Normalized Relevance Distance – A Stable Metric for Computing Semantic Relatedness over Reference Corpora

Christoph Schaefer[1] and **Daniel Hienert**[2] and **Thomas Gottron**[3]

Abstract. We propose the Normalized Relevance Distance (NRD): a robust metric for computing semantic relatedness between terms. NRD makes use of a controlled reference corpus for a statistical analysis. The analysis is based on the relevance scores and joint occurrence of terms in documents. On the basis of established reference datasets, we demonstrate that NRD does not require sophisticated data tuning and is less dependent on the choice of the reference corpus than comparable approaches.

1 Introduction

The knowledge of the semantic relatedness of two terms is of importance in many applications in the areas of linguistics, information retrieval and text mining. While humans can easily assess the semantic relatedness for terms they are familiar with, this task is difficult to solve for automatic approaches. Research has addressed this issue over the last decades using various methods. From statistical analysis of word co-occurrence, over models for latent topic spaces, approaches based on lexical databases, to solutions involving Web search engines as tools for analyzing the Web as a corpus.

Two well-known and well-performing approaches based on statistical analysis of reference corpora are the *Normalized Google Distance* (NGD) [2] and *Explicit Semantic Analysis* (ESA) [4]. NGD uses hit counts provided by Web search engines to estimate the probability of two terms to appear in a Web document individually as well as the probability to co-occur. ESA, instead, uses Wikipedia as a controlled reference corpus for computing a vector of relevance scores between a term and all articles in the corpus. The semantic relatedness of two terms is computed using the cosine similarity of their vectors of relevance values. ESA provides high quality values for semantic relatedness, outperforms NGD and has been adopted in many applications and methods. A drawback of ESA is that its performance depends on the choice and quality of the reference corpus. Thus, various papers have investigated ideal compositions for the reference corpus, preprocessing, and data tuning methods [6, 20, 22].

In this paper, we extend the theory behind NGD to incorporate relevance scores obtained over a controlled reference corpus. Our approach, the *Normalized Relevance Distance* (NRD), combines relevance weights of terms in documents and the joint relevance of the terms to identify not only co-occurrence but also correlation of importance of the terms in documents.

[1] University of Koblenz–Landau, Germany, email: chrisschaefer@uni-koblenz.de
[2] GESIS - Leibniz Institute for the Social Sciences, Germany, email: Daniel.Hienert@gesis.org
[3] University of Koblenz–Landau, Germany, email: gottron@uni-koblenz.de

In our evaluation, we empirically show that NRD is competitive with ESA in terms of computing semantic relatedness and significantly outperforms NGD when using the same reference corpus. Furthermore, we show that NRD is less susceptible to the choice and data tuning of the reference corpus.

The rest of this paper is structured as follows: In Section 2, we review related work and give an overview of state of the art approaches for computing semantic relatedness. We present our own approach in Section 3 and describe our empirical evaluation in Section 4. We conclude the paper in Section 5 with a summary of our findings and an outlook at future work.[4]

2 Related Work

Related work on semantic relatedness can be divided in knowledge-based, corpus-based and hybrid methods.

The use of knowledge bases such as thesauri or lexical databases is central to various approaches. Many essential contributions rely, for instance, on WordNet [15]. Such measures for semantic relatedness use different properties of the semantic network in WordNet, for instance, shortest path or PageRank information [1, 11, 21]. These approaches have the limitation that semantic relatedness can only be computed for concepts which are found in the network.

An alternative approach is to mine semantic relatedness from documents in a suitable reference corpus. Web-based measures use the Web as reference corpus to compute semantic relatedness. The rationale for using the Web is the huge amount of text in different languages, which can be used to extract new lexical semantic knowledge. However, directly accessing all information available on the Web is hardly feasible from a computational point of view. A common solution to circumvent this problem is to leverage the index of a Web search engine. NGD, for example, makes use of hit counts for terms appearing individually and together to compute an information theoretic distance measure. We will describe NGD in more detail in Section 3. As an alternative to the entire Web, also Wikipedia is frequently used as reference corpus due to its wide scope, high quality and public availability. Salient Semantic Analysis [9] and the Wikipedia Link-based Measure [30], for instance, exploit anchor texts and the link structure between different Wikipedia concepts to compute semantic relatedness between terms. WikiRelate! [29] and its successor WikiNet [18] search for appropriate Wikipedia articles for two related terms and then compute semantic relatedness based on the paths within the category hierarchy or text overlaps.

