Int J Digit Libr
DOI 10.1007/s00799-011-0069-5

A hybrid system for German encyclopedia alignment

Roman Kern · Christin Seifert · Michael Granitzer

Abstract Collaboratively created on-line 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 started an initiative to merge their corpora to create a single, more complete encyclopedia. The crucial step in this merging process is the alignment of articles. We have developed a two-step hybrid system to provide high-accurate alignments with low manual effort. First, we apply an information retrieval based, automatic alignment algorithm. Second, the articles with a low confidence score are revised using a manual alignment scheme carefully designed for quality assurance. Our evaluation shows that a combination of weighting and ranking techniques utilizing different facets of the encyclopedia articles allow to effectively reduce the number of necessary manual alignments. Further, the setup of the manual alignment turned out to be robust against inter-indexer inconsistencies. As a result, the developed system empowered us to align four encyclopedias with high accuracy and low effort.

Keywords Encyclopedia alignment · Semantic similarity · Hybrid alignment system

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) and information retrieval (IR) techniques.

The rise of the Internet and more specifically the popularity of on-line encyclopedias has put pressure on the producers of traditional printed encyclopedias. While initially there have been doubts whether the new form of collaboratively created resources can match the quality of the established encyclopedias (see for example [14]), more recently traditional publishers have changed their strategy. They have started to put their resources on-line and also started initiatives 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 encyclopedia containing more complete information. The most important step of this operation is the alignment of articles, i.e., the identification of corresponding

This article is a substantially revised and extended version of a article with the title "German Encyclopedia Alignment Based on Information Retrieval Techniques" originally appeared in the Proceedings of the 14th European Conference on Digital Libraries (ECDL 2010).

R. Kern (✉) · C. Seifert · M. Granitzer
Graz University of Technology, Knowledge Management Institute, Inffeldgasse 21a, 8010 Graz, Austria
e-mail: rkern@know-center.at

C. Seifert
e-mail: christin.seifert@tu-graz.at

M. Granitzer
Know-Center GmbH and Graz University of Technology, Knowledge Management Institute, Inffeldgasse 21a, 8010 Graz, Austria
e-mail: mgrani@know-center.at

Published online: 09 September 2011

and non-corresponding articles. Articles about the same person, entity or concept in both encyclopedias should be automatically assigned to each other. In addition, articles that only exist in one of the two encyclopedias should be identified and thus create a new entry in the merged one.

State-of-the-art methods in NLP and related techniques have not yet reached the level that such an alignment can be conducted completely automatically. Hence, manual intervention of human experts is still necessary in many cases. Combining automatic with manual alignment methods therefore raises a number of requirements to be considered:

- 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.
- Ensure high quality of human alignments and provide quality assurance means.

Given these requirements, we developed a system that consists of two parts, an automatic alignment component and a manual alignment procedure.

The algorithm for automatic alignment relies on recent developments in IR and provides fast and efficient alignments with high accuracy. During the development of the automatic alignment algorithm, domain experts assessed the algorithms alignment quality and provided the corresponding ground-truth for evaluating various algorithmic properties. Figure 1 shows a screen shot of the application for the experts. We evaluated properties like ranking-schemes, compound-word processing strategies and boosting of different facets to optimize the automatic alignment, which forms the foundation for the second part.

The second part of our system optimized human alignment support of all non-automatically aligned articles, i.e., articles which could not be reliably processed by the automatic algorithm. We developed an assignment scheme to distribute subsets of these articles to a group of alignment workers, referred to as users from now on. Via overlapping subsets and an equidistant distribution of the expert ground-truth into the manual alignment scheme, we have been able to detect mistakes among the alignment workers. From these mistakes we could identify several groups of common errors and estimate the inter-indexer inconsistency. The fact that even humans regularly produce wrong alignments acts as an indicator that the correct alignment of encyclopedic articles remains a challenging task, which can be eased with the help of our approach.

2 Related work

The most striking characteristic of many articles within traditional encyclopedias is their length. Due to space limitations the majority of all articles are relatively short compared to the covered information. Therefore, the crucial component of an encyclopedia alignment system is a reliable similarity measure for short texts.

In [13], an overview of methods to calculate similarities 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 pair-wise comparison to reference samples. To calculate the similarity between two short contexts according to the article the words can either be directly used or replaced by a 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 words within the context are usually expanded by exploiting an external resource, such as WordNet.

One of the approaches to integrate semantic information via WordNet is presented in [9]. They propose an algorithm to calculate the similarity between individual sentences. The distances between entries within the WordNet graph are taken as proxy for the semantic relatedness of words. In addition, the algorithm deviates from the unordered bag of words approach by incorporating the word order into their similarity calculation.

A similar approach is taken by [8] 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 [12]. 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. The authors found that the LSA-based algorithm produces a higher correlation than the similarities calculated as described in [8].

Various degrees of similarities are studied in [11], from a broad topical similarity at one end of the spectrum to document identity at the other end. Various measures to calculate the similarity of sentences and documents are presented, 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 [2]) for document similarity. In the evaluation, the performance of the different similarity measures for various degrees of similarity is reported. For encyclopedia alignment, the results for the "same facts" category are the most relevant, as articles from two different encyclopedias which cover the same topic are expected to be differently written while covering the same information. For this category, the

