CPRel: Semantic Relatedness Computation Using Wikipedia based Context Profiles

Shahida Jabeen, Xiaoying Gao and Peter Andreae

School of Engineering and Computer Science
Victoria University of Wellington, New Zealand
[shahidarao, Xiaoying.Gao, Peter.Andreae]@ecs.vuw.ac.nz

Abstract. Semantic relatedness is a well known problem with its significance ranging from computational linguistics to Natural language Processing applications. Relatedness computation is restricted by the amount of common sense and background knowledge required to relate any two terms. This paper proposes a novel model of relatedness using context profile built on features extracted from encyclopedic knowledge. Proposed research makes use of Wikipedia to represent the context of a word in the high dimensional space of Wikipedia labels. Semantic relatedness of a word pair is then assessed by comparing their corresponding context profiles based on three different weighting schemes using traditional Cosine similarity metrics. To evaluate proposed relatedness approach, three well known benchmark datasets are used and it is shown that Wikipedia article contents can be used effectively to compute term relatedness. The experiments demonstrate that the proposed approach is computationally cheap as well as effective when correlated with human judgments.

1 Introduction

Semantic relatedness is the process of quantifying the extent of semantic connection between two textual units [1,2]. Semantic relatedness is a well explored area. Many researchers have attempted to solve this problem by taking into account various aspects such as statistical relatedness, lexical relations, text contents, rhetorical relations and using external sources of world knowledge such as thesaurus, lexical databases, dictionaries and encyclopedia. Consequently, this problem is widely studied in a variety of applications ranging from computational linguistics to NLP and web mining to intelligent web. There are various applications of semantic relatedness in text summarization [3], information retrieval [4], topic identification [5,6], automatic keyphrase extraction [7], topic indexing [8], word sense disambiguation [9,10,11], document clustering [12,5] and spelling correction [13].

Traditional way of computing text relatedness is to represent context of individual words in a multidimensional space and computing the distance between their corresponding vectors. This paper introduces a new model of relatedness called *Context Profile based Relatedness (CPRel)*. CPRel improved the context

representation by constructing the context profile of each concept based on certain features derived from Wikipedia. The proposed research focuses on computing semantic relatedness of individual words. However, it can be conveniently used for text relatedness as well. It should also be noted that this work focuses on semantic relatedness computation rather then semantic similarity which is less generalized and is based on specific lexical relations such as synonymy, hypernymy or hyponymy. In all of the vector space inspired approaches, the selection of the high-dimensional context space plays a vital role in controlling the performance of relatedness computation. Proposed research analyzed this aspect of relatedness and with the help of simple features achieved a performance comparable to other well known Wikipedia based approaches.

The rest of the paper is organized as follows. Section 2 categorizes existing semantic measures proposed in literature and the corresponding research done in each category. Section 3 proposes and discusses a new relatedness approach. Section 4 analyzes the performance of proposed methodology using three well known datasets. Comparison of proposed approach with other existing strategies and discussion on results are also included in the same section. Finally, section 5 concludes this research and discusses some future research directions.

2 Related Work

With the exponential growth of World Wide Web and ever increasing importance of retrieving relevant information from web, contextual relatedness has become a critical research area. Prior work on relatedness computation can be divided into two main streams: *Statistical techniques based approaches*, where text content and corpus features are statistically analyzed to compute relatedness scores and *external knowledge source based approaches*, where repositories of human knowledge are used as a source of background knowledge to support relatedness computation.

Early research work based on Statistical techniques, introduced the concept of distributional similarity[14,15]. Later, Latent Semantic Analysis (LSA) [16] was proposed as a dimensionality reduction technique where latent concepts are represented by most prominent dimensions in the data using Singular Value Decomposition (SVD). Similarly, Hoffman proposed Probabilistic LSA [17] that constructs a low dimensional concept space. Another statistical technique used for relatedness computation is Latent Dirichlet Allocation (LDA)[18]. LDA represents a document as a mixture of words where each word is attributable to one of the document topics. Sun et al. [19] used LDA based Fisher Kernel for text segmentation.

