Non-Orthogonal Explicit Semantic Analysis

Nitish Aggarwal Kartik Asooja Georgeta Bordea Paul Buitelaar
Insight Centre for Data Analytics
National University of Ireland
Galway, Ireland
firstname.lastname@insight-centre.org

Abstract

Explicit Semantic Analysis (ESA) utilizes the Wikipedia knowledge base to represent the semantics of a word by a vector where every dimension refers to an explicitly defined concept like a Wikipedia article. ESA inherently assumes that Wikipedia concepts are orthogonal to each other, therefore, it considers that two words are related only if they co-occur in the same articles. However, two words can be related to each other even if they appear separately in related articles rather than co-occurring in the same articles. This leads to a need for extending the ESA model to consider the relatedness between the explicit concepts (i.e. Wikipedia articles in Wikipedia based implementation) for computing textual relatedness. In this paper, we present Non-Orthogonal ESA (NESA) which represents more fine grained semantics of a word as a vector of explicit concept dimensions, where every such concept dimension further constitutes a semantic vector built in another vector space. Thus, NESA considers the concept correlations in computing the relatedness between two words. We explore different approaches to compute the concept correlation weights, and compare these approaches with other existing methods. Furthermore, we evaluate our model NESA on several word relatedness benchmarks showing that it outperforms the state of the art methods.

1 Introduction

Significance of quantifying relatedness between two natural language texts has been shown in various tasks which deal with information retrieval (IR), natural language processing (NLP), or other related fields. The semantics of a word can be obtained from existing lexical resources like WordNet and FrameNet. However, such lexical resources require domain expertise for defining the hierarchical structure, which makes their creation very expensive. Therefore, distributional semantic models (DSMs) have achieved much attention as they utilize available document collections like Wikipedia, and do not depend upon human expertise (Harris, 1954). DSMs represent the semantics of a word by transforming it to a high dimensional distributional vector in a predefined concept space. Many models have been proposed that derive this concept space by using explicit concepts or implicit concepts. Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007) utilizes the concepts which are explicitly derived under human cognition like Wikipedia concepts (articles). However, Latent Semantic Analysis (LSA) derives a latent concept space by performing dimensionality reduction (Landauer et al., 1998).

Gabrilovich and Markovitch (2007) introduced ESA model in which Wikipedia and Open Directory Project were used to obtain the explicit concepts, however, Wikipedia has been a popular choice in further ESA implementations (Polajnar et al., 2013; Gottron et al., 2011; Aggarwal et al., 2014). ESA represents the semantics of a word with a high dimensional vector over the Wikipedia concepts. The tf-idf weight of the word with the textual content under a Wikipedia concept reflects the magnitude

Table 1: Top 5 Wikipedia concepts for "football" and "soccer" in the ESA vector

#	**football**	**soccer**
1	FIFA	History of soccer in the United States
2	Football	Soccer in the United States
3	History of association football	United States Soccer Federation
4	Football in England	North American Soccer League (196884)
5	Association football	United Soccer Leagues

of the corresponding vector dimension. To obtain the semantic relatedness between two words, it computes the vector dot product between their vectors. ESA considers the dimensions as orthogonal to each other. For instance, the synonyms like "soccer" and "football" are highly related, however, they may not co-occur together in many Wikipedia articles. Table 1 shows that the top 5 Wikipedia concepts retrieved for "football" and "soccer" do not share any concept, however, the concepts may exhibit relatedness to each other. Consequently, ESA model assumes that words can be related only if they co-occur in the same articles. However, two words can also be related even if they do not share the same articles at all, but appear in the related ones. LSA resolves the orthogonality issue to some extent by building latent concept space in an unsupervised way (Landauer et al., 1998). However, the resulting latent concepts are not as clearly interpretable as the human-labeled concepts in the ESA model. Previous studies (Gabrilovich and Markovitch, 2007; Cimiano et al., 2009; Hassan and Mihalcea, 2011) show that ESA performs better than LSA for computing text relatedness. Therefore, it is important to consider the relatedness between dimensions in the ESA model, rather than considering them orthogonal, and also without losing the explicit property of ESA model at the same time.

