German Encyclopedia Alignment Based on Information Retrieval Techniques

Roman Kern[1] and Michael Granitzer[1,2]

[1] Know-Center, Graz
[2] Graz University of Technology
{rkern,mgrani}@know-center.at

Abstract. Collaboratively created online encyclopedias have become increasingly popular. Especially in terms of completeness they have begun to surpass their printed counterparts. Two German publishers of traditional encyclopedias have reacted to this challenge and decided to merge their corpora to create a single more complete encyclopedia. The crucial step in this merge process is the alignment of articles. We have developed a system to identify corresponding entries from different encyclopedic corpora. The base of our system is the alignment algorithm which incorporates various techniques developed in the field of information retrieval. We have evaluated the system on four real-world encyclopedias with a ground truth provided by domain experts. A combination of weighting and ranking techniques has been found to deliver a satisfying performance.

1 Introduction

Printed encyclopedias have been the prime source of information for a long time. They are created by experts in their fields and therefore provide a high credibility. Due to their tradition as printed media, encyclopedias follow a particular structure and outline. Space is at prime and therefore articles tend to be terse. Still the articles should contain all available information resulting in a writing style specific to such corpora. Dealing with this kind of language poses an additional challenge for natural language processing (NLP), machine learning (ML) or information retrieval (IR) techniques.

The rise of the Internet and more specifically the popularity of online encyclopedias has put pressure on the producers of traditional printed encyclopedias. While initially there has been doubts whether the new form of collaboratively created resources can match the quality of the established encyclopedias (see for example [1]), more recently traditional publishers have changed their strategy. They have started to put their resources online and also started in parts to allow non-experts to contribute information.

Another way to improve the quality and especially the completeness of an encyclopedic resource is the combination of multiple sources. Starting with two encyclopedias one can create a merged resource that contains the combined and as a consequence a more complete information. The most important step of this

operation is the alignment of articles. Articles about the same person, entity or concept in both encyclopedias should be automatically assigned to each other. Additionally, those articles that only exist in one of the two encyclopedias should be identified and thus create a new entry in the merged corpus.

State-of-the-art methods in NLP and related techniques has not yet reached the level that such an alignment can be conducted completely automatically. Manual intervention of human experts is still necessary in many cases. Prior to developing an encyclopedia alignment system we have set-up a number of goals to achieve:

- The accuracy of the automatic article alignment should be maximized.
- The coverage of automatically aligned articles should be as high as possible, to minimize the number of articles required for manual assignment.
- "Keep the human in the loop" and support the manual alignment by providing an intuitive search infrastructure and useful recommendations.

Our system should also be used in an interactive manner to support manual alignment by domain experts. Therefore the alignment algorithm not only provides a high accuracy, it should also be fast and efficient, as finally that the algorithm should be integrated into a software tool targeted at desktop computers (Figure 1 depicts a prototype of the application). We have decided to choose techniques from the field of information retrieval as the base of our alignment algorithm for a couple of reasons. Search and indexing tools have been developed for a long time and have now reached a mature level. Retrieval algorithms are well studied and their behavior is well understood. In contrast to these methods, the results produced by many supervised classification methods are hardly traceable.

2 Related Work

The most striking characteristic of many articles within traditional encyclopedias is their length. Because of space limitations the majority of all articles are relatively short compared to the covered information.

In [2] an overview of similarity methods for various short contexts is given. Using the categorization presented by the authors, a single encyclopedia article can be classified as head-less context and the alignment can be seen as pairwise comparison to reference samples. To calculate the similarity betweeen two short contexts according to the paper the words can either be directly used or replaced by an representation. The first method is referred as first-order similarity, whereas the second method is called second-order similarity. For the second-order representation the individual word within the context are usually expanded by exploiting an external resource, for example WordNet.