Various attempts were made to incorporate human knowledge in a structured way in relatedness computation using external knowledge sources such as knowledge bases, dictionaries, thesauri and lexical databases. Ponzetto and Strube [2] used Wikipedia category network and calculated various statistical and structural measures from Wikipedia concepts. Yeh et al. [20] constructed Wikipedia graph and applied random walk with personalized page ranks to compute se-

mantic relatedness for words and texts. Gabrilovich and Markovich [21] proposed Explicit Semantic Analysis (ESA) to incorporate human knowledge into relatedness computation by constructing concept vectors and comparing them using Cosine Similarity. Milne and Witten [22] used Wikipedia hyperlink structure to compute semantic relatedness based on in-link and out-link overlaps. Temporal Semantic Analysis (TSA) [23] was proposed to incorporate temporal dynamics to enhance text relatedness models. TSA represented each input word as a concept vector and extended static representation with temporal dynamics. Jabeen et al. [24] used Wikipedia hyperlinks and disambiguation pages for relatedness computation. They used Dice Coefficient inspired measure of relatedness. Halawi et al. [25] proposed Constrained Learning of relatedness in which they learned a suitable word representation in a latent factor space. Hassan and Mihalceae [1] introduced Salient Semantic Analysis (SSA) by modeling frequently co-occurring words in a contextualized profile for each word. They only used words with high saliency or relevance to the document. Their approach works for both word pairs and text pair relatedness computation. Liu et al. [26] incorporated UMLS and WordNet definitions to generate context vectors for relatedness computation on biomedical data. Navigli and Ponzetto [27] proposed a graph based multilingual approach to compute semantic relatedness. They used BabelNet, a multilingual lexical knowledge source, to construct sub graphs for a word pair in different languages and computed semantic relatedness based on the subgraph intersection.

Proposed approach effectively bridges previously mentioned two research streams by augmenting the relatedness computation with statistics derived from Wikipedia as an external knowledge source. Frequency of occurrence and link probability are used as statistical features driven from Wikipedia article contents. Proposed research is similar to ESA with three main distinctions: First, proposed approach considers a different context for each input word, based on Wikipedia article contents. Second, proposed approach is computationally cheap as it does not preprocess the entire Wikipedia like ESA and third, proposed approach does label pruning to filter out unwanted context and showed that this technique is quite effective in improving the performance of proposed relatedness measure.

3 Context Profile based Relatedness Computation

The idea behind the proposed relatedness computation method is to construct a context profile of each input word based on the corresponding Wikipedia article. Label pruning is performed to weed out all unnecessary labels. In the context profile, each label is assigned a weight based on a hybrid weighting scheme. Semantic relatedness of word pair is then assessed by comparing their corresponding context profiles using traditional Cosine similarity metrics. The work flow of the proposed method is shown in Figure. 3.

To compute the relatedness score of two given terms, the first step is to identify their corresponding Wikipedia articles. The shear size of Wikipedia is

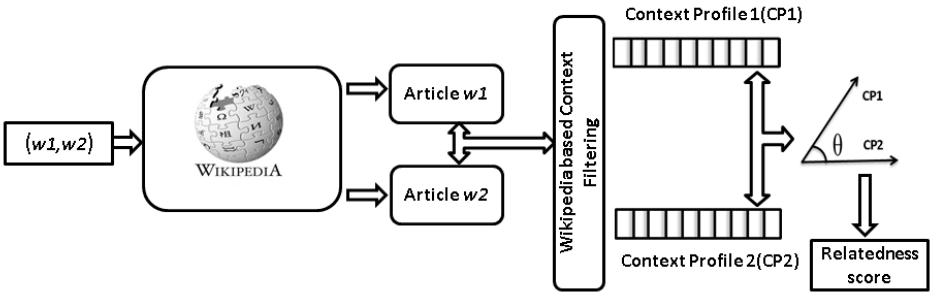

Fig. 1. Framework for Context Profile based Relatedness Computation

big enough to cope with the knowledge requirement of all datasets used in this research.

A series of preprocessing steps is followed to convert Wikipedia articles from *MediaWiki* format to plain text. These steps involve stripping external links, non-article links, special characters and extra spaces. After the article matching and preprocessing phase, article contents are gone through a number of filtering steps to eliminate unnecessary words from the context profile of each word.