In this paper, we present Non-Orthogonal ESA (NESA) model, an extension to ESA, which also uses relatedness between the explicit concepts for computing semantic relatedness between texts. The concepts in ESA model are clearly interpretable and they refer to the title of Wikipedia articles. This characteristic provides an opportunity to investigate different concept relatedness measures, such as relatedness between articles' content (document relatedness) or relatedness between corresponding Wikipedia titles. In order to investigate the performance of these concept relatedness measures, we evaluate them on an entity relatedness benchmark called KORE (Hoffart et al., 2012) as Wikipedia article title generally refers to an entity.

We then apply the different approaches for computing concept relatedness in our model NESA to compute text relatedness. We evaluate NESA on several word relatedness benchmarks to verify whether considering non-orthogonality in ESA model improves its performance.

2 Related Work

2.1 Text Relatedness

In recent years, there have been a variety of efforts to develop semantic relatedness measures. Classical approaches assess the relatedness scores by using existing knowledge bases or corpus statistics. Lexical resources such as WordNet or Roget thesaurus (Jarmasz and Szpakowicz, 2004) are used as knowledge bases to compute the relatedness scores between two words. Most of these approaches make use of the hierarchical structure present in the lexical resources. For instance, Hirst and St-Onge (1998), Leacock and Chodorow (1998), and Wu and Palmer (1994) utilize the edges that define taxonomic relations between words; Banerjee and Pedersen (2002) computes the scores by obtaining the overlap between glosses associated with the words; and some of the other approaches (Resnik, 1995; Lin, 1998) use corpus evidence with the taxonomic structure of WordNet. These approaches are limited to perform only for the lexical entries and thus do not work with non-dictionary words. Moreover, these measures rely on the manually constructed lexical resources and they are not portable to multiple languages due to unavailability of lexical resources in multiple languages.

Corpus-based methods such as LSA (Landauer et al., 1998), Latent Dirichlet Allocation (LDA) (Blei et al., 2003), and ESA (Gabrilovich and Markovitch, 2007) employ statistical models to build the semantic profile of a word. LSA and LDA generate unsupervised topics from a textual corpus, and represent the semantics of a word by its distribution over these topics. LSA performs singular value decomposition (SVD) to obtain a latent concept space. On the contrary, ESA directly uses supervised topics such as Wikipedia concepts that are built manually, and considers that every concept is orthogonal to each other. Polajnar at el. (2013) proposed an approach to improve ESA by considering the concept relatedness using word overlap in Wikipedia articles' content. Radinsky at el. (2011) introduced Temporal Semantic Analysis (TSA) also considers the concept relatedness in ESA model, which is computed by using their temporal distribution over the NewYork Times news archives from the last 100 years. Although, these approaches consider relatedness between explicit concepts (Polajnar et al., 2013; Radinsky et al., 2011) and improve the accuracy, however, either they define a weak concept relatedness measure or require an external corpus statistics. Our approach takes inspiration from them and uses more advanced concept relatedness measures that rely on the same corpus statistics, which is used to build the ESA model.

2.2 Concept Relatedness

As NESA model requires a concept relatedness measure to overcome orthogonality, we address here the existing methods of computing it (Strube and Ponzetto, 2006; Witten and Milne, 2008; Polajnar et al., 2013). Most of these approaches rely on Wikipedia and its derived knowledge bases such as DBpedia[1], YAGO[2] and FreeBase[3]. These knowledge bases provide immense amount of information about millions of concepts or entities which can be utilized for computing concept relatedness. Strube and Ponzetto (2006) proposed WikiRelate that counts the edges between two concepts in Wikipedia link structure, and also considers the depth of a concept in the Wikipedia category structure. Witten and Milne (2008) applied Google distance metric (Cilibrasi and Vitanyi, 2007) on incoming links in Wikipedia. Hoffart at el. (2012) utilized the textual content associated with the Wikipedia concepts. It observes the partial overlap between the concepts (key-phrases) appearing in the article content. The above mentioned approaches mainly exploit the article content or Wikipedia link structure for computing concept relatedness. In this paper, we also utilize the distributional information of the title and hyperlinks for computing concept relatedness.

