar, 2012; Pantel and Fuxman, 2011), where related concepts are concatenated with a query to improve the information retrieval performance. We propose three different methods to obtain related context: 1) WordNet-based Context Enrichment 2) Wikipedia-based Context Enrichment, and 3) WikiDefinition-based Context Enrichment.

4.1 WordNet-based Context Enrichment

WordNet-based context enrichment uses the WordNet synonyms to obtain the context, and concatenates them into the given word to build the ESA vector. However, WordNet may contain more than one synset for a word, where each synset represents a different semantic sense. Therefore, we obtain more than one contexts for a given word, by concatenating the different synsets. Further, we calculate ESA score of every context of a given word against all the contexts of the other word which is being compared, and consider the highest score as the final relatedness score. For instance, there is a given word pair "train and car", car has 8 different synsets that build 8 different contexts, and train has 6 different synsets that build 6 different contexts. We calculate the ESA score of these 8 contexts of car to the 6 contexts of train, and finally select the highest obtained score from all of the 24 calculated scores.

4.2 Wikipedia-based Context Enrichment

In this method, the context is defined by the word usage in Wikipedia articles. We retrieve top 5 Wikipedia articles by querying the articles' content, and concatenate the short abstracts of the retrieved articles to the given word to build the ESA vector. Short abstract is the first two sentences of Wikipedia article and has a maximum limit of 500 characters. In order to retrieve the top 5 articles from Wikipedia for a given word, we build an index of all Wikipedia articles and use TF-IDF scores. We further explain

our implementation in Section 5.1.

4.3 WikiDefinition-based Context Enrichment

This method uses the definition of a given word from Wikipedia. To obtain a definition from Wikipedia, we first try to find a Wikipedia article on the given word by matching the Wikipedia title. As definition, we take the short abstract of the Wikipedia article. For instance, for a given word "train", we take the Wikipedia article with the title "Train"[2]. If there is no such Wikipedia article, then we use the previous method "Wikipedia-based Context Enrichment" to get the context defining text for the given word. In contrary to the previous method for defining context, here we first try to get a more precise context as it comes from the Wikipedia article on that word only. After obtaining the definition, we concatenate it to the given word to build the ESA vector. At the time of experimentation, we were able to find 339 words appearing as Wikipedia articles out of 437 unique words in the WN353 dataset.

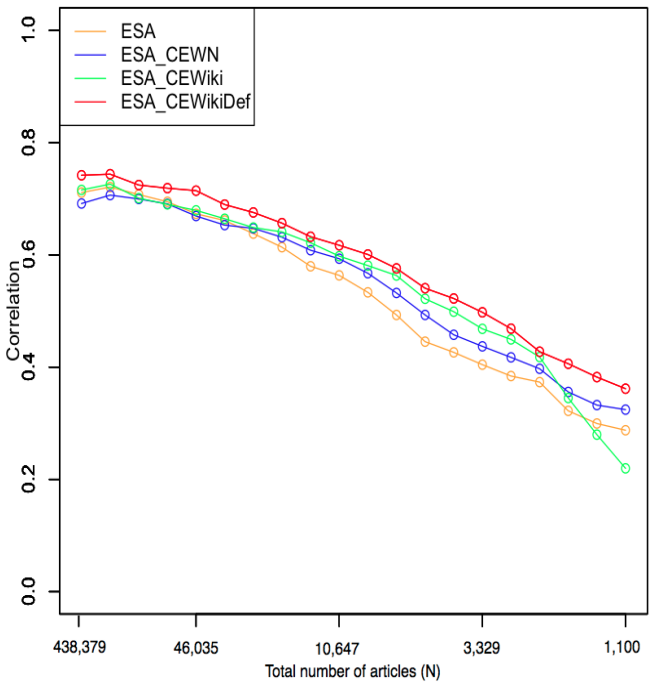

Figure 2: Effect of different types of context enrichments on WN353 gold standard

[2]http://en.wikipedia.org/wiki/Train

5 Experiment

5.1 ESA implementation

In this section, we describe the implementation of ESA and the parameters used to build the model. We build an index over all Wikipedia articles from the pre-processed Wikipedia dump from November 11, 2005 (Gabrilovich, 2006). We use Lucene[3] to build the index and retrieve the articles using TF-IDF scores. As described in section 3, we build 20 different indices with different values of total number of articles (N).