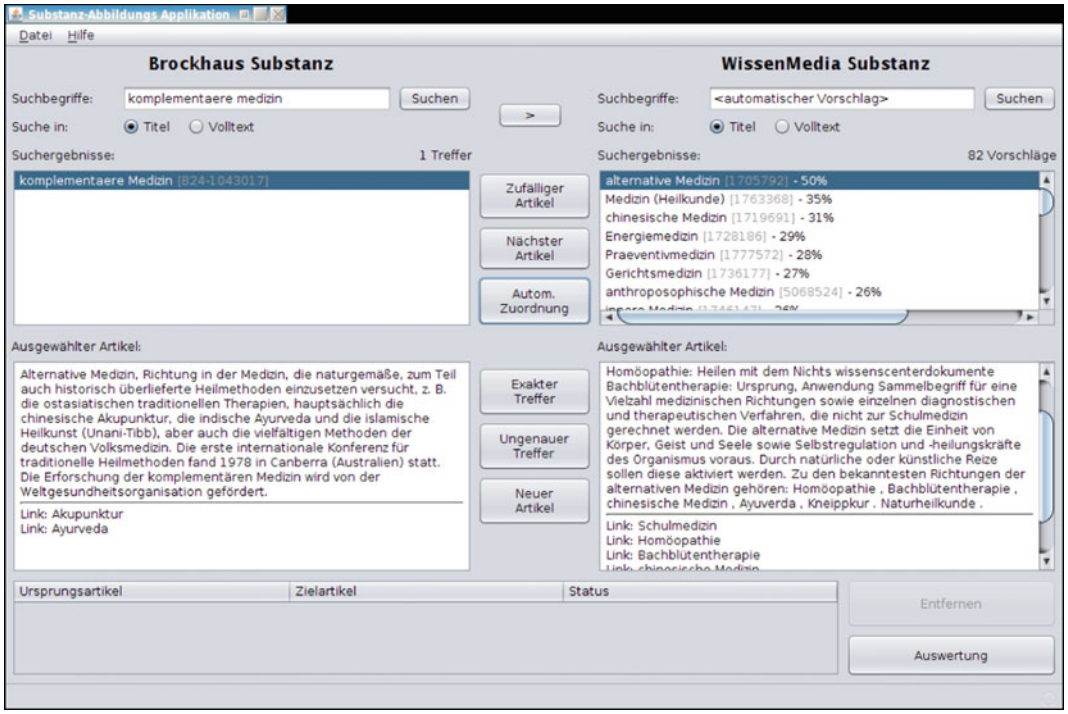

Fig. 1 Screenshot of an application developed to support domain experts in generating feedback to improve the quality of the automatic alignment algorithm. In addition, this tool has been used to assemble a test dataset that was used for the evaluation

machine translation model and the simple overlap measure provide the best performance.

Besides incorporating resources like WordNet and other thesauri into the similarity calculation, the web has become increasingly popular as knowledge base in recent works. In [16], 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 [17] by changing the weighting function and by integrating ML algorithms. Out of the surface matching, similarity measures the Jaccard Coefficient fared better than the Overlap and the Cosine similarity in their evaluation.

Another known approach to compute the semantic relatedness of texts is the Explicit Semantic Analysis (ESA) [1,5]. Here, texts are mapped to a reference corpus using standard text metrics like Cosine-similarity, and compared afterwards in the mapped representation. The ESA yields significant improvements compared to the vector space model and LSA especially on short text documents, but requires a corresponding reference corpus.

Marko et al. [10] present an alignment system for cross-lingual medical corpora. In contrast to our work, their corpora are domain-specific whereas our encyclopedias contain world knowledge. The system presented by Bouma et al. [3] aligns templates in the Wikipedia encyclopedia. While their content is also world knowledge, they align templates across different languages while in this article the encyclopedias are of a single language.

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.[1]

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 Fig. 2 should be assigned to the article depicted on the left.

It is expected that there is a certain degree of overlap between each of the encyclopedias. For example, each of the corporas contains an article for the city “New York”. In addition, each of the four encyclopedias contains unique articles

[1] Due to changes in the way how redirects and disambiguation pages are handled the number of articles varies slightly from previously published numbers.

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 I	Brockhaus II	Brockhaus III	WissenMedia
Number of articles	42,008	92,260	173,671	180,828
Number of unique words	129,478	362,773	788,593	370,076
Number of words	924,248	4,997,885	16,450,799	6,761,156
Average article length	22.00	54.18	94.73	38.42

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

not found in any of the other corporas. For example, only a single corpus contains an article dedicated to the "New York State Barge Canal".

3.1 Anatomy of an encyclopedia article

Each article in an encyclopedia consists of multiple parts, the title and the textual content being 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 additional annotations not visible in the plain text are available, 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 represent a person the sub-title contains the person's first name. The sub-title is sometimes also used for the purpose of disambiguation, for example, the articles with the title "Mexico" also carry the sub-title "city" or "country". 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 Automatic alignment

The task of the automatic alignment is to assign all articles from a source encyclopedia to a matching article in the target encyclopedia.

4.1 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 system's performance.

4.1.1 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.[2] The hyphenation patterns and the dictionary are available from the OpenOffice.org project.[3] The output of this processing for the word "Kaffeemaschine" is:

```
kaffee fee kaffeemaschi ma maschi schi
```

4.1.2 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 *Überseedépartement* 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 about 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 Eq. 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.

$$\text{Length ratio} = \min\left(\frac{\text{length}}{2 \cdot \text{averageDocLength}}, 1\right) \tag{1}$$

4.1.3 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 built 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. The function score(f, s, t) denotes the result of the similarity score function given a facet f and two articles s and t. A similarity score of 1 represents the highest possible similarity, while 0 indicates that the two articles are completely dissimilar. 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. Furthermore, a dynamic boost weighting—$b(f, s, t)$—has been integrated based on the intuition that similarity scores near the extremes are better suited to assess a similarity or dissimilarity. The range of the dynamic boost function lies between 0 and 1. The final formula to calculate the similarity between two articles can be formalized as:

$$b(f, s, t) = \text{boost}(\text{score}(f, s, t)) \tag{2}$$

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

[3] http://extensions.services.openoffice.org/dictionary.

Table 2 The values for the boost constants for the article facets, where a higher number results in a bigger influence of the facet in the final score

Facet	B_f
Title-Exact	20
Title	25
Sub-Title	40
Content	75
Date	50
Length	2

$$\text{Score}(s,t) = \frac{1}{B_{\text{sum}}} \sum_{f \in F} B_f \cdot b(f,s,t) \cdot \text{score}(f,s,t) \tag{3}$$

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. In the evaluation section, a number of boost function are given and compared against a baseline that just returns a constant value for each similarity score. See Table 2 for the actual values for the boost constants B_f that were determined on a preliminary test-set of 100 randomly drawn articles.