3 Non-Orthogonal Explicit Semantic Analysis

To compute text relatedness, NESA uses relatedness between the dimensions of the distributional vectors to overcome the orthogonality in ESA model. In addition to represent the words as distributional vectors, where each dimension is associated with a Wikipedia concept as in ESA model, NESA also utilizes a square matrix C$_{n,n}$ (n is the total number of dimensions) containing the correlation weights between the dimensions. Thus, to obtain the relatedness score between the words *w1* and *w2*, NESA formulates the measure as follows:

$$rel_{NESA}(w1, w2) = \mathbf{w1}^T_{1,n}.C_{n,n}.\mathbf{w2}_{n,1} \qquad (1)$$

where $\mathbf{w1}_{n,1}$ and $\mathbf{w2}_{n,1}$ are the corresponding distributional vectors consisting of n dimensions. Every concept dimension can be further semantically interpreted as a distributional vector in some other vector space of m dimensions. This transformation allows the computation of the correlation weights between the concept dimensions. Thus, a transformation matrix E$_{m,n}$ can be built, where each column corresponds to a transformation vector for each concept dimension. Using the matrix E$_{m,n}$, we can compute the matrix C$_{n,n}$ by multiplying E$_{m,n}$ with its transpose as in equation 2. In the next section, we discuss the different approaches used for computing C$_{n,n}$ containing the relatedness between the concept dimensions .

$$C_{n,n} = E^T_{n,m}.E_{m,n} \qquad (2)$$

[1] http://dbpedia.org/About

[2] http://yago-knowledge.org/

[3] https://www.freebase.com/

4 Computing Concept Relatedness

NESA requires the relatedness scores between Wikipedia concepts (articles), therefore we present the different approaches for computing $C_{n,n}$ matrix using $E_{m,n}$. Every Wikipedia article consists of different fields to represent the semantics of the concept dimensions, such as Wikipedia title, textual description and hyperlinks. We utilize this information to implement four different concept relatedness measures: VSM-Text, VSM-Hyperlinks, ESA-WikiTitle, and DiSER. These approaches represent the semantics of a concept with a distributional vector of m dimensions. All such vectors combined as column vectors for n concept dimensions form the matrix $E_{m,n}$.

4.1 VSM-Text

This approach is based on plain Vector Space Model (VSM) for text. It calculates the relatedness scores between concepts by taking word overlap between their corresponding Wikipedia article content. The concept is transformed to a column vector mx1, where m is the total number of unique words in the Wikipedia corpus. The magnitude of each dimension is calculated on the basis of the number of occurrences of the different words in the associated Wikipedia article content.

4.2 VSM-Hyperlink

Similar to the VSM-Text, this approach calculates the concept relatedness by taking the overlap between the hyperlinks present in their corresponding Wikipedia articles' content. The concept is transformed to a column vector mx1, where m is the total number of hyperlinks in the whole Wikipedia. The magnitude of each dimension is calculated on the basis of the number of occurrences of the different hyperlinks in the associated Wikipedia article content.

4.3 ESA-WikiTitle

One intuitive way of obtaining concept relatedness scores is by using ESA itself for calculating the relatedness between the concepts. We use the associated Wikipedia article title for this purpose. ESA represents the semantics of a word with a high dimensional vector over the Wikipedia concepts. Therefore, each concept dimension is transformed into a column vector of mx1, where m is the total number of Wikipedia concepts. The magnitude of each dimension is computed by using the term frequency (tf) and inverse document frequency (idf) for the terms appearing in the Wikipedia article title over the Wikipedia corpus (Gabrilovich and Markovitch, 2007).