[4] Our approach has been prototypically implemented and is available at `https://github.com/chrip/SemanticRelatedness`

ESA uses weighted vectors of Wikipedia concepts to represent terms. Semantic relatedness is then computed by comparing these vectors with a cosine metric. For this approach, an inverted index has to be created that maps terms to concepts, which needs preprocessing of the whole text corpus. ESA can be applied to both, single words and text fragments. The overall performance of ESA can be optimized by a number of factors. For example, by choosing an adequate article selection strategy [20] or by the topic composition and the size of the index collection [6]. A pruning of the concept vector entries and an improved length normalized tf-idf score has been used by [28]. In [22], the authors apply several optimization measures such as replacing tf-idf by BM25F, pruning the ordered term-document vectors to 250 entries, using only the 10,000 longest articles, indexing only the top 100 terms of each article and only terms which occur at least in 10 articles. Further optimizations techniques incorporate semantic properties like article link structure and categorization in their approach [26] or use PageRank weights instead of tf-idf scores [17].

A survey of the quality of different approaches for determining semantic relatedness is given in [19]. The authors show that hybrid measures, which use multiple corpora or combine lexical knowledge bases and reference corpora, outperform other measures for semantic relatedness. In fact, currently the state of the art approaches for computing semantic relatedness between terms achieve their improvements by combining several sources of background knowledge. Temporal Semantic Analysis (TSA) [23] improves ESA by additionally incorporating the co-occurrence of terms over time. To this end, the authors use Wikipedia and The New York Times newspaper archive from the past 130 years. In [31] a refinement of ESA is presented, which combines multiple vector space models build over a text corpus, thesauri and Web search results. CLEAR [7], which stands for Constrained LEArning of Relatedness, achieves its high performance with a machine learning algorithm trained on data obtained from WordNet and three text corpora from very different domains. However, improving single corpus based metrics—as presented in this paper—can be seen as a foundation for improvements in multi corpus or hybrid approaches.

3 Normalized Relevance Distance

We now introduce our NRD approach and develop it using the theoretical background of NGD. In Section 3.2, we will go into details of how we implemented NRD on top of an inverted index.

3.1 Theoretical Motivation

At the core of NGD lies the *Normalized Compression Distance*. The Normalized Compression Distance measures the distance between two strings on the basis of a compression algorithm [2]. This distance metric compares two strings x and y using the length of the compressed encoding of their concatenation xy in relation to the length of their individual encodings. If the function $C(x)$ provides the length of the encoding of a string x using a given compression algorithm then the Normalized Compression Distance for strings x and y is defined as:

$$\mathrm{NCD}(x,y) = \frac{C(xy) - \min\left(C(x), C(y)\right)}{\max\left(C(x), C(y)\right)} \tag{1}$$

The numerator in Equation (1) indicates that the distance of x and y is small if the length overhead of encoding xy is small compared to the shortest encoding of either of the two strings alone. The denominator provides a normalization factor to ensure consistent values independent of the length of x and y.

The Normalized Google Distance follows the approach of employing a prefix-free code [10] as compression scheme underlying $C(x)$. The advantage of this approach is that Shannon's source code theorem [27] provides an optimal lower bound for the length of the prefix-free code words if a distribution $P(x)$ over the strings is known. In this case, it is possible to use the entropy value $-\log(P(x))$ as the optimal length of the encoding $C(x)$. Furthermore, in the context of NGD we are not interested in arbitrary strings x, but rather in terms t_x. Accordingly, we need to estimate a probability distribution $P(t_x)$ over terms.

For NGD, this distribution $P(t_x)$ is estimated from the hit count given by web search engines (e.g. Google) when querying for documents containing the term t_x. If the function $f(t_x)$ provides the count of relevant documents returned for query t_x and N is the aggregated total number of documents provided for all terms[5], then the probability of t_x can be estimated by $P(t_x) = \frac{f(t_x)}{N}$. The concatenation $t_x t_y$ in this setting is interpreted as querying for the boolean conjunction $t_x \wedge t_y$, i.e $f(t_x, t_y)$ provides the number of documents containing both terms.