One of the approaches to integrate semantic information via WordNet is presented in [3]. They propose an algorithm to calculate the similarity between individual sentences. The distance between entries within the WordNet graph are taken as proxy for the semantic relatedness of words. Additionally, their

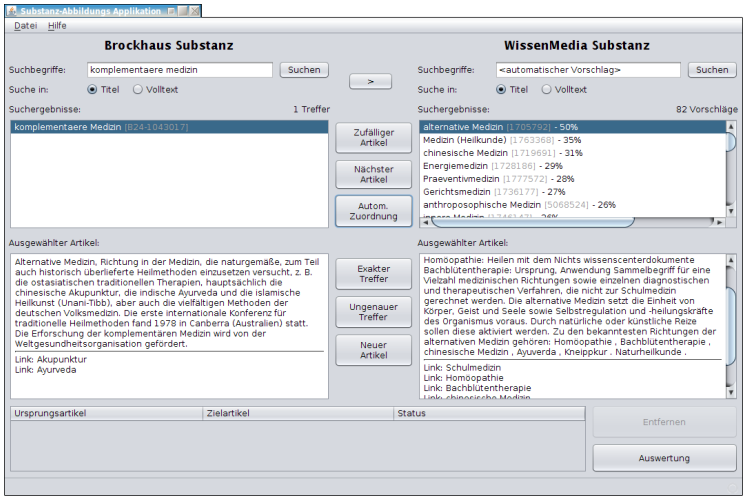

Fig. 1: Prototype of the application that should support domain experts in the process of encyclopedia alignment. The quality and the number of automatically aligned articles has a high impact on the productivity of the experts.

algorithm deviates from the unordered bag of words approach by incorporating the word order into their similarity calculation.

A similar approach is taken by [4] where position information and lexical distance serve as base for the similarity. The performance of this algorithm is compared against a system which employs Latent Semantic Analysis in [5]. In this comparative study they created a benchmark dataset of 30 sentence pairs. At first humans assigned a similarity for each of the sentence pairs which served as ground truth. Finally, they compared the mean similarity from the human judgments with the results of the two approaches. They found that the LSA based algorithm produces a higher correlation than the similarities calculated as described in [4].

Various degrees of similarities are studied in [6], from a broad topical similarity at one end of the spectrum to document identity at the other end. They present various measures to calculate the similarity of sentences and documents, for example the overlap of common words and a *TFIDF* based weighting of shared words. Probabilistic translation models are also investigated in their study together with the DECO system (see [7]) for document similarity. In their evaluation they report the performance of the different similarity measures for various degrees of similarity. For encyclopedia alignment the results at the "same facts" category are the most relevant. For this category the machine translation model and the simple overlap measure provides the best performance.

Beside incorporating resources like WordNet and other thesauri into the similarity calculation, the web has become increasingly popular as knowledge base in recent works. In [8] the authors incorporate the results of web searches into a

similarity kernel function. Their method is targeted at finding similar short text snippets, especially substitution candidates for search queries. This approach is further improved in [9] by changing the weighting function and by integrating machine learning algorithms. Out of the surface matching similarity measures the Jaccard Coefficient fared better than the Overlap and the Cosine similarity in their evaluation.

3 Encyclopedic Corpora

We have had the opportunity to have four encyclopedic corpora at our disposal when developing and evaluating our encyclopedia alignment system. Three of them are from *Brockhaus* and vary in the number of articles and average length of the articles. The fourth corpus, the *WissenMedia* encyclopedia, is comparable in number of articles with the largest of the Brockhaus corpora. Table 1 gives an overview of the statistics of the four datasets.

The task of our alignment system is to merge all articles from the three Brockhaus corpora and the WissenMedia corpus to create one single and complete encyclopedia. For example, the article on the right side of figure 2 should be assigned to the article depicted on the left.

3.1 Anatomy of an Encyclopedia Article

Each article in an encyclopedia consists of multiple parts, whereas the title and the textual content are the most important ones. The main content does not only contain the plain text of the article but also links to other articles and may also feature references to pictures and other media. Depending on the actual encyclopedia it may contain additional annotations not visible in the plain text, for example the number of inhabitants in an article that describes a specific city or country. Other data can be extracted directly from the text, for example the date of birth of a person.