5.2 Results and Discussion

To evaluate the effect of the aforementioned approaches for context enrichment, we compare the results obtained by them against the results generated by ESA model as a baseline. We calculated the scores on WN353 word pairs dataset by using ESA, WordNet-based Context Enrichment (ESA_CEWN), Wikipedia-based Context Enrichment (ESA_CEWiki) and WikiDefition-based Context Enrichment (ESA_CEWikiDef). Further, we examine the performance of context enrichment approaches by reducing the total number of articles taken to build the model. Figure 2 shows that the proposed methods of context enrichment significantly improve over the ESA scores for different values of N.

Table 2 reports the results obtained by using different context enrichments and ESA model. It shows Spearman's rank correlation on four different values of N. All the proposed context enrichment methods improve over the ESA baseline scores. Context enrichments based on Wikipedia outperforms the other methods, and ESA_CEWikiDef significantly improves over the ESA baseline. Moreover, given a very less number of Wikipedia articles used for building the model, ESA_CEWikiDef obtains a correlation score which is considerably higher than the one obtained by ESA baseline. ESA_CEWN and ESA_CEWiki can include some unrelated context as they do not care about the semantic sense of the given word, for instance, for a given word "car", ESA_CEWiki

[3]https://lucene.apache.org/

K	Total articles (N)	ESA	ESA_CEWN	ESA_CEWiki	ESA_CEWikiDef
100	438,379	0.711	0.692	0.724	**0.741**
200	221,572	0.721	0.707	0.726	**0.743**
500	46,035	0.673	0.670	0.679	**0.698**
1000	10,647	0.563	0.593	0.598	**0.614**

Table 2: Spearman rank correlation scores on WN353 gold standard

includes the context about the word "car" at surface level rather than at the semantic level. However, ESA_CEWikiDef only includes the definition if it does not refer to more than one semantic sense, therefore, ESA_CEWikiDef outperforms all other types of context enrichment.

We achieved best results in all the cases by taking all the articles which has a minimum of 200 unique words (K=200). This indicates that further increasing the value of K considerably decreases the value of N, consequently, it harms the overall distributional knowledge of the language, which is the core of ESA model. However, decreasing the value of K introduces very small Wikipedia articles or stubs, which do not provide enough content on a subject.

6 Conclusion

In this paper, we investigated the ESA performance for three different types of text pairs: word-word, phrase-phrase and document-document. We showed that ESA scores varies significantly for word relatedness measure with the change in the number (N) and length (≈K which is the number of unique words) of the Wikipedia articles used for building the model. Further, we proposed context enrichment approaches for improving word relatedness computation by ESA. To this end, we presented three different approaches: 1) WordNet-based, 2) Wikipedia-based, and 3) WikiDefinition-based, and we realized that concatenating the context defining text improves the ESA performance for word relatedness task.

Acknowledgments

This work has been funded in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (INSIGHT) and by the EU FP7 program in the context of the project LIDER (610782).

References

Nitish Aggarwal and Paul Buitelaar. 2012. Query expansion using wikipedia and dbpedia. In *CLEF*.

Nitish Aggarwal, Kartik Asooja, and Paul Buitelaar. 2012. DERI&UPM: Pushing corpus based relatedness to similarity: Shared task system description. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 643–647, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Evgeniy Gabrilovich. 2006. *Feature generation for textual information retrieval using world knowledge*. Ph.D. thesis, Technion - Israel Institute of Technology, Haifa, Israel, December.

Thomas Gottron, Maik Anderka, and Benno Stein. 2011. Insights into explicit semantic analysis. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1961–1964. ACM.

Zellig Harris. 1954. Distributional structure. In *Word 10 (23)*, pages 146–162.

Samer Hassan and Rada Mihalcea. 2011. Semantic relatedness using salient semantic analysis. In *AAAI*.

Michael David Lee, BM Pincombe, and Matthew Brian Welsh. 2005. An empirical evaluation of models of text document similarity. *Cognitive Science*.

Patrick Pantel and Ariel Fuxman. 2011. Jigs and lures: Associating web queries with structured entities. In *ACL*, pages 83–92.

Tamara Polajnar, Nitish Aggarwal, Kartik Asooja, and Paul Buitelaar. 2013. Improving esa with document similarity. In *Advances in Information Retrieval*, pages 582–593. Springer.

Philipp Scholl, Doreen Böhnstedt, Renato Domínguez García, Christoph Rensing, and Ralf Steinmetz. 2010. Extended explicit semantic analysis for calculating semantic relatedness of web resources. In *Sustaining TEL: From Innovation to Learning and Practice*, pages 324–339. Springer.

Philipp Sorg and Philipp Cimiano. 2010. An experimental comparison of explicit semantic analysis implementations for cross-language retrieval. In *Natural Language Processing and Information Systems*, pages 36–48. Springer.