The most important part of Eq. 3 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 for a given term t and article d, with N being the total number of articles in the encyclopedia, and docFreq_t the number of articles the term occurs in:

$$w^{\text{local}}_{\text{TFIDF}}(t,d) = \sqrt{\text{termFreq}_t} \tag{4}$$

$$w^{\text{global}}_{\text{TFIDF}}(t) = \log\left(\frac{N}{\text{docFreq}_t + 1} + 1\right) \tag{5}$$

$$\text{weight}_{\text{TFIDF}}(t,d) = w^{\text{local}}_{\text{TFIDF}}(t,d) \cdot w^{\text{global}}_{\text{TFIDF}}(t) \tag{6}$$

The next term weighting function has been developed using an axiomatic approach to IR, see [4]. 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 [4].

$$w^{\text{length}}_{\text{Axio}}(d) = \frac{\text{docLength}_d}{\text{avgDocLength}} \tag{7}$$

$$w^{\text{local}}_{\text{Axio}}(t,d) = \frac{\text{termFreq}_t}{\text{termFreq}_t + 0.5 + \text{w}^{\text{length}}_{\text{Axio}}(d)} \tag{8}$$

$$w^{\text{global}}_{\text{Axio}}(t) = \left(\frac{N}{\text{docFreq}_t}\right)^{\alpha} \tag{9}$$

$$\text{weight}_{\text{Axio}}(t,d) = w^{\text{local}}_{\text{Axio}}(t,d) \cdot w^{\text{global}}_{\text{Axio}}(t) \tag{10}$$

The BM25 retrieval function, see [15], has proven to provide state-of-the-art performance in a number of scenarios. We used the recommended default values for the parameters: $k_1 = 2, b = 0.75$

$$w^{\text{length}}_{\text{BM25}}(d) = k_1\Big((1-b) + b \cdot \frac{\text{docLength}_d}{\text{averageDocLength}}\Big) \tag{11}$$

$$w^{\text{local}}_{\text{BM25}}(t,d) = \frac{\text{termFreq}_t}{\text{termFreq}_t + \text{w}^{\text{length}}_{\text{BM25}}(d)} \tag{12}$$

$$w^{\text{global}}_{\text{BM25}}(t) = \log\frac{\text{N} - \text{docFreq}_t + 0.5}{\text{docFreq}_t + 0.5} \tag{13}$$

$$\text{weight}_{\text{BM25}}(t,d) = w^{\text{local}}_{\text{BM25}}(t,d) \cdot w^{\text{global}}_{\text{BM25}}(t) \tag{14}$$

For the final term weighting function, we modified the $BM25$ weighting scheme to incorporate the degree of dispersion (DP) of terms. The DP measure has been proposed by [6] and successfully used by [7] 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.

$$\text{weight}_{\text{BM25DP}}(t) = \text{weight}_{\text{BM25}}(t) \cdot DP_t^{\alpha} \tag{15}$$

4.2 Evaluation and discussion

The evaluation of our automatic alignment algorithm 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. Each alignment, therefore, maps an article from one of the three Brockhaus encyclopedias to a WissenMedia article or is marked as new article. A total of 605 articles were manually processed. For 64 Brockhaus articles, the experts have not found a corresponding article in the WissenMedia encyclopedia. The mappings contain a number of highly ambiguous concepts, for example, there are seven mappings for source articles with the title "Baum" (tree). The ground-truth does not contain any redirects or disambiguation articles.

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).

There are 147 mappings within the ground-truth where both the source and the target article carry the same title. For 38 of these articles the target encyclopedia contains multiple articles with identical titles. A system that would just compare the titles to find alignments would produce a precision of 36% and a recall of 27% when counting ties as misalignment.

The harmonic mean of precision and recall, called F1, is the base of the 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 Fig. 3. The highest F1 measures define the best achievable performance when both precision and recall should be equally optimized.

Another characteristic of a system configuration is the number of aligned articles without a single misalignment. This can be achieved by rising the threshold as long as there are wrongly aligned articles according to the ground-truth. At this point, the precision reaches 100% and the recall value is measured. 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 post-processed 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 Fig. 3. While all four methods appear comparable when using only the F1 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 system's performance together with a configuration without any compound-word splitting, see Fig. 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 are 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, IR.

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 measures on different facets could be one possible way to further improve the quality of the alignment algorithm.

Then, 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. The results of this analysis is given in Table 3. 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% precision measure. Only the combination of all facets provides the best overall performance of the automatic encyclopedia alignment algorithm.

5 Manual alignment

All articles that could not be confidently mapped or marked as new by the automatic alignment algorithm are further processed manually. The task of the human alignment workers is to find the corresponding target article for a given source article if it exits and marks the article as new otherwise. Obviously, humans would need support in finding candidate matches due to the sheer size of the encyclopedias. We developed an application supporting alignment workers in finding candidate matches and collecting their decisions. Figure 6 shows a screenshot of the application. The graphical user interface is divided into two parts, each part represents an encyclopedia. The left part corresponds to the source encyclopedia (Brockhaus I, II or III) and the right part corresponds to the WissenMedia encyclopedia. The articles of the source encyclopedia are listed alphabetically. The application is equipped with a retrieval interface to search for articles in the target encyclopedia. The search functionality applies the same technologies for retrieving relevant articles as described in the automatic alignment process (see Sect. 4.1.3). The search terms are defined by the user. The user can also chose whether a title search or a full-text search should be performed. The three buttons on the top right allow the user to indicate the relation of the selected source and target articles: "exakt" means that source and target articles are exact matches, "ungenau" means that source and target article are near matches and "nicht vorhanden" means that no match has been found in the target encyclopedia. An exact match