The lower boundary length for an optimal prefix-free code word for t_x is then $\log(N) - \log(f(t_x))$. Substituting C in Equation (1) with the corresponding value for the length of an encoding leads to the following final formula for NGD:

$$\mathrm{NGD}(t_x,t_y) = \frac{\max(\log(f(t_x)),\log(f(t_y))) - \log(f(t_x,t_y))}{\log(N) - \min\left(\log(f(t_x)),\log(f(t_y))\right)} \tag{2}$$

NGD as a semantic relatedness metric can easily be transferred to any other indexed corpus than Google's search index. For instance, it can be computed using a search index built over the documents of Wikipedia. We will refer to this variation as *Normalized Wikipedia Distance* (NWD) and use it for comparison over a controlled reference corpus. Equation (2) remains unchanged for NWD and the frequency functions f are still based only on the binary notion of term presence in a document.

However, it is long known in Information Retrieval that words can also occur in a document "by chance" [8]. In this case, a term t_x is not really relevant to the description of the document. Accordingly, one should not consider these documents in estimating the probability $P(t_x)$, or at least to a lower degree. Probabilistic relevance models for Information Retrieval have been developed to identify the probability of relevance of a document and a specific term. The history of the developed models goes far beyond the scope of this paper and we refer to [24] for a detailed summary of the findings.

A result of the analysis are tf-idf based models assigning a weight to each term in each document. These weights can be considered a metric for the probability of relevance for a given term and document[6]. In this way, we can specify the probability of term t_x to occur in a document to be a joint probability of t_x to appear in a document d and the probability $P(\mathit{Rel}|d, t_x)$ of t_x to be actually relevant for document d. Given again a total number of N documents in the index, this leads to the probability $P(t_x)$ to be estimated by:

[5] N is a large number which is difficult to obtain. However, it has been shown, that the concrete choice of N has no effect on the quality of the results, but simply scales all relatedness scores. Thus, it is in practice often set to the total number of documents in the search index [2].

[6] Due to transformations and simplifications under computational aspects the actual values do not comply with the formal characteristics of a probability density function.

$$P(t_x) = \frac{\sum_{d:t_x \in d} P(Rel|d, t_x)}{N} \quad (3)$$

Using the normalized tf-idf weight $\textit{tf-idf}_{norm}(t_x, d)$ as an approximation for $P(Rel|d, t_x)$ leads to a better approximation of the probability $P(t_x)$. Incorporating this value for $P(t_x)$ into the compression schema in [2] leads to a substitute for the frequency functions f. As a result, in our NRD approach we use the functions f_{NRD} for single and combined terms as follows:

$$f_{NRD}(t_x) = \sum_{d \in D} \textit{tf-idf}_{norm}(t_x, d) \quad (4)$$

$$f_{NRD}(t_x, t_y) = \sum_{d \in D} \textit{tf-idf}_{norm}(t_x, d) \cdot \textit{tf-idf}_{norm}(t_y, d) \quad (5)$$

This leads to an adaptation of Equation (2) and we obtain the final formula for computing NRD as follows:

$$\text{NRD}(t_x, t_y) = \frac{\max(\log(f_{NRD}(t_x)), \log(f_{NRD}(t_y))) - \log(f_{NRD}(t_x, t_y))}{\log(N) - \min(\log(f_{NRD}(t_x)), \log(f_{NRD}(t_y)))} \quad (6)$$

3.2 Implementation

To access relevance scores over terms and documents we leverage the mature and widely adopted text retrieval software Lucene[7]. Lucene implements a length-normalized tf-idf variant as relevance scores [14, p. 86] which suits our needs for estimating the probabilities of relevance.

To ensure that all Lucene scores $\textit{tf-idf}_{Lucene}(t, d)$ are in a range between 0 and 1 we divide all scores by the largest score occurring for term t:

$$\textit{tf-idf}_{norm}(t, d) = \frac{\textit{tf-idf}_{Lucene}(t, d)}{\max\{\textit{tf-idf}_{Lucene}(t, d') \mid d' \in D\}} \quad (7)$$

Lucene makes use of an inverted index which maps each term to the vector of documents in which it occurs. Therefore, all relevance scores for a given term can be accessed very efficiently. As a consequence, our NRD approach is also computationally attractive.

4 Evaluation

We empirically evaluate our NRD approach under three aspects. First of all, we compare its performance in assessing the semantic relatedness of given word pairs with other single corpus-based and non-hybrid approaches. Furthermore, we are interested in the influence of the quality of the reference corpus on NRD—especially in comparison to ESA. In particular, we want to evaluate the impact of data tuning methods applied to the reference corpus as well as corpus size and corpus domain on the performance of NRD.