Additionally to the title each article may also contain a sub-title. For articles that represents a person the sub-title contains the person's first name. The sub-title is sometimes also used for disambiguational purpose, for example the articles with the title "Mexico" also carry the sub-title "city" or "country".

Table 1: Overview of the statistics of the three *Brockhaus* and the *WissenMedia* encyclopedias. The average length of an article is less then 100 words for each corpus.

	Brockhaus 1	Brockhaus 2	Brockhaus 3	WissenMedia
Number of articles	42,450	94,854	176,963	176,011
Number of unique words	137,929	395,685	840,577	370,076
Number of words	979,958	5,256,761	16,920,079	6,761,156
Average article length	23.16	55.47	95.62	38.42

Fig. 2: The same article in two different encyclopedias. Although the title of the two articles differ, they cover the same topic and should therefore be assigned to each other.

Unfortunately this disambiguation information is not standardized and is used differently in each encyclopedia.

Finally, the article may also carry a wide array of additional meta-data, which is not exploited by our system, for example the pronunciation of the article's title, assignments to classification taxonomies and hints how the article should look like in printed form.

4 Algorithm

The alignment algorithm operates in two stages: a retrieval and a ranking stage. In the first stage for a particular source article a list of candidate target articles is generated. Each of the candidate articles are individually weighted in the second stage. The output of the final stage is a ranked list of possible target articles, where each article's weight ranges between 0 and 1. The highest ranked article is marked as the alignment match for the source article if the weight exceeds a predefined threshold. By choosing a low threshold the number of automatically aligned articles will rise. A high threshold will lead to fewer aligned articles but the number of misalignments will also decline. In the evaluation section we study the influence of this parameter on the systems performance.

4.1 Text Processing

In contrast to the English language, in German noun word-compounds are frequently used. For example the English phrase "coffee maker" can be translated as the single German word "Kaffeemaschine". In encyclopedias these compound words are even more common than in general German due to the terse nature of articles.

In our system we have implemented two different strategies to deal with these compound words. The first is a simple character n-gram approach that splits words into n-grams of up to 3 consecutive characters. For example the 3-grams of "Kaffeemaschine" are: `kaf aff ffe fee eem ema mas asc sch chi hin ine`

The second approach is more sophisticated. Each tokenized word is first split into syllables based on hyphenation patterns. Each syllable is looked up in a dictionary to detect whether the syllable can be used as a single standalone word. After the syllable has passed this check it is finally stemmed[3]. The hyphenation patterns and the dictionary are available from the OpenOffice.org project[4]. The output of this processing for the word "Kaffeemaschine" is: `kaffee fee kaffeemaschi ma maschi schi`

4.2 Article Facets

The basic data-structure of our alignment algorithm is a search index, which is populated by all articles of the encyclopedias. To capture the different aspects of an article we split the article into different facets:

Title-Exact The article title is tokenized and normalized. All characters were transformed to lower case, umlauts were replaced with their corresponding digraphs and diacritics were removed. For example the word *Übersee-département* is normalized to: `ueberseedepartement`

Title The tokenized, normalized title is further processed using one of the two compound words processing algorithms.

Sub-Title The sub-title (if available) is tokenized and processed like the title.

Content The body of the article is again split into normalized tokens which were consecutively processed by one of the word-compound processing approaches.