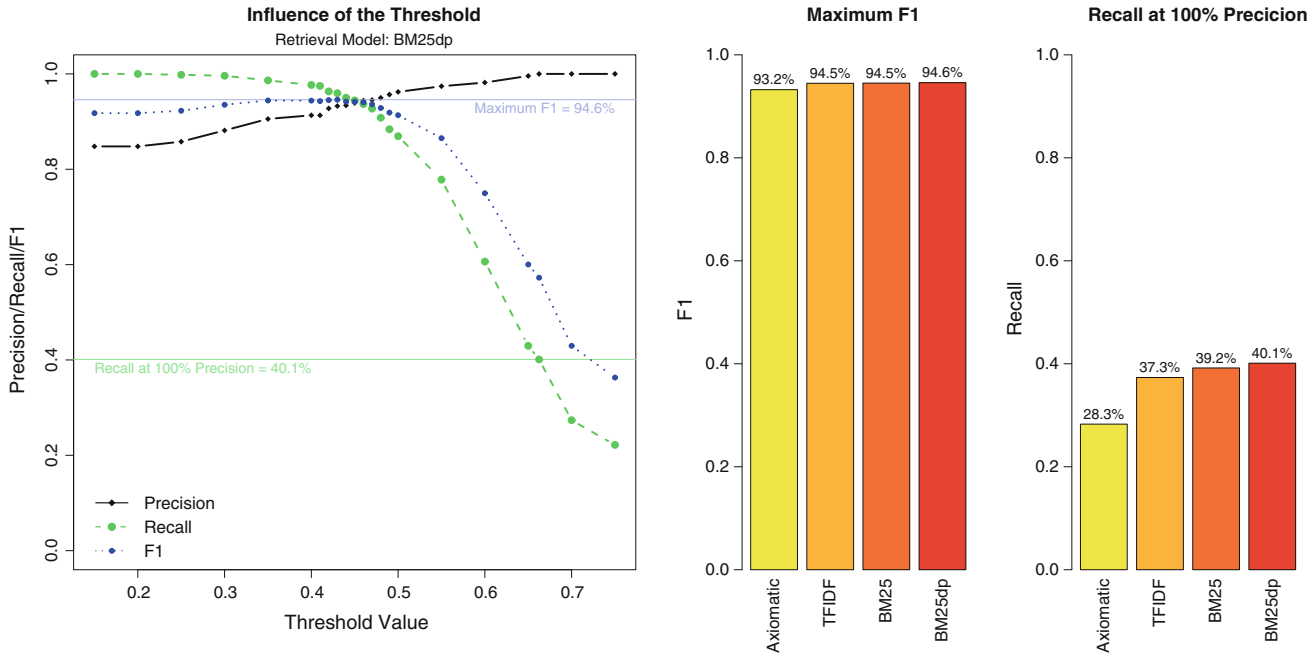

Fig. 3 Precision/Recall/F1 curve for various threshold values for a single retrieval model (BM25dp) on the *left side*. On the *right side*: Comparison of all evaluated retrieval models using the two main quality indicators. The modified BM25 retrieval model achieves the highest overall performance

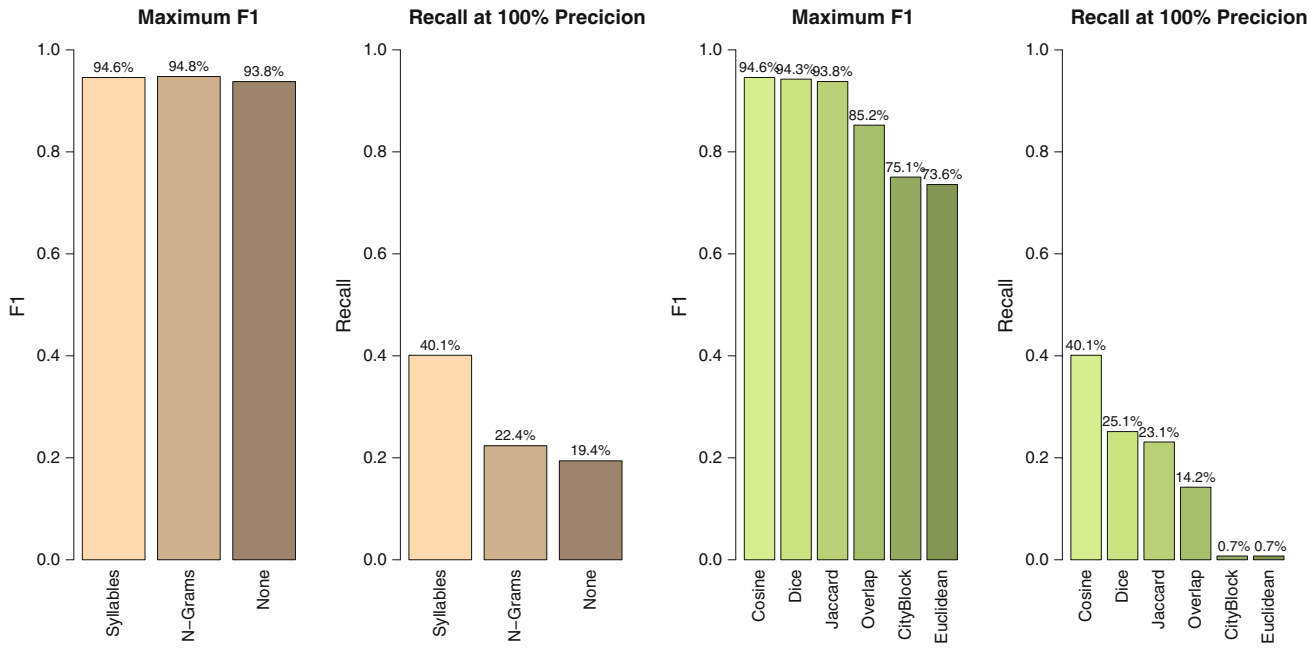

Fig. 4 Comparison of the word-compound processing strategies on the *left side*, and the performance of the various similarity measures on the *right side*. The compound splitting method based on hyphenation pattern outperforms the *n*-gram-based method, which is still better than no splitting at all.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

occurs, if the two selected articles describe the same person, entity or concept. A new article is an article in one encyclopedia that has no counterpart in the other encyclopedia, i.e., it describes a person, entity or concept that is not described in the other encyclopedia. A near match occurs, if two articles are related, but do not describe the same person, entity or concept and no better match exists. We provided examples to illustrate the meaning of "near" matches to the users. Such examples for near matches are: article "ä" in Brockhaus II is a near match to the article "Umlaute" (German umlauts) in