4.1 Quality in Measuring Semantic Relatedness

With this experiment we want to compare NRD with other single corpus-based and non-hybrid approaches for determining the semantic relatedness of word pairs.

[7] http://lucene.apache.org

Table 1. Evaluation results (Spearman's correlation) of ESA, NWD, and NRD. The upper part of the table shows other non-hybrid metrics for semantic relatedness reproduced from [19] for comparison.

Similarity Measure	MC ρ	RG ρ	WS ρ
Random	0.056	-0.047	-0.122
N-WuPalmer	0.742	0.775	0.331
N-Leack.Chod.	0.724	0.789	0.295
W-NGD-GoogleWiki	0.334	0.502	0.251
C-BowDA	0.693	0.782	0.466
C-SynDA	0.790	0.786	0.491
C-NGD-Factiva	0.603	0.599	0.600
C-PatternWiki	0.461	0.542	0.357
D-ExtendedLesk	0.792	0.718	0.409
ESA	0.793	0.803	0.744
NWD	0.742	0.742	0.743
NRD	**0.811**	**0.821**	**0.756**

Experimental Setup: In our experiments we rely on established datasets for assessing the quality of automatic approaches for computing semantic relatedness of term pairs. For this purpose, we make use of the evaluation framework sim-eval [19]. The sim-eval framework provides three openly available datasets and scripts for the comparison of semantic relatedness measures. The datasets MC [16], RG [25] and WordSim353 [3] cover 30, 65 and 353 term pairs which have been assessed by human experts for their semantic relatedness. The quality of automatic semantic relatedness measures can then be evaluated based on the Spearman's correlation between values provided by human experts and the automatic approach. State of the art approaches that achieve high Spearman's correlation to the RG dataset are [1, 9, 11, 21, 31]. Approaches with high correlation to WordSim353 are [7, 23, 31]. All these methods achieve slightly higher correlation values than our approach. However, we argue in this paper for the simplicity and robustness of our method that uses only one single corpus, is stable over different preprocessing steps and does not use any hybrid or machine learning methods.

In the context of this framework we evaluate NRD, ESA and NWD. To be able to compare our approach in the most competitive setting, we employ an existing weighted inverted index, which is available as part of an open source implementation of ESA [12]. The index contains tf-idf scores as weights and is highly optimized for ESA making use of the most common data tuning techniques.

Results: Our results for the performance of NRD, NWD and ESA are shown in Table 1, the performance metrics for other approaches are taken from [19]. Consistent with other surveys, we observe that ESA performs better than all other single measures evaluated in [19] in assessing the semantic relatedness on the datasets MC, RG and WS; especially in comparison with the Normalized Google Distance on the Google index (W-NGD-GoogleWiki) and the Normalized Google Distance on the Factiva corpus (C-NGD-Factiva).

NWD achieves comparable results to ESA on the WordSim353 dataset, but is clearly outperformed on the smaller datasets MC and RG. However, our novel NRD approach performs best for all of the three datasets for semantic relatedness.

Discussion: NWD, ESA and NRD all make use of a single text corpus representing the background knowledge in the system. They share the same hypothesis that the co-occurrence of two words in the same document indicates a semantic relatedness. From a technical point of view they use an inverted index which provides for each term a list of the documents, i.e. Wikipedia articles, wherein it occurs as well as a relevance weight. The difference between the three ap-

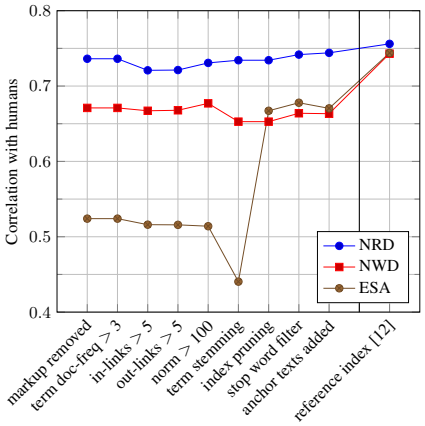

Figure 1. Impact of accumulated preprocessing steps, applied on the WikiPrep2005 dataset.

proaches lies in how they use this information. This difference also explains the performance of the algorithms.

NWD ignores the relevance weights. The frequency functions $f(t_x)$ and $f(t_y)$ only utilize the number of documents in which t_x and t_y appear. The combined frequency function $f(t_x, t_y)$ counts the documents in which both terms are present. Not making use of relevance information leads to the lower performance.