4.4 DiSER

Distributional Semantics for Entity Relatedness (DiSER) (Aggarwal and Buitelaar, 2014) is a model for computing relatedness scores between entities. DiSER considers every Wikipedia concept as an entity. Therefore, it can be used for computing concept relatedness matrix $C_{n,n}$, as required by the NESA model. In contrast to text relatedness measures based on DSMs such as ESA, which do not distinguish between entity and text, DiSER differentiate between entity and its surface forms by using unique hyperlinks referring to entities in Wikipedia for encoding entities while building DSMs. It uses the distributional information of such hyperlinks only over the whole Wikipedia corpus for representing a concept by a high dimensional distributional vector. Therefore, each concept dimension is transformed into a column vector of mx1, where m is the total number of Wikipedia concepts. The magnitude of each dimension is computed by using the concept frequency (ef) and inverse document frequency (idf) for an concept in the Wikipedia corpus. The concept frequency (cf) is a slight variation of term frequency. It computes the frequency of a concept appearing as hyperlink in the Wikipedia articles. To obtain the DiSER based relatedness scores between Wikipedia concepts, we use Entity Relatedness Graph (EnRG)[4] (Aggarwal et al., 2015), which is a focused related entities explorer based on DiSER scores.

5 Evaluation of Concept Relatedness Measures

In this section, we evaluate the different approaches defined for computing concept relatedness measures in the previous section. For our evaluation, we use the snapshot of English Wikipedia from 1^{st} October, 2013. This snapshot consists of 13,872,614 articles,

[4] EnRG demo: http://enrg.insight-centre.org

in which 5,934,022 are Wikipedia redirects. We filtered out all the namespace[5] pages by using the articles' titles as they have specific namespace patterns. There are 3,571,206 namespace pages in this snapshot. We remove all those articles which contain less than 100 unique words or less than 5 hyperlinks; such articles are too specific and may generate some noise. We perform further filtering by removing all the articles if their titles are numbers like "19", dates like "June 1", or if the title starts with "list". We finally obtain a total of 3,635,833 Wikipedia articles for our experiment. We implement all the concept relatedness measures by using these obtained Wikipedia articles.

VSM-Text represents the semantics of a concept with a column vector of mx1, where m is the total number of unique words appear in Wikipedia. Wikipedia contains more than 2.5 billion unique words, therefore, to reduce the matrix size, we use only 5 million most frequent words. ESA-WikiTitle represents the semantics of a concept with a column vector of mx1, where m is 3,635,833 in our implementation. In order to obtain the hyperlinks for VSM-Hyperlink and DiSER, we retain only those text segments which have manually defined links provided by Wikipedia volunteers. However, the volunteers may not create the link for every surface form appearing in the article content. For instance, "Apple" occurs 213 times in "Steve Jobs" Wikipedia page in our corpus, but only 7 out of these 213 are linked to the "Apple Inc." Wikipedia page. The term frequency of "Apple" is calculated without considering the partial string matches, for example, we do not count if "apple" appears as a substring of any annotated text segment like "Apple Store" or "Apple Lisa". To obtain the actual frequency of every hyperlink for computing the magnitude of the dimension, we apply "one sense per discourse" heuristic (Gale et al., 1992), which assumes that a term tends to have the same meaning in the same discourse. We link every additional un-linked occurrence of the text segment with the same hyperlink appearing most of the times for the same segment in the article. The total number of hyperlinks possible in our corpus would be equal to the total number of Wikipedia articles i.e. 3,635,833.

5.1 Dataset

In order to evaluate the concept relatedness measures, we performed our experiments on the gold standard benchmark dataset KORE (Hoffart et al., 2012). The KORE dataset consists of 21 seed Wikipedia concepts selected from the YAGO knowledge base[6]. Every seed concept has a ranked list of 20 related Wikipedia concepts. In order to build this dataset, 20 concept candidates are selected and ranked by human evaluators on crowdsourcing platforms to give the relative comparison between two candidates against the corresponding seed Wikipedia concept. For instance, human evaluators provide their judgement if "Mark Zuckerberg" is more related to "Facebook" than "Sean Parker". With the answers for such binary questions, a ranked list is prepared for every seed Wikipedia concept. The KORE dataset[7] consists of 21 seed candidates, thus forming 420 concept pairs with their relatedness scores assigned by 15 human evaluators.