Classical ways of text filtering is to weed out all stop words for which the list of common English words[1] is used. Stemming is also performed to convert inflectional words to their roots. N-grams up to 3-grams are extracted from article contents. Clearly, many n-grams would be of no help in supporting the context of a specific term. To prune such words, Wikipedia labels are used. In Wikipedia anchor texts, also called labels, are the hypertexts used to link one Wikipedia article to other articles within the same context. This way all the Wikipedia articles are linked to other articles through the hyperlink structure, making Wikipedia a graph of links. These anchor texts are a very good source of encoding synonyms and other variations in the title of an article. They are an additional source of finding those synonyms which are not covered by the redirects. They are an extremely useful component of Wikipedia because Wikipedia contributors modify them according to the context of the article in which they are used. They not only encode the synonyms and surface forms but also the polysemy and the likeliness of each sense [28]. Matching with Wikipedia labels is a good way of judging whether a word or phrase is useful or not. *Link Probability* (LP) [11] is a proven measure to signify keyphraseness of a word. This research uses Wikipedia labels and link probability for context filtering. filtering of unwanted context is performed in two phases. In the first filtering pass, all words which are not valid Wikipedia labels are discarded, leaving only those keywords that which match with Wikipedia labels. In the second pass, all labels having

[1] available at http://www.db-net.aueb.gr/gbt/resources/stopwords.txt

LP values below a certain cutoff threshold α are discarded. So, $k \epsilon CP(w1)$ if $LP(k) > \alpha$ where, $CP(w1)$ is the context profile of input word $w1$ and $LP(k)$ represents the LP value of a label k. Labels with LP value above threshold α are used to populate the context profile of each input word. This way context profile of each input word is represented in high dimensional space of Wikipedia labels.

3.1 Normalized Term frequency and Link Estimation

Each label in the context profile of each input word is assigned a hybrid weight based on two features:

- **Term frequency**: If a word occurs in a good proportion to the total size of a specific article then it is considered important for that article. Based on this assumption, Term Frequency (NTF) of a word w is computed as the number of times w occurs in a specific Wikipedia article normalized by the size of that article and is given by:

$$TF(w) = \frac{Count(w)}{|W|} \tag{1}$$

  Where $|W|$ is the total number of words in the article.

- **Link probability**: If a keyword occurs more number of times as a label in Wikipedia then it is significant. Based on this assumption, Link probability is used to signify the importance of a keyword as a label. It is the defined as an estimation of probability of a keyword to be used as a link in Wikipedia. It is defined as a ratio of the number of Wikipedia documents having a keyword as a link to the number of Wikipedia documents in which that keyword occurs in any form (as a link or a word) and is given by:

$$P(keyword|w) = \frac{count(Dkey)}{count(Dw)} \tag{2}$$

Where, $count(Dkey)$ represents the number of documents having a word w as an label and $count(Dw)$ is the number of documents in which the word appears. In general, the more generic a Wikipedia label is, the less is its link probability. So a label *Car* gets a lower LP value(0.01) than a label *Sports Car*, which gets a lower Lp value (0.17) than *Ferrari* with LP value 0.27, whereas, most generic labels such as *the* gets extremely lower LP value (8.7×10^{-6}).

These two features are modified and combined to assign weights to individual labels of context profiles. So, after stemming, each root word r is assigned a weight w based on individual weights of its inflectional words set [*w1,w2....wn*].

$$w(r) = LE(r) \times NTF(r) \tag{3}$$

where, *Normalized Term Frequency (NTF)* is defined as the sum of frequencies of all the inflectional words divided by total number of words in an article and is given as:

$$NTF(r) = \frac{\sum_{i=1}^{k} TF(w_i)}{|W|} \tag{4}$$

In general, *NTF* is good for finding out frequently occurring relevant words in a document but its not always helpful. Some of the labels may still exist in an article with high frequency count but of not much relevance. To counter such keywords, *Link Estimation (LE)* of a root word r is used as the measure of popularity of a root word being used as a link in the whole corpus. It is defined as the ratio of sum of link document count (Number of documents where the word occur as a link) of each inflectional form to the sum of total document count (Number of document where a word occurs at all) of them. *LE* is computed as below:

$$LE(r) = \frac{\sum_{i=1}^{k} count(Dkey_{w_i})}{\sum_{i=1}^{k} count(Dw_{w_i})} \tag{5}$$

where, k represents the number of inflectional forms of a root word r. This measure penalizes all unwanted common words which succeeded in passing through stop word and label filters and have higher *NTF*.