ESA, instead, uses the specific relevance weights for each document. The motivation is to interpret the vector of weights over all Wikipedia documents as a concept vector and to compute similarity in this concept space. This interpretation is easy to follow and ESA has shown to perform well in many settings.

NRD combines the benefits of NGD and ESA. On the one hand, it integrates the metric underlying NGD. On the other hand, it increases the performance of NGD by incorporating relevance weights as done in ESA. The frequency functions $f_{NRD}(t_x)$ and $f_{NRD}(t_y)$ sum up the tf-idf scores of all entries in the concept vectors belonging to t_x and t_y. The combined frequency $f_{NRD}(t_x, t_y)$ is calculated by the scalar product of two vectors, which exhibits strong similarities to the cosine metric used in ESA. The biggest difference lies in the normalization of the two approaches. NRD is based on a normalized information distance, while ESA uses vector length normalization.

We attribute the high quality of both NRD and ESA to this similar use of relevance information. However, ESA has been extended and optimized using several data tuning techniques. The question of the impact of these optimizations motivates our next expriment.

4.2 Reference Corpus: Impact of Data Tuning Methods

In the context of ESA, it has been observed that performance can be improved by tuning the reference corpus and the resulting index. In our previous experiments we used such a fine-tuned reference index. With our second experiment we intend to identify, if and to which degree also NRD benefits from such work-intensive preprocessing of the reference corpus.

Experimental Setup: As reference corpus we used the original dataset employed in [5] which is available for download on the author's website[8]. This dataset is an already slightly preprocessed Wikipedia snapshot from 2005: Wikipedia templates and redirection links are resolved, wiki markup has been removed, date formats are normalized, and meta data is added to each article. We refer to this dataset as "WikiPrep2005".

Starting from this dataset, [5] list further data tuning tasks which they performed to improve the semantic relatedness performance. The tuning tasks given are: (a) considering only terms which occur in more then three articles, (b) requiring articles in the reference corpus to have at least five in- and out-links, (c) requiring the article's minimum word length to be 100, (d) applying a stemming algorithm, (e) performing an index pruning step, (f) removing stop words and (g) cross-document smoothing techniques based on adding anchor texts to referenced article contents. In this list of tasks, the most complex processing step is the index pruning. This pruning step operates on an inverted index, which provides for each search term its ESA vector of tf-idf weights for the documents. This is done using a sliding window algorithm [5], which cuts away the long tail of descending ordered tf-idf scores in the vectors. If, for example, a term occurs in a few documents very frequently and in a large number of other documents only rarely, the documents with only few occurrences are cut away. Please note, that this pruning technique has been empirically motivated by the improved performance of ESA. For NRD there is no theoretic motivation to perform pruning and we omit this step in the data preparations for NRD.

In our experiment we successively perform each of these well-documented data tuning steps on the WikiPrep2005 dataset. After each step we built a Lucene index over the optimized reference corpus to obtain tf-idf scores for each term and document. We then use this index as background knowledge for NRD, NWD and ESA to compute semantic relatedness values for the WordSim353 dataset and compare them to the values provided by human experts. In this way we can measure the impact and improvement obtained by each data tuning method. For comparison, we employed for the last data point in this experiment again the fine-tuned reference index used in the previous experiment.

Results: The plot in Figure 1 shows how the performance of NRD, NWD and ESA is affected by each of the data tuning steps. As one can see, even without any further data tuning, NRD reaches a high correlation of 0.736.

With the same un-optimized reference corpus, ESA obtains a correlation of 0.524. This low value is even higher than the values produced with the ESA reimplementation in [9] where a correlation of 0.435 is reported. While the performance of NRD and NWD is very stable and changes only little, ESA is affected most by the two steps of stemming and index pruning. Applying a stemmer causes the performance of ESA to drop, while the index pruning boosts ESA to perform better than NWD. NRD, however, consistently performs better than ESA.

We also noticed, that for ESA we could not reproduce the correlation value of 0.744 observed with the given fine-tuned reference index. The best ESA results achieved with our implementation are limited to 0.678 after we had successively applied all steps up to the stop word filter. Also for NRD we observe a insignificantly lower performance of 0.744 on our own index, after anchor texts were added, in comparison to the correlation of 0.756 obtained on the pre-computed reference index. This gap, mainly between the results of our ESA implementation and the values obtained with the reference index can be attributed either to a variation for computing tf-idf scores used in the preprocessed corpus or further, not documented data tuning steps.