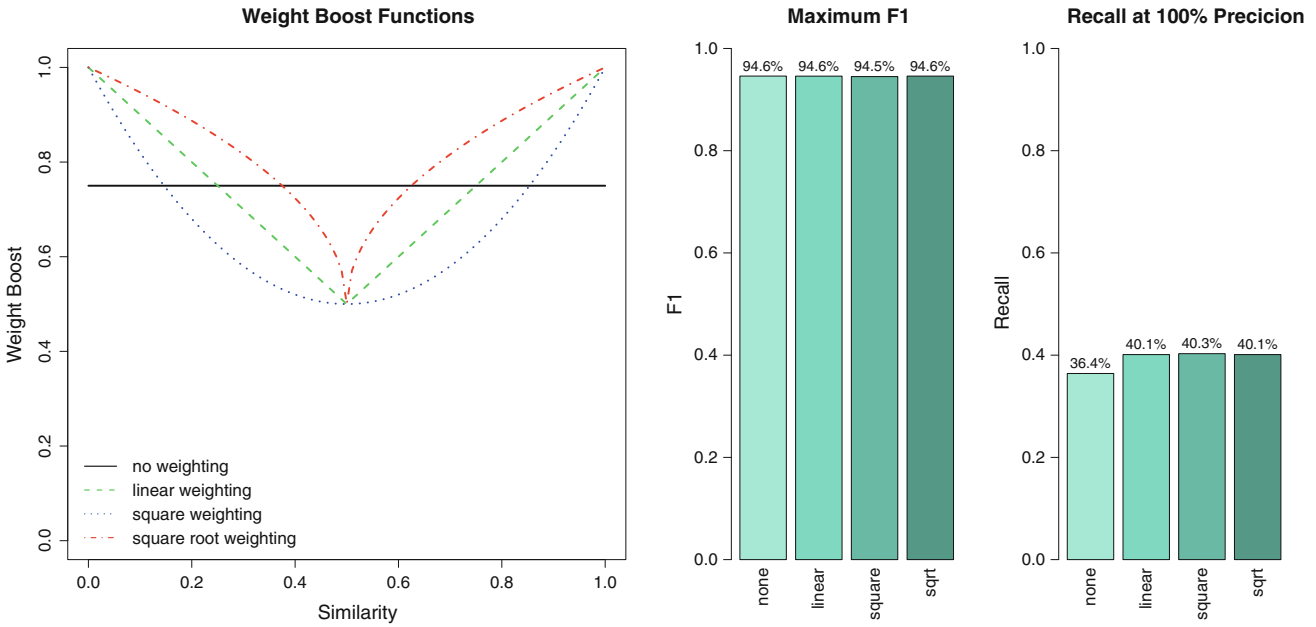

Fig. 5 Weighting functions that capture the intuition that similarities close to the extremes should have a higher impact on the final assignment. The shapes of the boosting function are depicted on the *left* and the performance indicators on the *right*. The improved recall at 100% precision indicates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

Table 3 Performance indicators for each facet when left out of the similarity calculation

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

Each facet appears to contribute to the quality of the alignment with the *Content* facet being the most important one

WissenMedia. Article "Adam und Eva" in Brockhaus III is a near match to both, the article "Eva (Theologie)" and the article "Adam (Theologie)". The latter example shows that there does not always exits one exact mapping, i.e., some articles are undecidable.

5.1 Quality assurance

Including quality assurance methods in the manual alignment process is necessary to find and correct errors, which finally leads to a higher quality of the resulting encyclopedia. Our methodology aims at preventing errors, finding errors and probable reasons of the errors as early as possible in the process. Thereby, we not only corrected the errors but also aimed at detecting and correcting the cause of these errors. For instance, if we found that a user obviously misunderstood the meaning of a "near match" we not only corrected the erroneous mappings but also notified the user about this misunderstanding.

The quality of the resulting alignment is influenced by three aspects: the correctness of the task performed by users, the level of support users get from the application and the means to detect and correct errors (quality control).

Correctness of the tasks Probably the most important part is to make sure that users (human alignment workers) fully understand the tasks. The users need to develop an understanding of what it means to correctly align two articles, what is the meaning of an exact match and when an article in the source encyclopedia is regarded as not available in the target encyclopedia.

Supporting application The application needs to support the subtasks of finding relevant articles and storing the users' decisions. Further, a user interface should make it easy and efficient for the user to achieve the desired result. Some of the features for the manual alignment are frequently used (search, alignment buttons). These features should be easily accessible to support the repetitive task of the user.

Quality control One needs to have a possibility to check the assignments of users and correct errors if needed. There are several reasons why errors might occur: the user is

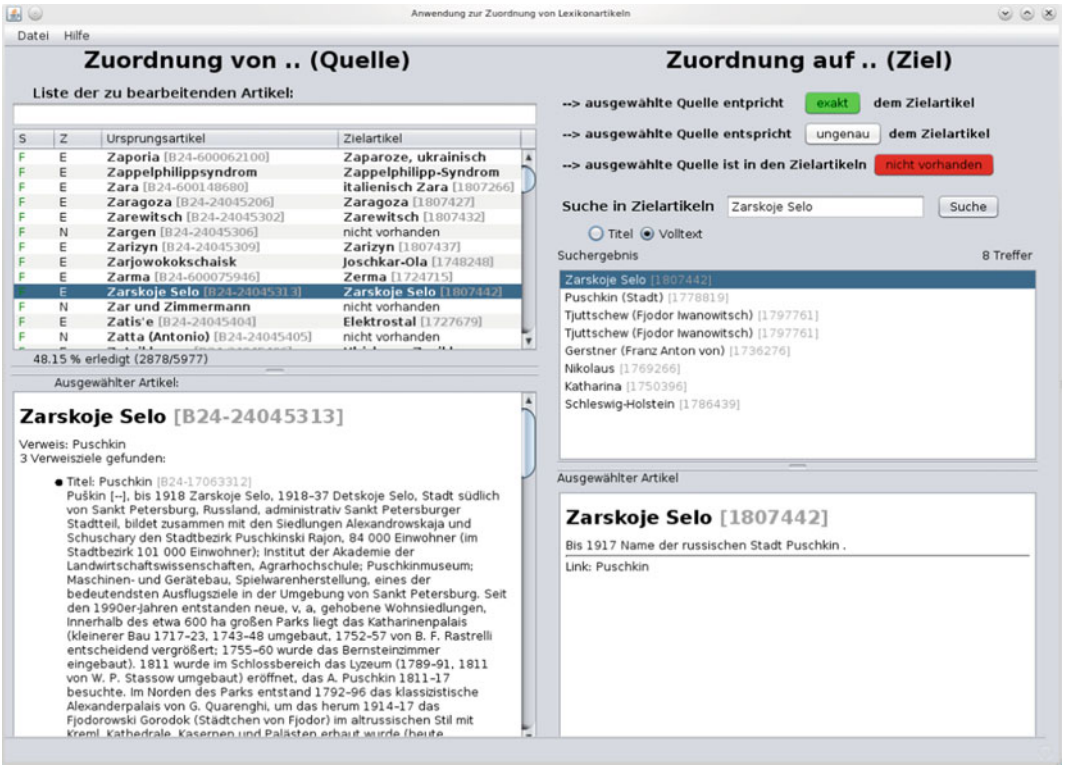

Fig. 6 Screenshot of the application used for the manual alignment of repositories. The *left part* shows the articles to process by the user (encyclopedia Brockhaus I, II or III), the *right part* shows possible matches in the WissenMedia encyclopedia. The three buttons on the *top right* are used to mark the selected articles either as match, as probable match or to mark the selected Brockhaus article as not available in the WissenMedia encyclopedia

inattentive, the tasks were not fully understood or the interface is not handled correctly (for instance, the user only searches the title and forgets to also search the full-text).