4 Evaluation

For relatedness computation based on proposed approach, the version of Wikipedia released in July 2011 is used. At this point, it contains 33GB of uncompressed XML markup which corresponds to more than five million articles, sufficiently covering all concepts for which manual judgment are available.

According to Budanitsky and Hirst [29,30], there are three methods of semantic relatedness evaluation: *Mathematical analysis*, where formal properties of relatedness measure are assessed, *application specific evaluation*, where the measure is applied in a real world application and tested indirectly and *comparison with human judgment*, where human judgments are used as gold standard for evaluation. Third method is the most widely used and best suited application independent evaluation method for relatedness computation. Proposed research also followed the same method for evaluation of relatedness computation.

In this experiment, three standard datasets, which have been widely used in the existing relatedness research, are used:

R&G dataset: Rubenstein and Goodenough (R&G) dataset consists of 65 words pairs sorted in an increasing order of relatedness. These 65 word pairs are scored by 51 human judges on a scale of 0-4 where 0 means unrelated and 4 means exactly the same.

M&C dataset: Miller and Charles dataset is a noun subset of R&G dataset and consists of 30 word pairs which are scored by 38 human subjects on the scale of 0-4.

WordSimilarity-353: WordSimilarity-353 also known as Finkelstein-353 is a dataset of 353 word pairs scored by 13 human experts on a scale of 0-10. It also includes 30 word pairs of M&C dataset but unlike M&C it includes diverse range of word pairs from proper nouns like "Yasser Arafat" to phrases like "Wednesday News" and abbreviations like "FBI" and "OPEC" which adds extra difficulty to the relatedness measure evaluation.

Many word pairs in these sets include ambiguous word like (*Crane,tool*). Since disambiguation is beyond the scope of this research so manually disambiguated versions of M&C and WordSimilarity-353[2]datasets were used and R&G dataset was manually disambiguated based on Wikipedia articles. Some of word pairs in these datasets do not have corresponding Wikipedia articles so such word pairs were also removed from each dataset, resulting in 24 word pairs in M&C, 58 word pairs in R&G dataset and 314 pairs in WordSimilarity-353 dataset.

Table 1. Best performance of three variants of CPRel on three benchmark datasets

Dataset	CPRel(hybrid)		CPRel(NTF)		CPRel(LE)	
	r_s	r	r_s	r	r_s	r
M&C	0.83	0.70	0.83	0.79	0.81	0.66
R&G	0.79	0.66	0.79	0.71	0.79	0.63
WordSimilarity-353	0.66	0.43	0.69	0.52	0.64	0.53

Two other variants of proposed relatedness measure are used to compare and analyze the influence of individual and combined features on the relatedness computation: CPRel with NTF weighting scheme and CPRel with LE weighting scheme. The comparison of CPRel with three different weighting schemes using Spearman's Correlation Coefficient (r_s) and Pearson Correlation (r) is shown in Table. 1. In general, the performance of all three variants of CPRel is highest on M&C dataset using both correlation variables whereas, both CPRel (Hybrid) as well as CPRel (NTF) achieved highest correlation values on M&C dataset overall. This solidify the fact that the proposed approach works quite well on noun-noun word pairs since M&C is a noun subset of R&G dataset.

One of the main features of proposed method is context filtering based on LP cutoff threshold. Performance of the proposed system varies according to the chosen cutoff threshold value. To understand the impact of label pruning on improving relatedness, the behavior of proposed relatedness approach on various cutoff threshold values is tested. For each dataset, six different threshold values (between 0 and 1) were chosen randomly and both Spearman's Correlation and Pearson's Correlation values were computed for CPRel with three different weighting schemes. The effect of cutoff threshold on each of the dataset is shown

[2] The manually disambiguated WordSimilarity-353 dataset is available at: http://www.nzdl.org/wikipediaSimilarity

in Table. 5. Only the best and worst r_s values of each variant of CPRel are shown[3].