Discussion: The negative impact of term stemming on ESA can be attributed to semantically different terms being reduced to the same syntactical stem. However, in the overall process, term stemming

[8] `http://www.cs.technion.ac.il/~gabr/resources/code/wikiprep/`

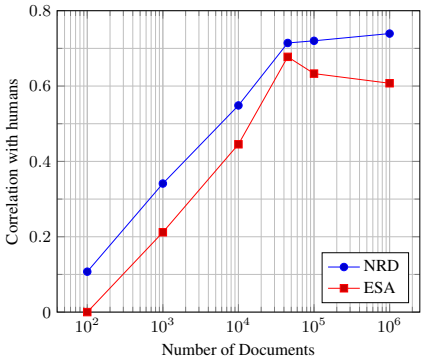

Figure 2. Impact of increasing number of concepts (dataset WikiPrep2005).

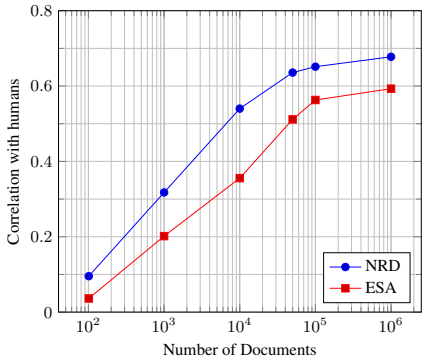

Figure 3. Impact of increasing number of concepts (dataset Wikipedia 2013).

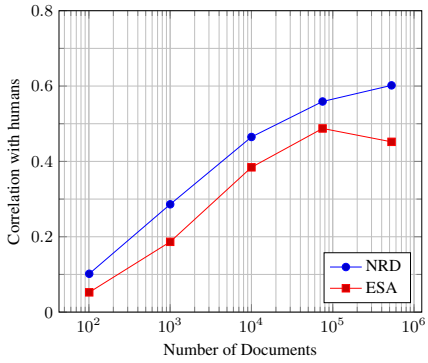

Figure 4. Impact of increasing number of concepts (dataset Reuters corpus 1996-1997).

seems to be of negligible importance. The drop in performance is recovered by the next step of index pruning. Interestingly, if applying index pruning without a prior stemming, the quality of the results does not improve, but is actually marginally below the combination of applying stemming and pruning.

The positive impact of index pruning, instead, can best be explained by reducing noise in the data. Filtering out low relevance scores in the concept vectors probably eliminates articles which contain a term by chance.

We hypothesize, that the fact that none of the data tuning tasks has a large effect on NRD or NWD can again be explained by the conceptually different approach. The sound theoretic foundation of NRD makes it less susceptible to noise. Small deviations in the relevance weights over multiple documents do not have a high impact on the overall probability distribution $P(t_x)$ underlying the assumed compression scheme. In ESA, instead the length and direction of the concept vectors is affected stronger by this noise.

4.3 Reference Corpus: Impact of Domain and Size

In the last experiment, we look at the relation between the size and domain of the reference corpus and the semantic relatedness performance.

Experimental Setup: We implement this experiment using three different text corpora: WikiPrep2005, a more recent Wikipedia snapshot from June 4, 2013, and Reuters CV1 [13] from 1996-1997. We incorporated the more recent Wikipedia snapshot to additionally confirm our observations also over reference corpora from different points in time. For each of these datasets we sampled smaller reference copora of increasing size. Each of the samples were used to build the index structure for NRD and ESA. We evaluated the performance of NRD and ESA on the WordSim353 dataset.

For this experiment, we use only the most important preprocessing steps found in the experiment above, which are applicable on all three corpora. These steps are term stemming, discarding all articles with less than 100 words and the index pruning for ESA. Because of the lack of links in the Reuters dataset, the contraints on the in- and outlink structure cannot be considered. However, as we have seen in Section 4.2 this preprocessing step does not have a noticeable impact on the performance of ESA.

Results: In Figure 2, we see how the performance of NRD and ESA develops when using more and more documents of WikiPrep2005 as reference corpus. ESA reaches a maximum correlation of 0.677 with the human assessments in WordSim353 by employing a subset of 45,000 articles. In this setting, the increase to 100,000 or 1 million articles leads to an even lower performance of ESA, whereas NRD archieves its maximum of 0.739 with 1 million concepts.