Date This facet is filled by extracting the birth and death dates out of the content by applying a pattern based approach. This faced is populated only for articles about persons. For example the article about *Johann Wolfgang von Goethe* contains the dates: `*1749 †1832`

Length This facet is in contrast to the other facets not filled with textual content. It captures the intuition that articles about important concepts tend to contain more words than minor topics. For example the article about famous persons will tend to be longer than articles of people who have not gained huge popularity. Two corresponding articles from two encyclopedias are thus expected to have similar length in relation to the average length of articles within the encyclopedia. The content of the *Length* facet is calculated as defined in equation 1. Important topics have a length ratio close to 1, the ratio for short articles is close to zero and a ratio of 0.5 reflects an article of average length.

$$lengthRatio = min\left(\frac{length}{2 * averageDocLength}, 1\right) \quad (1)$$

[3] Stemmer and token splitting algorithms are taken from the open-source Lucene project: `http://lucene.apache.org/java/docs/`

[4] `http://extensions.services.openoffice.org/dictionary`

4.3 Candidate Selection & Candidate Weighting and Ranking

Once the search index is created the matching target articles for a source article can be searched. The first step is the selection of a list of possible candidates. Out of the features of the source article a query is build and the top 100 results are selected for further investigation. This query is a disjunction of the facets *Title-Exact*, *Title*, *Sub-Title* and *Content*. In case of the *Content* facet only the 10 tokens with the highest weight are taken, using the weighting scheme described in the following section.

In the article weighting step, each candidate is compared with the source article and a similarity score is calculated. The similarity score for each target article is computed by combining the individual similarities of the facets. For each facet - f - out of the set of facets - F - a similarity score is computed for a pair of source and target articles. Not all facets should contribute equally to the final score, thus a predefined boost constant for each facet B_f is incorporated into a weighted mean for the final score:

$$S(s,t) = \frac{1}{B_{sum}} \sum_{f \in F} B_f * boost(score(f,s,t)) * score(f,s,t) \tag{2}$$

The B_{sum} is the sum of all boost constants and serves as a normalization factor for the score to fall between 0 and 1. The $boost()$ function is based on the intuition that similarity scores near the extremes are better suited to assess a similarity or dissimilarity. In the evaluation section a number of boost function are compared against a baseline that just returns a constant value for each similarity score. The actual values for the boost constants B_f were determined on a preliminary test of 100 randomly drawn articles: $B_{TitleExact} = 20, B_{Title} = 25, B_{SubTitle} = 40, B_{Content} = 75, B_{Date} = 50, B_{Length} = 2$.

The most important part of equation 2 is the $score()$ function that calculates the similarity of corresponding facets of two articles. Each facet is transformed into a weighted vector so that different similarity measures can be used, namely: Cosine, City-Block, Euclidean, Jaccard, Dice and Overlap. Distance measures were transformed to similarities via: $sim = 1/(1 - distance)$

To create the weighted term vector for each facet we have integrated a number of weighting functions. The first is a simple *TFIDF* weighting scheme based on the number of articles - N - within the encyclopedia and the number of articles the term t occurs in - $docFreq_t$:

$$weight_{TFIDF}(t) = log(\frac{N}{docFreq_t + 1} + 1) * \sqrt{termFreq_t} \tag{3}$$

The next term weighting function has been developed using an axiomatic approach to information retrieval, see [10]. This weighting scheme also incorporates the actual length of the article (in this case the number of terms within a facet) and the average length of articles. For the parameter α we used the value 0.32 as suggested by the authors of [10].

$$weight_{Axiomatic}(t) = (\frac{N}{docFreq_t})^{\alpha} \frac{termFreq_t}{termFreq_t + 0.5 + \frac{docLength}{avgDocLength}} \tag{4}$$