Table 2. Performance variations of three versions of CPRel on M&C dataset

LP Cutoff Threshold	CPRel(hybrid)		CPRel(NTF)		CPRel(LE)	
	r_s	r	r_s	r	r_s	r
0.001	0.83	0.70	0.80	0.75	0.81	0.66
0.005	0.76	0.70	0.83	0.79	0.81	0.66
0.01	0.81	0.70	0.77	0.74	0.81	0.66
0.05	0.76	0.70	0.74	0.69	0.81	0.66
0.1	0.79	0.64	0.76	0.74	0.81	0.66
0.5	0.68	0.70	0.69	0.65	0.70	0.65

Table 3. Performance variations of three versions of CPRel on R&G dataset

LP Cutoff Threshold	CPRel(hybrid)		CPRel(NTF)		CPRel(LE)	
	r_s	r	r_s	r	r_s	r
0.001	0.78	0.66	0.67	0.70	0.79	0.63
0.005	0.78	0.66	0.64	0.71	0.79	0.63
0.01	0.78	0.66	0.76	0.70	0.78	0.63
0.05	0.76	0.65	0.78	0.66	0.78	0.63
0.1	0.71	0.66	0.70	0.66	0.77	0.63
0.5	0.79	0.65	0.79	0.65	0.79	0.63

Table 4. Performance variations of three versions of CPRel on WS-353 dataset

LP Cutoff Threshold	CPRel(hybrid)		CPRel(NTF)		CPRel(LE)	
	r_s	r	r_s	r	r_s	r
0.001	0.65	0.43	0.62	0.47	0.61	0.53
0.005	0.65	0.43	0.59	0.51	0.60	0.53
0.01	0.66	0.43	0.69	0.51	0.64	0.53
0.05	0.60	0.42	0.61	0.52	0.60	0.53
0.1	0.62	0.41	0.58	0.35	0.58	0.51
0.5	0.61	0.42	0.59	0.48	0.59	0.50

In general, high cutoff value means more keywords are filtered out in the context pruning phase. The effect of LP cutoff threshold on performance of CPRel using M&C dataset is shown in Table. 2. Clearly, with all three variants of

[3] Bold values in Table. 5 indicate a specific performance pattern of CPRel. Clearly, on M&C dataset the performance of all variants of CPRel was consistently worst on LP cutoff threshold value of 0.5 whereas for R&G dataset, the best correlation values were achieved on α=0.5. In case of WS-353, the best performance was observed on α=0.01

Table 5. Correlation based Performance of three weighting schemes of CPRel on three benchmark datasets

Dataset	CPRel(hybrid)			CPRel(NTF)			CPRel(LE)		
	Result	Cutoff	r_s	Result	Cutoff	r_s	Result	Cutoff	r_s
M&C	Best	0.001	0.83	Best	0.005	0.83	Best	0.001	0.81
	Worst	**0.5**	0.68	Worst	**0.5**	0.69	Worst	**0.5**	0.70
R&G	Best	**0.5**	0.79	Best	**0.5**	0.79	Best	**0.5**	0.79
	Worst	0.1	0.71	Worst	0.005	0.64	Worst	0.1	0.77
WS-353	Best	**0.01**	0.66	Best	**0.01**	0.69	Best	**0.01**	0.64
	Worst	0.05	0.60	Worst	0.1	0.58	Worst	0.1	0.58

CPRel on M&C datset, the highest correlation value is achieved on lowest cutoff value and vice versa. But, in case of R&G dataset, shown in Table. 3, the behavior of CPRel was entirely opposite. All variants of CPRel, achieved the highest correlation on highest threshold value. For CPRel (Hybrid) and CPRel(LE) on R&G dataset , the lowest and highest correlation values were observed on the top two threshold values, though the difference among correlation values on all cutoffs was not very significant. For WordSimilarity-353 dataset, as indicated by Table. 5, the highest correlation value was achieved on 0.01 threshold for all variants of CPRel but for other thresholds, their correlation changed randomly. Overall, the performance of CPRel (LE) remained consistent on all three datasets with minor changes in correlation values. It was found that the performance of each dataset was different on different threshold values. This elucidated the fact the there is no unique threshold value which could be used as a discriminator for good or bad relatedness performance on all datasets.