On the 2013 Wikipedia snapshot, both algorithms perform best when using 1 million articles (cf. Figure 3). The highest correlation observed for NRD is 0.678, ESA archieves a correlation of 0.593 at its best.

For the Reuters corpus we observed again a performance decrease for ESA after exceeding a certain threshold, which on this corpus is reached at 75,000 articles (cf. Figure 4). In contrast, the performance of NRD is as on the other two corpora always monotonically increasing when increasing the amount of articles. The absolute numbers of 0.487 for ESA and 0.602 for NRD are slightly lower compared to the Wikipedia-based corpora evaluated above.

Discussion: Also on corpora of different size and domain, we observe NRD to perform consistently better than ESA. An interesting fact is that NRD also always benefits from using a larger reference corpus, as the performance increases when adding more documents. For ESA, however, we observed a drop in performance in two of our three experiments, when exceeding a certain size of the reference corpus. We hypothesize that this is an artifact of the index pruning step for ESA. Pruning has a stronger effect when the reference corpus, and thereby the concept vectors gets larger. While in general ESA benefits from pruning the index, this step also potentially discards some valuable information contained in the smaller relevance values.

The observation that using Reuters as reference corpus leads to an overall lower performance for both ESA and NRD is consistent with previous results on ESA. The Reuters corpus, in this case, has a low topic coverage of the terms in the evaluation dataset. Thus, it is more difficult to have reliable statistics about term correlation.

5 Conclusion

In this paper, we presented NRD, a robust approach for computing semantic relatedness between terms. NRD makes use of a reference corpus and extends NGD by incorporating relevance scores. We described the theoretical motivation and showed in an empirical evaluation, that NRD outperforms other single corpus approaches for determining semantic relatedness.

Furthermore, we demonstrated that the quality of NRD does not depend on fine-tuning and optimization of the reference corpus as required, e.g., for an optimal performance of ESA. Finally, we revealed that relative to NGD and ESA, NRD performs consistently better on

text corpora of all sizes and different domains. At the same time, we observed that the performance of NRD increases monotonically with the size of the underlying corpus.

In future work, we will incorporate NRD into hybrid methods like CLEAR and methods making use of multiple corpora like TSA. In this way we will evaluate how the improvements of our single corpus method can boost the state of the art of more complex approaches. Furthermore, first attempts to extend our approach from single words to longer text sequences look promising as well.

Acknowledgments

This work was supported by the EU 7th FP under grant number IST-FP7-288815 in project Live+Gov (liveandgov.eu).

REFERENCES

[1] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa, '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*, NAACL '09, pp. 19–27, Stroudsburg, PA, USA, (2009). Association for Computational Linguistics.

[2] Rudi L. Cilibrasi and Paul M.B. Vitanyi, 'The google similarity distance', *Knowledge and Data Engineering, IEEE Transactions on*, **19**(3), 370–383, (2007).

[3] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin, 'Placing search in context: the concept revisited', in *Proceedings of the Tenth International World Wide Web Conference*, pp. 116 – 131, (2001).

[4] Evgeniy Gabrilovich and Shaul Markovitch, 'Computing semantic relatedness using Wikipedia-based explicit semantic analysis', in *Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI 07)*, pp. 1606–1611, (2007).

[5] Evgeniy Gabrilovich and Shaul Markovitch, 'Wikipedia-based semantic interpretation for natural language processing', *Journal of Artificial Intelligence Research*, **34**(2), 443, (2009).

[6] Thomas Gottron, Maik Anderka, and Benno Stein, 'Insights into Explicit Semantic Analysis', in *CIKM'11: Proceedings of 20th ACM Conference on Information and Knowledge Management*, pp. 1961–1964, (2011).

[7] Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren, '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, pp. 1406–1414, New York, NY, USA, (2012). ACM.

[8] Stephen P. Harter, 'A probabilistic approach to automatic keyword indexing. part i. on the distribution of specialty words in a technical literature', *Journal of the American Society for Information Science*, **26**(4), 197–206, (1975).

[9] Samer Hassan and Rada Mihalcea, 'Semantic relatedness using salient semantic analysis.', in *AAAI*, eds., Wolfram Burgard and Dan Roth. AAAI Press, (2011).

[10] David A. Huffman, 'A method for the construction of minimum redundancy codes', *Proceedings of the I.R.E.*, **40**(9), 1098–1101, (1952).

[11] Thad Hughes and Daniel Ramage, 'Lexical semantic relatedness with random graph walks.', in *EMNLP-CoNLL*, pp. 581–589. ACL, (2007).