5.2 Experiment

We compare the concept relatedness measures described in section 4 against other existing methods. Hoffart at el. (2012) proposed KORE and KPCS which use the article content to compute the concept relatedness. They use Mutual Information (MI-weight) to capture the importance of the hyperlink for a Wikipedia concept. To evaluate the concept relatedness measures using KORE dataset, we compute the concept relatedness scores for all the concept pairs and rank the list of 20 candidates for each seed Wikipedia concept. We calculated Spearman Rank correlation between the gold standard dataset and the results obtained from VSM-Text, VSM-Hyperlink, ESA-WikiTitle and DiSER.

5.3 Results and Discussion

Experimental results are shown in Table 2. We compare our results with the other existing methods of computing concept relatedness: WLM, KORE, and KPCS. WLM is the Wikipedia Link-based approach by Witten and Milne (2008). KPCS and

[5] http://en.wikipedia.org/wiki/Wikipedia:Namespace

[6] http://datahub.io/dataset/yago

[7] http://www.mpi-inf.mpg.de/yago-naga/aida

Table 2: Spearman rank correlation of concept relatedness measures with gold standard

Concept Relatedness Measures	Spearman Rank Correlation with human
VSM-Text	0.510
VSM-Hyperlink	0.637
ESA	0.661
DiSER	**0.781**
WLM	0.610
KPCS	0.698
KORE	0.673

KORE are the approaches proposed in (Hoffart et al., 2012), where KPCS is the cosine similarity on MI-weighted keyphrases while KORE represents the keyphrase overlap relatedness. These keyphrases can be the text segment with hyperlinks in the article content. Therefore, KPCS is a similar approach to VSM-Hyperlink, besides KPCS assigns MI-weights to capture the generality and specificity of concept in the Wikipedia article. Many concepts in the gold standard dataset are defined by ambiguous surface forms such as "NeXT" and "Nice", or they have ambiguous text segments in their surface forms like "Jobs" in "Steve Jobs" and "Guitar" in the "Guitar Hero" video game. Therefore, the effect of using only hyperlinks can be observed with the remarkable difference between the results obtained by ESA and DiSER. DiSER improves the accuracy over ESA by 20%. These scores illustrate that ESA fails in generating the appropriate semantic profiles for ambiguous terms. VSM-Text does not capture the semantics of Wikipedia concepts as the textual description in Wikipedia article also contains generic terms which are not enough to specify the precisely semantics of Wikipedia concepts. Therefore, VSM-Hyperlink achieved noticeable improvement over VSM-Text as VSM-Hyperlink builds the semantic profile by using hyperlinks in the article content. These hyperlinks are created by Wikipedia volunteers, therefore, it can be assumed that the text segments which are linked to other Wikipedia article, are more important than un-linked ones. However, KPCS and KORE achieved significantly higher accuracy in comparison to VSM-Hyperlink, which indicates that generality and specificity of hyperlinks in the article content are very influential features for concept relatedness measures.

6 Evaluation of NESA for Word Relatedness

In this section, we evaluate NESA for word relatedness. We experiment by using different concept relatedness measures as explained in section 4 for building the $C_{n,n}$ in NESA model as shown in equations 1 and 2. We use the same filtered Wikipedia articles as used for evaluating the concept relatedness measures in the previous section.

6.1 Dataset

We use 6 different word relatedness benchmarks to evaluate NESA.

WN353 consists of 353 word pairs annotated by 13-15 human experts on a scale of 0-10. 0 refers to un-related and 10 stands for highly related or identical. This dataset mainly contains generic words like "money", "drink", "movie", etc.. It also contains named entities such as "Jerusalem", "Palestinian" and "Israel", which makes this dataset more challenging for approaches that use only the lexical resources.