In another experiment, performance of each approach with and without Context Filtering (CF) was compared. In case, when no CF was applied and all the words that matched to Wikipedia labels were considered, the behavior of each dataset was again different. For CPRel (NTF) there was a significant improvement in the correlation values (r_s) on all three datasets when CF was applied. On average, there was an increase of 15% in the correlation value on all three datasets, weighted by their sizes. In case of other two approaches, there is no significant improvement in correlation values with CF. It means that the CPRel (NTF) performs quite well with context filtering.

Table 6. Best performance of three variants of CPRel on three benchmark datasets with and without Context Filtering (CF)

Dataset	CPRel(hybrid)		CPRel(NTF)		CPRel(LE)	
	CF	No CF	CF	No CF	CF	No CF
M&C	0.83	0.83	0.83	0.72	0.81	0.81
R&G	0.79	0.78	0.79	0.62	0.79	0.78
WordSimilarity-353	0.66	0.64	0.69	0.54	0.64	0.62

It was also found that the relatedness computation not only depends on the context filtering but also on the nature of the dataset and type of weighting scheme. In comparison with other existing Wikipedia based approaches, shown in Table. 7, CPRel performed significantly better then ESA on M&C dataset. For other two datasets, it performed better then other Wikipedia based approaches but was still behind ESA. There are two reasons for this: First, to identify the context of each input word CPRel focus only on corresponding Wikipedia article contents whereas, ESA makes a good use of whole Wikipedia corpus to mine the context of each word. This is a limitation of CPRel and in future it is intended to modify this approach so that the context spread of each word in the whole corpus may effectively be used to improve relatedness computation. Second, CPRel approach works quite well on noun word pairs in particular, justifying best performance on M&C dataset (which is a noun subset of R&G dataset). The advantage of CPRel is that it does not require preprocessing of the whole Wikipedia corpus like ESA which is computationally quite expensive and laborious. It is proved that good relatedness scores can be achieved following a simple and computationally inexpensive approach. Another advantage of CPRel is that it can be effectively used for document relatedness also. As a future work, it is intended to test document relatedness based on the same approach. To top this, other features of Wikipedia such as hyperlink structure, category network and corpus statistics etc. could be considered to improve context profiles for getting better relatedness scores.

Table 7. Best performance comparison of CPRel with existing Wikipedia based approaches on three benchmark datasets

Dataset	WikiRelate	ESA	WLM	**CPRel(proposed)**
M&C	0.45	0.73	0.70	**0.83**
R&G	0.52	0.82	0.64	**0.79**
WordSim-353	0.49	0.75	0.69	**0.64**

5 Conclusions

This paper describes CPRel, a measure of semantic relatedness using Wikipedia. Semantic relatedness of words is computed by constructing context profile of each word based on Wikipedia article content and labels. The influence of different factors like types of weighting scheme, nature of dataset, nature of knowledge source and impact of cutoff threshold value on the performance of relatedness computation was analyzed. The impact of a cutoff threshold LP value on each weighting scheme was tested using various threshold values and it was found that context filtering was helpful in improving the relatedness scores in case of CPRel (NTF). When evaluated on three benchmark datasets of term relatedness, CPRel performed quite well in comparison with other Wikipedia based approaches.