[12] Petr Knoth, Lukas Zilka, and Zdenek Zdrahal, 'KMI, The Open University at NTCIR-9 CrossLink: Cross-Lingual Link Discovery in Wikipedia Using Explicit Semantic Analysis ', in *Proceedings of NTCIR-9 Workshop Meeting*, Tokyo, Japan, (2011).

[13] David D. Lewis, Y. Yang, T. Rose, and F. Li, 'Rcv1: A new benchmark collection for text categorization research', *Journal of Machine Learning Research*, **5**, 361–397, (2004).

[14] Michael McCandless, Erik Hatcher, and Otis Gospodnetic, *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*, Manning Publications Co., Greenwich, CT, USA, 2010.

[15] George A. Miller, 'WordNet: a lexical database for English', *Commun. ACM*, **38**(11), 39–41, (November 1995).

[16] George A. Miller and Walter G. Charles, 'Contextual correlates of semantic similarity', *Language and Cognitive Processes*, **6**(1), 1–28, (1991).

[17] Zsolt Minier, Zalan Bodo, and Lehel Csato, 'Wikipedia-based kernels for text categorization', in *Symbolic and Numeric Algorithms for Scientific Computing, 2007. SYNASC. International Symposium on*, pp. 157–164. IEEE, (2007).

[18] Vivi Nastase and Michael Strube, 'Transforming Wikipedia into a large scale multilingual concept network', *Artif. Intell.*, **194**, 62–85, (January 2013).

[19] Alexander Panchenko and Olga Morozova, 'A study of hybrid similarity measures for semantic relation extraction', in *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, HYBRID '12, p. 10–18, Stroudsburg, PA, USA, (2012). Association for Computational Linguistics.

[20] Anil Patelia, Sutanu Chakraborti, and Nirmalie Wiratunga, 'Selective integration of background knowledge in tcbr systems', in *Case-Based Reasoning Research and Development*, eds., Ashwin Ram and Nirmalie Wiratunga, volume 6880 of *Lecture Notes in Computer Science*, pp. 196–210. Springer Berlin / Heidelberg, (2011). 10.1007/978-3-642-23291-6_16.

[21] Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli, 'Align, disambiguate and walk: A unified approach for measuring semantic similarity.', in *ACL (1)*, pp. 1341–1351. The Association for Computer Linguistics, (2013).

[22] Tamara Polajnar, Nitish Aggarwal, Kartik Asooja, and Paul Buitelaar, 'Improving ESA with document similarity', in *Advances in Information Retrieval*, 582–593, Springer, (2013).

[23] Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch, '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, pp. 337–346, New York, NY, USA, (2011). ACM.

[24] Stephen Robertson and Hugo Zaragoza, 'The probabilistic relevance framework: Bm25 and beyond', *Foundations and Trends in Information Retrieval*, **3**(4), 333—389, (2009).

[25] Herbert Rubenstein and John B. Goodenough, 'Contextual correlates of synonymy', *Commun. ACM*, **8**(10), 627–633, (October 1965).

[26] Philipp Scholl, Doreen Böhnstedt, Renato Domínguez García, Christoph Rensing, and Ralf Steinmetz, 'Extended explicit semantic analysis for calculating semantic relatedness of web resources', in *Proceedings of the 5th European conference on Technology enhanced learning conference on Sustaining TEL: from innovation to learning and practice*, EC-TEL'10, pp. 324–339, Berlin, Heidelberg, (2010). Springer-Verlag.

[27] ClaudeE. Shannon, 'A mathematical theory of communication', *Bell System Technical Journal*, **27**, 379–423 and 623–656, (July and October 1948).

[28] Philipp Sorg and Philipp Cimiano, 'Cross-lingual information retrieval with explicit semantic analysis', in *Working Notes for the CLEF 2008 Workshop*, (2008).

[29] Michael Strube and Simone Paolo Ponzetto, 'WikiRelate! computing semantic relatedness using Wikipedia', in *proceedings of the 21st national conference on Artificial intelligence - Volume 2*, AAAI'06, p. 1419–1424. AAAI Press, (2006).

[30] Ian H. Witten and David Milne, '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*, pp. 25–30, (2008).

[31] Wen-tau Yih and Vahed Qazvinian, 'Measuring word relatedness using heterogeneous vector space models', in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pp. 616–620, Stroudsburg, PA, USA, (2012). Association for Computational Linguistics.