WN353Rel and WN353Sim datasets are the subsets of WN353. As WN353 contains similar and related word pairs, Agirre at el. (2009) refine the WN353 gold standard by splitting it in two parts: related word pairs and similar word pairs. The notion of similarity and relatedness are defined as follow: two words are similar if they are connected through the taxonomic relations like synonym or hyponym in lexical resources, while two words can be considered related if they are connected through other relations such as meronym and holonym. For instance, "football" and "soccer" are two similar words while "computer" and "software" can be considered as related. Finally, WN353Rel and WN353Sim contain 252 and 203 word pairs respectively.

MC30 is the dataset build by Miller and Charles (1991) that contains the selected word pairs of WN353. The relatedness scores of these words are

Table 3: Spearman rank correlation of relatedness measures with gold standard datasets

#	**WN353**	**WN353Rel**	**WN353Sim**	**MC30**	**RG65**	**MT287**
H&S	0.347	0.142	0.497	0.811	0.813	0.278
L&C	0.302	0.172	0.412	0.793	0.823	0.284
Lesk	0.337	0.125	0.511	0.583	0.5466	0.271
W&P	0.316	0.131	0.461	0.784	0.807	0.331
Resnik	0.353	0.184	0.535	0.693	0.731	0.234
J&C	0.317	0.089	0.442	0.820	0.804	0.296
Lin	0.348	0.154	0.483	0.750	0.788	0.286
Roget	0.415	- -	- -	**0.856**	0.804	- -
SSA	0.629	- -	- -	0.810	0.830	- -
Polajnar et al.	0.664	- -	- -	- -	- -	- -
ESA	0.660	0.643	0.663	0.765	0.826	0.507
NESA (VSM-Text)	0.666	0.648	0.669	0.768	0.827	0.509
NESA (VSM-Hyperlink)	0.670	0.649	0.672	0.768	0.828	0.516
NESA (ESA-WikiTitle)	0.681	0.652	0.684	0.774	0.830	0.541
NESA (DiSER)	**0.696**	**0.663**	**0.719**	0.784	**0.839**	**0.572**

provided by 38 human experts on a scale of 0-4.

RG65 is a collection of 65 non-technical word pairs. These word pairs are annotated by 51 human experts (see for more detail (Rubenstein and Goodenough, 1965)).

MT287 is a relatively newer dataset that contains 287 word pairs. This dataset is prepared mainly to study the effect of temporal distribution (Radinsky et al., 2011) of a word over several years. The relatedness scores of the word pairs are obtained from 15-20 mechanical turkers.

6.2 Experiment

We compare the NESA model with other state of the art methods of calculating word relatedness: Explicit Semantic Analysis (ESA), Salient Semantic Analysis (SSA), and several WordNet-based similarity measures. Hassan and Mihalcea (2011) reported SSA performance on WN353, MC30 and RG65 datasets as shown in table 3. The WordNet-based similarity measures are implemented using WS4J (WordNet Similarity for Java)[8] library built on WordNet 3.0.

6.3 Results and Discussion

Table 3 shows the results of the NESA model with different concept relatedness approaches and other state of the art methods of calculating word relatedness. The knowledge-based methods that use lexical resources like WordNet or Roget thesaurus (Jarmasz and Szpakowicz, 2004), achieve higher accuracy if the words in benchmark datasets are available in the knowledge bases. For instance, WordNet-based measures (H&S (Hirst and St-Onge, 1998), L&C (Leacock and Chodorow, 1998), Lesk (Banerjee and Pedersen, 2002), W&P (Wu and Palmer, 1994), Resnik (Resnik, 1995) J&C (Jiang and Conrath, 1997), Lin (Lin, 1998)) and Roget thesaurus-based measure (Jarmasz and Szpakowicz, 2004) achieved higher accuracy on MC30 and RG65 datasets. However, these approaches may not fit well for the datasets that contain non-dictionary words, therefore, the accuracy of knowledge-based measures decrease significantly on other datasets. Corpus-based measures ESA and SSA achieved higher scores than knowledge-based methods on WN353, WN353Rel, WN353Sim and MT287 datasets. Moreover, corpus-based methods performed comparable to knowledge-based methods on MC30 and RG65. Most of the knowledge-based measures use the taxonomic relations for computing word relatedness. Therefore, these measures

[8] https://code.google.com/p/ws4j/

obtained poor results on WN353Rel in contrast to WN353Sim dataset. However, corpus-based measures performed well for both type of relations i.e. similarity and relatedness.