References

1. Hassan, S., Mihalcea, R.: Semantic relatedness using salient semantic analysis. In: AAAI. (2011)
2. Strube, M., Ponzetto, S.P.: Wikirelate! computing semantic relatedness using wikipedia. In: Proceedings of Association for the Advancement of Artificial Intelligence(AAAI). (2006)
3. Barzilay, R., Elbadad, M.: Using lexical chains for text summarization (1997)
4. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. In: Readings in information retrieval. Morgan Kaufinann, USA (1997) 323–328
5. He, X., Ding, C.H.Q., Zha, H., Simon, H.D.: Automatic topic identification using webpage clustering. In: Proceedings of IEEE International Conference on Data Mining(ICDM). (2001) 195–202
6. Coursey, K., Mihalcea, R.: Topic identification using wikipedia graph centrality. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. NAACL-Short '09 (2009) 117–120
7. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: Proceedings of the 17th ACM conference on Information and knowledge management. CIKM '08 (2008) 509–518
8. Medelyan, O., Witten, I.H., Milne, D.: Topic indexing with Wikipedia. In: Wikipedia and Artificial Intelligence: An Evolving Synergy. Papers from the 2008 AAAI Workshop, Menlo Park, CA, USA, Proceedings of AAAI (2008) 19–24
9. Agirre, E., Unibertsitatea, E.H., Rigau, G.: A proposal for word sense disambiguation using conceptual distance. In: Proceedinds of 1st international conference on recent advances in natural language processing (RANLP). (1990)
10. Patwardhan, S., Banerjee, S., Pedersen, T.: Using measures of semantic relatedness for word sense disambiguation. In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Mexico (2003) 241–257
11. Mihalcea, R., Csomai, A.: Wikify!: linking documents to encyclopedic knowledge. In: Proceedings of the sixteenth ACM conference on information and knowledge management. CIKM '07, New York, NY, USA (2007) 233–242
12. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: KDD Workshop on Text Mining. (2000)
13. Budanitsky, A., Hirst, G.: Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In: Workshop on wordNet and other lexical resources, second meeting of the North American Chapter of the Association for Computational Linguistics. (2001)
14. Dagan, I., Lee, L., Pereira, F.C.N.: Similarity-based models of word cooccurrence probabilities. Journel of Machine Learning Research **34** (1999) 43–69
15. Lee, L.: Measures of distributional similarity. In: Proceedings of the ACL. (1999) 25–32
16. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. American Society for Information Science, JASIS **41** (1990) 391–407
17. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '99, New York, NY, USA (1999) 50–57

18. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3** (2003) 993–1022
19. Sun, Q., Li, R., Luo, D., Wu, X.: Text segmentation with lda-based fisher kernel. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, Stroudsburg, PA, USA (2008) 269–272
20. Yeh, E., Ramage, D., Manning, C.D., Agirre, E., Soroa, A.: Wikiwalk: random walks on wikipedia for semantic relatedness. In: 2009 Workshop on Graph-based Methods for Natural Language Processing. TextGraphs-4 (2009) 41–49
21. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. (2007) 1606–1611
22. Milne, D., Witten, I.H.: 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. (2008) 25–30
23. Radinsky, K., Agichtein, E., Gabrilovich, E., Markovitch, S.: A word at a time: computing word relatedness using temporal semantic analysis. In: Proceedings of the 20th international conference on World wide web. WWW '11, New York, NY, USA (2011) 337–346
24. Jabeen, S., Gao, X., Andreae, P.: Harnessing wikipedia semantics for computing contextual relatedness. In: PRICAI 2012: Trends in Artificial Intelligence. Volume 7458 of Lecture Notes in Computer Science. Springer (2012) 861–865
25. Halawi, G., Dror, G., Gabrilovich, E., Koren, Y.: Large-scale learning of word relatedness with constraints. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD '12, New York, NY, USA (2012) 1406–1414
26. Liu, Y., McInnes, B.T., Pedersen, T., Melton-Meaux, G., Pakhomov, S.: Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, umls and wordnet. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. IHI '12, New York, NY, USA (2012) 363–372
27. Navigli, R., Ponzetto, S.P.: Babelrelate! a joint multilingual approach to computing semantic relatedness. In: AAAI. (2012)
28. Milne, D.: An open-source toolkit for mining wikipedia. In: Proceeding of New Zealand Computer Science Research Student Conference. Volume 9. (2009)
29. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of lexical semantic relatedness. Comput. Linguist. **32** (2006) 13–47
30. Zesch, T., Gurevych, I.: Automatically creating datasets for measures of semantic relatedness. In: Proceedings of the Workshop on Linguistic Distances, ACL. (2006) 16–24