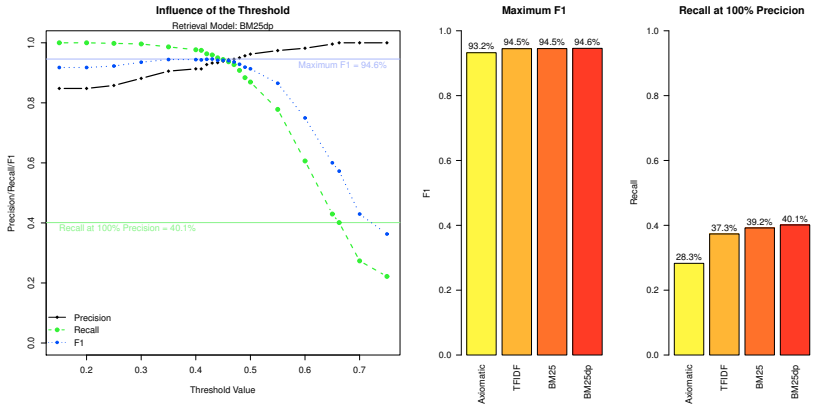

Fig. 3: Precision/Recall/F1 curve for various threshold values for a single retrieval model (BM25dp) on the left side. On the right side: Comparison of all evaluated retrieval models using the two main quality indicators. The modified BM25 retrieval model achieves the highest overall performance.

The BM25 retrieval function, see [11], has proven to provide state-of-the-art performance in a number of scenarios. We used the recommended default values for the parameters: $k_1 = 2$, $b = 0.75$

$$weight_{BM25}(t) = \frac{termFreq_t}{termFreq_t + k_1((1-b) + b * \frac{docLength}{averageDocLength})} * log \frac{N - docFreq_t + 0.5}{docFreq_t + 0.5} \quad (5)$$

For the final term weighting function we modified the $BM25$ weighting scheme to incorporate the degree of dispersion of terms. The DP measure has been proposed by [12] and successfully used by [13] to separate function words from content words. The dispersion degree is low for words with an even frequency distribution, which is expected for words with little semantics but with a grammatical function. The parameter α has been set to -0.3 based on the results of the preliminary tests.

$$weight_{BM25dp}(t) = weight_{BM25}(t) * DP_t^{\alpha} \quad (6)$$

5 Evaluation & Discussion

The evaluation of our encyclopedia alignment system is based on a ground truth generated by domain experts, which were asked to pick representative articles from their respective domains. The three Brockhaus corpora serve as source and the WissenMedia corpus as target of the alignment. A total of 605 articles were manually processed. For 64 Brockhaus articles the experts have not found a corresponing article in the WissenMedia encyclopedia.

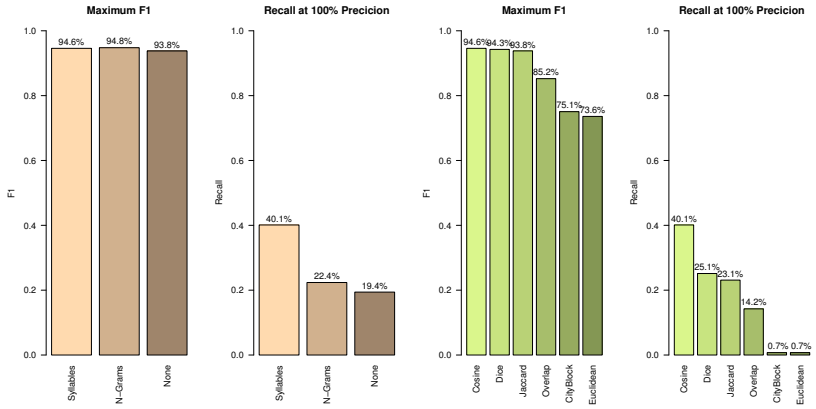

Fig. 4: Comparison of the word-compound processing strategies on the left side, and the performance of the various similarity measures on the right side. The compound splitting method based on hyphenation pattern outperforms the n-gram based method, which is still better than no splitting at all. Out of the similarity measures, the cosine similarity provides by far the best results.

With this ground truth the precision and recall of each configuration of the system can be calculated. The precision is calculated as the number of correctly assigned articles in relation to the total number of automatically assigned articles. Recall defines the ratio of correct assignments to the number of possible assignments (the number of manually assigned articles). The harmonic mean of precision and recall, called F1, is the base of first main indicator for the quality of the results of the algorithm. Running the evaluation with different thresholds generates a series of F1 measures, see left chart in figure 3. The highest F1 measures defines the best achievable performance when both precision and recall should be equally optimized.