The NESA model combined with any concept relatedness measure outperforms ESA for all the word relatedness benchmark datasets. It shows that considering non-orthogonality between explicit concepts in ESA model improves the accuracy. NESA-VSM-Hyperlink performs better than NESA-VSM-Text implying that considering only the hyperlinks from the article content works better than taking the overlap of whole content. NESA-ESA-WikiTitle and NESA-DiSER achieved higher scores than both NESA-VSM-Text and NESA-VSM-Hyperlink. It shows that the distributional representation of the article title captures the semantic information better than considering only the corresponding article content. Another interesting thing to note is that the correlation scores obtained by NESA model with the four concept relatedness measures follow the same order in table 3 as of the correlation scores obtained in evaluating concept relatedness shown in table 2. It represents the consistency of proposed concept relatedness measures in two different experiment settings. NESA-DiSER achieved the highest correlation scores in all the word relatedness benchmark datasets.

7 Conclusion

We presented Non-Orthogonal ESA which introduces the relatedness between the explicit concepts in the ESA model for computing semantic relatedness, without compromising with the explicit property of the ESA concept space. We showed that the word relatedness results vary with the different concept relatedness measures. NESA outperformed all state of the art methods, in particular, NESA-DiSER achieved the highest correlation with the gold standard. We also evaluated the different concept relatedness measures using benchmark dataset KORE, in which DiSER outperformed all others.

8 Acknowledgements

This work has been funded by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (INSIGHT).

References

Nitish Aggarwal and Paul Buitelaar. 2014. Wikipedia-based distributional semantics for entity relatedness. In *2014 AAAI Fall Symposium Series*.

Nitish Aggarwal, Kartik Asooja, and Paul Buitelaar. 2014. Exploring esa to improve word relatedness. *Lexical and Computational Semantics (* SEM 2014)*, 51.

Nitish Aggarwal, Kartik Asooja, Housam Ziad, and Paul Buitelaar. 2015. Who are the american vegans related to brad pitt? exploring related entities. In *24th International World Wide Web Conference (WWW 2015), Florence, Italy*.

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Computational linguistics and intelligent text processing*, pages 136–145. Springer.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Rudi L Cilibrasi and Paul MB Vitanyi. 2007. The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):370–383.

Philipp Cimiano, Antje Schultz, Sergej Sizov, Philipp Sorg, and Steffen Staab. 2009. Explicit versus latent concept models for cross-language information retrieval. In *IJCAI*, volume 9, pages 1513–1518.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th IJCAI*, pages 1606–1611.

William A Gale, Kenneth W Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, pages 233–237. ACL.

Thomas Gottron, Maik Anderka, and Benno Stein. 2011. Insights into explicit semantic analysis. In *Proceedings of the 20th CIKM*, 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*.

Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*, 305:305–332.

Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. Kore: Keyphrase overlap relatedness for entity disambiguation. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 545–554. ACM.

Mario Jarmasz and Stan Szpakowicz. 2004. Rogets thesaurus and semantic similarity1. *Recent Advances in Natural Language Processing III: Selected Papers from RANLP*, 2003:111.

Jay J Jiang and David W Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

T K Landauer, P. W. Foltz, and D Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

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

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *20th WWW*, pages 337–346.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Michael Strube and Simone Paolo Ponzetto. 2006. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*, volume 6, pages 1419–1424.

I Witten and David Milne. 2008. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pages 25–30.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on ACL*, pages 133–138. ACL.