We propose the following steps to account for these three aspects:

- Usability testing of the application and using the gained insights of these tests to improve the application.
- Perform a kick-off session with users to explain the task, demonstrate the application and answer questions. Hand out a written task description and example mappings.
- Define a set of articles that are processed by more than one user (overlap set). Check the finished mappings of each user at an early stage, identify the types of errors, the causes of the errors and give feedback to the users.
- Check the final mappings to get an estimate of the remaining misalignments.

5.1.1 Usability test

Usability testing aims at gaining an understanding of how "real" users would use an application and to uncover usability problems. These findings can then be used to finally improve the application. We performed an informal qualitative usability test with two test users. The users were provided with a prototype of the application (a predecessor of the application depicted in Fig. 6) and introduced to the task. They were asked to perform as many mappings as possible within 30 min. The test was repeated two times with the different configurations of the field "Suchergebnis" (the list of search results, text area in the middle right in Fig. 6). In one configuration, this field was left empty. In the other configuration, the field was filled with the results of the automatic alignment algorithm of the currently selected article in the table of the source encyclopedia (left in Fig. 6). In an informal interview, the users were asked what could be improved in the application and which of the two configurations for the search result text area they preferred.

The users made the following suggestions, which we implemented in the final application: (i) The assignment buttons should be colorized to better distinguish between them; (ii) The list of articles from the source encyclopedia should be ordered alphabetically, such that users can use their short-term memory for subsequent, similar articles; (iii) The title of the article currently selected in the source encyclopedia should be automatically copied to the search field; (iv) The results of the automatic alignment algorithm should not be displayed in the search result box. Users tended to blindly

select one of those articles without checking the alignment and without any further investigation of whether there exists a more appropriate match.

5.1.2 Quality control

For controlling the quality of the manual alignments, it is necessary to have a ground-truth, i.e., a subset of articles with correct mappings. These articles are then mixed into the set of articles given to the human alignment worker. To compute the errors, the mapping made by the human alignment worker is compared with the known correct mapping. An error is defined as difference between the mapping of the human alignment worker and the correct mapping. Note, that we assume that the ground-truth (which also was generated by a human alignment worker) always represents the correct alignment. If such a ground-truth can not be generated (e.g., due to missing human experts), an error estimate is still possible by evaluating a set of articles processed by two users. The mappings for these articles are compared and disagreements are counted. The error rate is then lower or equal than the number of disagreements. Note that we only know that there is a disagreement, we do not know which of the alignment workers made the wrong alignment. We even do not know if there is an error at all (as mentioned above, there are undecidable articles). We refer the set of articles that are either part of ground-truth or are processed by two users as "overlap set". In the following we describe a scheme to partition the article assigned to the users and how to integrate the overlap sets. Assume n users. Let S be the set of articles of the source encyclopedia. Divide the set into i disjoint subsets S_i with equal numbers of articles. $S = \bigcup S_i,\ S_i \cap S_j = \emptyset,\ |S_i| = |S_j|,\ 1 \leq i, j \leq n, i \neq j$. For each subset, S_i define a subset C_i which contains the overlapping articles. $C_i \subset S_i$. The overlap set C is then given by $C = \bigcup C_i$. By definition, all articles in the overlap set are distinct. The overlap set can either be processed by one single expert user (generation of the ground-truth) or be equally distributed among all users, such that user k gets $S_i \cup C_j,\ i \neq j$ assigned. In other words, the articles form the ground-truth are equally distributed over all users, such that each user has the same number of articles.

As we mentioned before, we want to detect and correct errors as early as possible in the process. Therefore, we propose an error check after the users have finished a part of their assigned articles. The number of required finished articles depends on the size of the overlap set and whether the ground-truth is available in advance or the users are checked against each other. With the size of the overlap set being 10% of the articles and the ground-truth being available in advance, we propose a checkpoint after 1000 articles resulting in approximately 100 overlap articles per user. The differences between the ground-truth and the user's assignment are counted and manually investigated in detail. Table 4 shows an overview of typical types of errors. If one user tends to repeat the same type of error frequently, it might be the case that the user did not fully understand the task. For instance, often occurring errors of type UE and EU suggest that the user did not fully comprehend the difference between exact and near mapping. We propose to notify users about their overall performance, their most frequent types of errors and their erroneous mappings. For users whose mapping performance is significantly below the average, we propose to perform another check point. After they have processed twice as many articles as for the first check point, their alignment quality is assessed again.

After all users finished their assigned mappings, a final quality check should be performed to get an estimate of the number of errors remaining in the resulting encyclopedia.