Another characteristic of an evaluation run is the number of aligned articles without a single misalignment. The recall value at the point where the precision reaches 100% captures this property. This measure reflects the usefulness of the configuration if the emphasis lies on optimizing the precision. The higher the recall the less articles have to be manually postprocessed and therefore this indicator plays an important role when choosing a configuration.

The first components of our system to be compared are the different retrieval models, see figure 3. While all four methods appear comparable when using only the F1 based measure as quality criteria, the recall measure reveals that the axiomatic approach falls behind the other retrieval models in terms of performance. The modified BM25 weighting function, which incorporates the dispersion of terms, appears to provide the best results and for this reason this configuration is taken as baseline for all other evaluations.

The next evaluation compares the consequences of the two different word-compound processing methods on the systems performance together with a con-

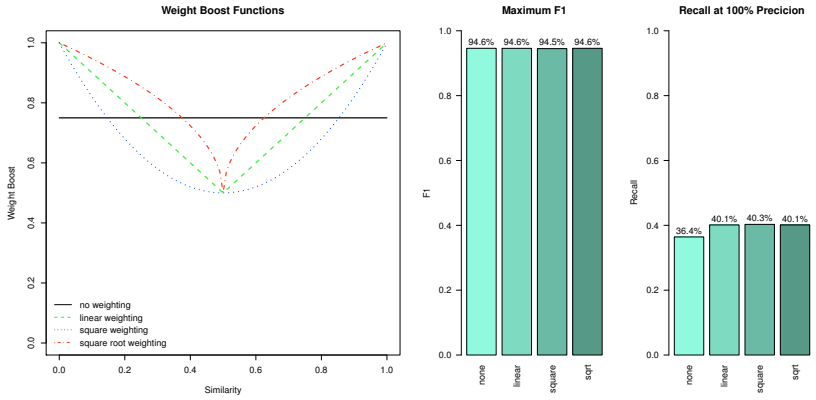

Fig. 5: Weighting functions that capture the intuition that similarities close to the extremes should have a higher impact on the final assignment. The shapes of the boosting function are depicted on the left and the performance indicators on the right. The improved recall at 100% precision indicates that this intuition is indeed sound.

figuration without any compound-word splitting, see figure 4. While the F1 measure for the n-gram based method is slightly higher, the syllable based approach achieves a higher recall value and thus is better suited for our use case. Still the n-gram based methods is able to perform better than using no splitting at all. This corroborates the need to process compound-words in the German language not only for encyclopedia alignment, but also in other areas, like for example information retrieval.

The results of the evaluation of the different similarity measures are simple to interpret. The cosine measure outperforms all other similarity measures considerably. One reason is the fact that some facets are very sparse, for example the *Title* facet. Applying different similarity measure on different facets could be one possible way to further improve the quality of the alignment algorithm.

Next we tried to assess whether the intuition that similarity measures near the extreme ends are better suited as indicator for similarity. This should especially help in situations where there is an exact match for one of the facets. Figure 5 depicts the baseline (the similarity value has no influence on the boost) and three weight boosting methods. Although the difference between boosting methods appears to be negligible, the boosting approach itself improves the recall by about 4%.

Finally, we investigated the relative influence of each facet. To measure the individual contribution of facets we have repeated the evaluation while removing one facet at a time. As expected the content of the article is by far the most important factor. Still all other facets contribute to the quality of the result, whereas the two facets generated from the title appear to be slightly redundant. The date and the length information have little influence on the maximum F1 measure, removing them has a pronounced negative effect on the recall at 100%

Table 2: Performance indicators for each facet when left out of the similarity calculation. Each facet appears to contribute to the quality of the alignment with the *Content* facet being the most important one.

Facet	Maximum F1	Recall at 100% Precision
Title-Exact	93.4%	40.5%
Title	92.5%	40.9%
Sub-Title	94.3%	39.0%
Content	86.4%	17.7%
Date	94.6%	35.7%
Length	94.4%	39.6%

precision measure. Only the combination of all facets provides the best overall performance of our encyclopedia alignment system.