Table 4 Possible categories of error-types

Error type	Description
NN	An article that should have an exact or near mapping was not mapped
UE	Article was mapped as an exact match but should have been a near match
EU	Article was mapped as near match but should have been an exact match
F1	Article was mapped but should have been not mapped at all
F2	Article was mapped to the wrong target article (which is not a synonym)
S	Article was mapped to a synonym instead of the article with the same title
X	Not decidable (different correct mappings are possible)

From the detected mistakes made by the users we identified several groups of errors

6 Combined alignment

To build an efficient alignment system, we had to find a balance in providing a high quality and minimizing the tedious manual alignment effort by humans. Therefore, we supplemented the automatic alignment algorithm with an subsequent manual alignment step. To keep the number of articles for manual alignment low, we tried to create a system that automatically aligns as many article as possible with high precision. For each article from the source encyclopedia, the alignment algorithm produces a ranked list of matching articles from the target encyclopedia. The ranking of the articles is based on the similarity weight as computed by the algorithm. This weight also serves as confidence of the alignment. We defined two thresholds to determine whether the result

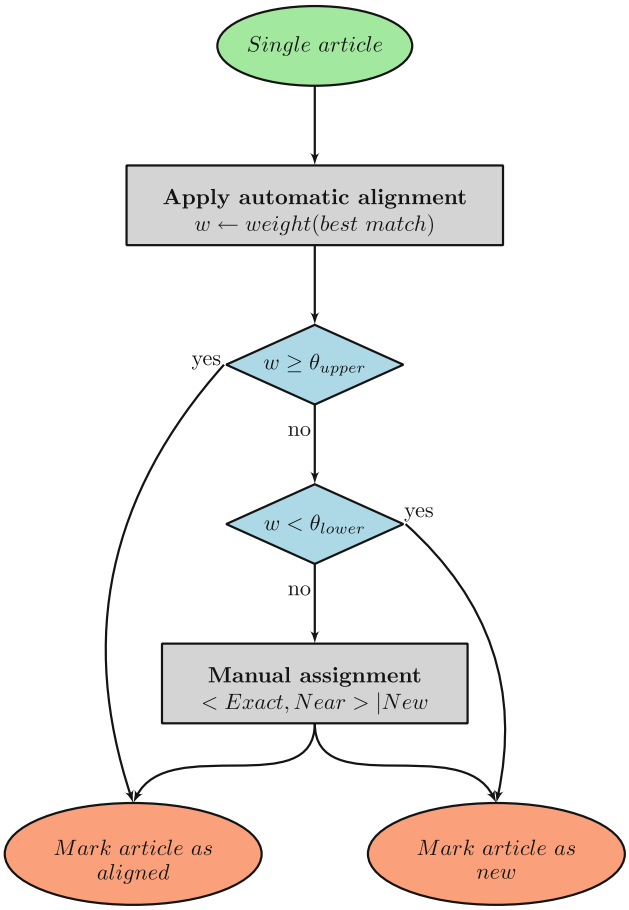

Fig. 7 Overview of the alignment of a single article for the combined system. At first the automatic alignment is employed, and if weight of the top ranked alignment candidate lies between the two thresholds, it is scheduled for manual alignment

of the automatic alignment should be used as a final decision or a manual alignment is necessary. If the weight is higher or equal to the upper threshold—θ_{upper} —the alignment candidate is marked as successfully aligned. If the weight is below the lower threshold—θ_{lower}—then the article is marked as new article without any matching article in the target encyclopedia. The article has to be manually aligned if the weight lies between the two thresholds (Fig. 7).

6.1 Automatic alignment

For the automatic alignment, we resorted to the configuration that yielded the highest performance in the evaluation. We used the BM25 weighting scheme applied on terms generated by the hyphenation-based compound word splitting processing. Based on the results of the evaluation and additional samples evaluated by domain experts, we set the thresholds values to $\theta_{\text{upper}} = 0.42$ and $\theta_{\text{lower}} = 0.3$ for the final alignment process.

Table 5 Overview of the number of articles in each category after applying the thresholds on the results of the automatic alignment process

	Brockhaus I	Brockhaus II	Brockhaus III
Exact	33,002	61,123	86,376
Unsure	7,330	22,937	55,977
New	1,676	8,200	31,368
Total	42,008	92,260	173,671

Exact: successfully aligned by the automatic method; *New*: there is no match in the WissenMedia encyclopedia; *Unsure*: the confidence value from the algorithm is too low for a definitive decision

Table 6 Number of articles for manual assignment, number of articles processed by two users (overlap) and total number of users for a specific encyclopedia

Encyclopedia	Articles	Overlap	Users
Brockhaus I	7,330	500	2
Brockhaus II	22,937	2,000	5
Brockhaus III	55,977	5,000	12

Table 5 shows the number of exact matches determined by the automatic alignment algorithm by applying the thresholds θ_{upper} (exact), the number of articles not available in the target encyclopedia by applying the threshold θ_{lower} (new) and the number of articles for which no decision could be made (unsure). Articles between the two thresholds were subsequently processed manually.

6.2 Manual alignment

6.2.1 Process

Each of the users had to align a set of 5000 articles. 10% of the articles were randomly selected for the overlap sets C_i. The overlap sets C_i were assigned to a separate reference user. Table 6 shows the number of articles for manual processing for each encyclopedia, the size of the overlap set and the number of users.

In a project kick-off meeting with all users, the task was explained, the application was demonstrated, example mappings were given and the users had time to ask questions. Further, a document consisting of the task description and example mappings was distributed to the users. The first check point was scheduled after 20% (1000) articles. If a second check point was necessary, it was scheduled after 40% (2000) articles.

6.2.2 Results

First check point The intermediate results of the users were checked after they had finished 1000 articles. The results

Table 7 Error types and occurrences

Error type	Number	% of Articles	% of Errors
NN	72	6.7	37.5
UE	7	0.6	3.6
EU	14	1.3	7.3
F1	17	1.6	8.9
F2	17	1.6	8.9
S	29	2.7	15.1
X	36	3.4	18.8

Occurrences are summed over three example users. 1,073 number of checked articles were checked, and 192 errors were found in total

were checked against the ground-truth: First, the assignments were automatically compared. Then, the articles for which the user and the ground-truth disagreed were manually checked again. For some users, we were able to identify types of errors that occurred regularly, e.g., one user regularly mapped articles to synonyms instead to the article with the same title. Each user was sent an error report and requested to correct the errors. The average error rate after 20% of the articles were processed was 17.3%. The results varied widely over the different users, the standard deviation of the error rate was 14.6% (min: 2.6% max: 45.5%). For two users, the alignment quality was significantly below the average, thus the mappings of these users were re-examined at the second check point.

Second check point The alignment quality of the two worst performing users did improve at the second check point. The error rate of User 2 dropped from 45.5 to 10.10%, and the error rate of User 10 dropped from 40.6 to 4.6%.

Final results The final error rate was about 10%. For approximately 4% of the articles, an agreement on the correct mapping could not be found. For three example users, the errors were evaluated in detail. Table 7 shows the occurrence of the different types of errors, the percentage of this type in the total number of errors and the percentage in the total number of articles. As can be seen the error type NN (a mapping was not found) was the dominant type of error (37.5% of the total errors, 6.7% of all articles). 15.1% of the errors resulted in assigning synonym articles instead of the article with the same title (error type S). 18.8% of the deviations from the ground-truth are due to non-decidable articles (error type X).

In the following some examples of errors that occurred for the three example users are listed.

- The article “Lazaro Cardenas“ was mapped to ”Cardenas (Lazáro)“. This mapping is wrong, because the former article describes a location, the latter a person. (error type: F1)
- The article about the town ”Kolarowgrad“ was not mapped, but the article ”Shuman“ describes the same town, but uses the older name of the town as title. (error type: NN)
- The article ”Konzeptalbum“ (concept album) was not mapped but the article ”Konzept-Album“ describing the same concept does exist. (error type: NN)
- The article ”Kochtla-Jarwe“ was mapped to the synonym ”Kohtla-Jaerve“ although the article with the same title ”Kochtla-Jarwe“ exists in the target encyclopedia. (error type: S)
- The article ”Kerkyra“ describing the town was assigned as a near match to ”Kerkyra“ describing the island although the article ”Kerkyra (Stadt)“ describing the town exists in the target encyclopedia. (error type: F2)

Table 8 Overview result of the manual alignment step

	Brockhaus I	Brockhaus II	Brockhaus III
Exact	4,334	11,083	20,459
Near	581	4,395	7,743
New	2,415	7,459	27,775
Total	7,330	22,937	55,977

Table 8 shows the final numbers of exact and near matches and new articles after the manual alignment step.

6.3 Integrating results of manual and automatic alignment step

The results of the automatic and manual alignment step were integrated into one single result by adding the exact matches from the manual step to the exact matches from the automatic step, adding the new articles from the manual steps to the new articles from the automatic step and generating a separate list for the near matches from the manual step. Table 9 shows the results of the integration step. The final goal of our alignment system is to provide an infrastructure to merge multiple encyclopedias into one corpus. The articles unique to each of the single encyclopedias are highly relevant, as they immediately help to increase the coverage of the merged encyclopedia.

7 Conclusion and future work

Currently, the state-of-the-art in machine text understanding has not reached a level which would allow us to use a completely automatic system for encyclopedia alignment. Nevertheless, algorithmic approaches can be utilized to align a

Table 9 Overview results of the combined alignment step

	Brockhaus I	Brockhaus II	Brockhaus III
Exact	37,336	72,206	106,835
Near	581	4,395	7,743
New	4,091	15,650	59,143
Total	42,008	92,260	173,671

About one third of the articles in the largest Brockhaus encyclopedia are not contained in the WissenMedia encyclopedia, which shows that the two encyclopedia to a certain extend cover different topics

subset of all articles and to support humans in their task of finding a correct match.

There are a number of challenges for the automatic alignment, some of them are unique to the way encyclopedic corpora are written. For example, the style of the German language within encyclopedias differs from the common language usage because of the terse nature of the articles.

We created an encyclopedia alignment algorithm by applying techniques developed in the field of IR. 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 a 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.

For articles that yielded a low confidence score in the automatic alignment process, we have set-up a manual alignment procedure. Articles were assigned to more than one person to detect and also to prevent errors. We did an in-depth error analysis and identified several groups of common mistakes. Finally, we applied our combined system of automatic and manual alignment to align four German encyclopedias.

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.

Acknowledgments 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. Anderka, M., Stein, B.: The ESA retrieval model revisited. In: Sanderson, M., Zhai, C., Zobel, J., Aslam, J. (eds.) 32th Annual International ACM SIGIR Conference (SIGIR 09), pp. 670–671. ACM (2009). doi:10.1145/1571941.1572070
2. Bernstein, Y., Zobel, J.: A scalable system for identifying co-derivative documents. In: String Processing and Information Retrieval Symposium, pp. 55–67 (2004)
3. Bouma, G., Duarte, S., Islam, Z.: Cross-lingual alignment and completion of wikipedia templates. In: Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies, CLIAWS3 '09, pp. 21–29. Association for Computational Linguistics, Stroudsburg, PA (2009)
4. 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, pp. 480–487. ACM (2005)
5. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the Twentieth International Joint Conference for Artificial Intelligence, pp. 1606–1611. Hyderabad (2007)
6. Gries, S.: Dispersions and adjusted frequencies in corpora. Int. J. Corpus Linguist. **13**(4), 403–437 (2008). doi:10.1075/ijcl.13.4.02gri
7. 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, pp. 167–171. ACM, New York, NY (2009). doi:10.1145/1643823.1643854
8. Li, Y., McLean, D., Bandar, Z.: Sentence similarity based on semantic nets and corpus statistics. IEEE Trans. Knowl. Data Eng. **18**(8), 1138–1150 (2006)
9. Liu, X., Zhou, Y., Zheng, R.: Measuring semantic similarity within sentences. In: Proceedings of the 7th International Conference on Machine Learning and Cybernetics, ICMLC, vol. 5, pp. 2558–2562 (2008). doi:10.1109/ICMLC.2008.4620839
10. Marko, K., Baud, R., Zweigenbaum, P., Merkel, M., Gronostaj, M.T., Kokkinakis, D., Schulz, S.: Cross-lingual alignment of medical lexicons. In: Workshop on Acquiring and Representing Multilingual, Specialized Lexicons: the Case of Biomedicine (2006)
11. 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, pp. 517–524. ACM (2005)
12. 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, vol. 4953, pp. 172–181. Springer (2008)
13. Pedersen, T.: Computational approaches to measuring the similarity of short contexts: a review of applications and methods. Comput. Res. Repos. (CoRR) abs/0806.3 (2008)
14. Rector, L.H.: Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. Ref. Serv. Rev. **36**(1), 7–22 (2008). doi:10.1108/00907320810851998
15. Robertson, S., Gatford, M.: Okapi at TREC-4. In: Proceedings of the Fourth Text Retrieval Conference, pp. 73–97 (1996)
16. 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, pp. 377–386. ACM (2006)
17. Yih, W., Meek, C.: Improving similarity measures for short segments of text. In: AAAI'07: Proceedings of the 22nd National Conference on Artificial Intelligence, pp. 1489–1494. AAAI Press (2007)