6 Conclusion & Future Work

The automatic alignment of encyclopedic corpora poses a number of challenges. The style of the German language differs from the common language usage because of the terse nature of the articles. Additionally the alignment process should be fast and efficient to be used in an interactive manner. Furthermore, the results produced by the system should be predictable and easy to interpret.

We created such an encyclopedia alignment system by applying techniques developed in the field of information retrieval. Domain experts manually aligned over 600 articles of four real-world encyclopedias. Using this ground truth we evaluated a number of configurations with different retrieval functions, similarity measures and text processing methods. The combination of a modified BM25 weighting function, the cosine similarity and a dictionary based word splitting algorithm provided the best overall performance. This configuration achieved an maximum F1 measure of over 94% and over 40% recall without a single misalignment.

To further improve the quality we could exploit the internal link structure and integrate external resources, for instance thesauri. Incorporating machine translation techniques and language models are among the possible candidates for future improvements.

Although our system has been developed to align articles from different encyclopedias, it should be easy to adapt for other purposes. The detection of duplicates is probably the most obvious application. Other areas are the named entity recognition and disambiguation, which could be integrated into a link recommendation system. Some aspects of our alignment system should not only apply to encyclopedias, but to other textual resources as well. The word-compound splitting method and the dispersion based term weighting should be helpful in other text processing applications as well.

Putting technical aspects aside we believe that our alignment system also serves as good example how science and industry can work together to create solutions and insights beneficial for both sides.

Acknowledgements

We would like to thank Kai-Ingo Neumann and his team at wissenmedia for their support in providing the datasets. The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

1. Rector, L.H.: Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. Reference Services Review **36**(1) (2008)
2. Pedersen, T.: Computational Approaches to Measuring the Similarity of Short Contexts: A Review of Applications and Methods. CoRR **abs/0806.3** (2008)
3. Liu, X., Zhou, Y., Zheng, R.: Measuring semantic similarity within sentences. In: Proceedings of the 7th International Conference on Machine Learning and Cybernetics, ICMLC. Volume 5. (2008) 2558–2562
4. Li, Y., McLean, D., Bandar, Z.: Sentence similarity based on semantic nets and corpus statistics. IEEE Transactions on Knowledge and Data Engineering **18**(8) (2006) 1138–1150
5. O Shea, J., Bandar, Z., Crockett, K., McLean, D.: A Comparative Study of Two Short Text Semantic Similarity Measures. In: Agent and Multi-Agent Systems: Technologies and Applications: Second KES International Symposium. Volume 4953., Springer (2008) 172–181
6. Metzler, D., Bernstein, Y., Croft, W., Moffat, A., Zobel, J.: Similarity Measures for Tracking Information Flow. In: CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management, ACM (2005) 517–524
7. Bernstein, Y., Zobel, J.: A Scalable System for Identifying Co-derivative Documents. In: String Processing and Information Retrieval. (2004) 55–67
8. Sahami, M., Heilman, T.: A web-based kernel function for measuring the similarity of short text snippets. In: WWW '06: Proceedings of the 15th international conference on World Wide Web, ACM (2006) 377–386
9. Yih, W., Meek, C.: Improving similarity measures for short segments of text. In: AAAI'07: Proceedings of the 22nd national conference on Artificial intelligence, AAAI Press (2007) 1489–1494
10. Fang, H., Zhai, C.: An exploration of axiomatic approaches to information retrieval. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, ACM (2005) 480–487
11. Robertson, S., Gatford, M.: Okapi at TREC-4. In: Proceedings of the Fourth Text Retrieval Conference. (1996) 73–97
12. Gries, S.: Dispersions and adjusted frequencies in corpora. International Journal of Corpus Linguistics **13**(4) (2008) 403–437
13. Kern, R., Granitzer, M.: Efficient linear text segmentation based on information retrieval techniques. In: MEDES '09: Proceedings of the International Conference on Management of Emergent Digital EcoSystems, ACM (2009) 167–171