Methods for Collection and Evaluation of Comparable Documents

Monica Lestari Paramita, David Guthrie, Evangelos Kanoulas, Rob Gaizauskas, Paul Clough, and Mark Sanderson

University of Sheffield, Regent Court,
211 Portobello Street, Sheffield S1 4DP, UK {m.paramita,d.guthrie,e.kanoulas,
r.gaizauskas,p.d.clough,m.sanderson}@sheffield.ac.uk

Abstract. *Considerable attention is being paid to methods for gathering and evaluating comparable corpora, not only to improve Statistical Machine Translation (SMT) but for other applications as well, e.g. the extraction of paraphrases. The potential value of such corpora requires efficient and effective methods for gathering and evaluating them. Most of these methods have been tested in retrieving document pairs for well resourced languages, however there is a lack of work in areas of less popular (under resourced) languages, or domains. This chapter describes the work in developing methods for automatically gathering comparable corpora from the Web, specifically for under resourced languages. Different online sources are investigated and an evaluation method is developed to assess the quality of the retrieved documents.*

Keywords: Comparable corpora, retrieval methods, evaluation

1 Introduction

The Web contains a vast number of texts authored in a multitude of languages. Crucially, some of these texts are available in multiple languages with varying degrees of correspondences, ranging from parallel versions to describing similar concepts or themes. Texts with a high degree of correspondence can be used to improve Statistical Machine Translation (SMT) systems and work has been undertaken in the past decade to develop methods to automatically retrieve such texts in order to build parallel and comparable corpora. However, most of these methods have only been tested in the context of retrieving document pairs for well resourced languages. The performance and applicability of these methods can differ significantly when they are applied to under resourced languages.

ACCURAT[1], which stands for Analysis and evaluation of Comparable Corpora for Under Resourced Areas for machine Translation, is an EU project which aims to investigate the use of comparable documents when parallel corpora are not readily available. This project focuses on under resourced Eastern European

[1] http://www.accurat-project.eu/

languages, which include Croatian, Estonian, Greek, Latvian, Lithuanian, Romanian and Slovenian. The aim of ACCURAT is to analyse the use of comparable documents for under resourced languages in order to improve the performance of machine translation. The work presented in this chapter is one of the tasks in the ACCURAT project which specifically focuses on developing methods to locate and download comparable documents from the Web.

In the ACCURAT project three levels of comparability have been used. The first level (parallel corpora) represents parallel documents, which are direct translations, although minor language-specific variations are allowed. The second level (strongly comparable) contains pairs of documents about the same topic or derived from the same source. The third level (weakly comparable) represents documents about different topics, but from similar domains or genres.

This chapter starts with a review of related research: previous retrieval approaches are detailed and assessed for their suitability for under resourced languages in Section 1. In Section 3, a series of novel approaches to retrieve comparable documents and preliminary results from each Web source are discussed. In Section 4, we describe our evaluation methods of the comparable corpora, which are focused on measuring the effectiveness of the retrieval methods.

2 Literature Review

There has been a range of previous work related to the tasks of gathering or identifying comparable corpora. The methods involved in the retrieval process can be categorised into two major processes: Web crawling and alignment. The use of Web crawling techniques is described in Section 2.1. The crawling process results in a large collection of unaligned multilingual texts. Several methods are then performed to identify comparable documents. These methods are described in Section 2.2.

2.1 Web Crawling

Many tools and approaches have been developed to build comparable corpora using techniques based on retrieval. For example, BootCat [6] retrieves documents from a list of seed words. Outputs can then be used to bootstrap the process by inserting more seed words to improve the recall of the document retrieval stage. This approach assumes that the retrieved results are relevant and satisfy the requirements of the query. Other approaches, on the other hand, perform an evaluation to check the relevance of results. If relevant, the result is used to enhance some underlying language model, or included in the collection to generate a query; otherwise, the results are not considered. This approach is referred to as focused crawling [11] and is shown to retrieve relevant documents in narrow domains more effectively than using general purpose crawlers.

Talvensaari et al. [41] implemented focused crawling using keywords as the input seeds. In an approach that differed from BootCat, they did not specifically look for relevant documents; rather they used the retrieval result to look for websites that consistently produced top results over the majority of these queries. These websites were seen as good resources for that particular domain and were crawled to retrieve all documents within it. Language was detected using a simple n-gram based algorithm [10].

Ghani et al. [19] implemented a different approach, which they call CorpusBuilder. Instead of using query seeds, they used a set of documents previously judged as relevant and non-relevant to a given query(or set of queries). To focus retrieval on documents of under resourced languages, in this case Slovenian, they used Slovenian documents for the relevant documents and those from other languages as the non-relevant documents. They investigated the performance of several query generation methods; and found that an approach based on odds-ratio resulted in the highest performance, compared to term frequency or random sampling baselines. The odds-ratio of each word is calculated by using the probability of the word occurring in a relevant and non-relevant document. A further difference of this method compared to others was that the query used both inclusion and negation of terms. Highest performance was obtained by using 3 positive and 3 negative keywords, each chosen based on the highest odds-ratio score of the sets of relevant and non relevant documents. After each retrieval operation, the first document was passed to a language filter. If this document was identified to be in Slovenian, the set of documents was updated, and query generation was performed again. In case the new document did not change the query, the next ranking document in the result was taken as a result, and this process was performed iteratively. This method managed to retrieve general corpora from minority languages effectively.

2.2 Identifying Comparable Text

Given a large collection of unaligned multilingual texts, a range of approaches have been used to align parallel documents or sentences automatically. For example, [35] and [51] used the HTML structure and URL paths of documents in order to find parallel texts on the Web. These approaches are language-independent, however they are not applicable to retrieve comparable documents since such documents do not always share the same URL paths or contain similar HTML structure. Other approaches to align comparable documents require a range of linguistic resources, from bilingual dictionaries, parallel corpora to machine translation systems.

Dictionaries can be used in a straightforward manner to translate the words (and phrases) in a document and these terms can be used as the query to an IR system in the target language. However, ambiguity can be an issue if a word (or phrase) has multiple interpretations, and therefore, translations. Problems also occur when a word does not exist in the dictionary, i.e. out-of-vocabulary term. To solve the latter problem, cognate matching can be used to identify the translation of a word (e.g. “colour” in English and “couleur” in French). However, this method only applies for languages with the same etymological roots and using the same writing system. If multilingual parallel corpora exist alignments can be computed and the resulting aligned texts can be used to build a statistical machine translation system. Unfortunately, this approach is computationally expensive and still has to deal with the problem of out-of-domain vocabulary. In addition, it can also be difficult to gather enough resources for machine translation because of the limited amount of parallel corpora available and accessible on the Web, particularly for under resourced languages. Most parallel corpora only cover a specific domain, such as law, which may cause problems with translation as the system may perform poorly when used to translate documents from a different domain [41].

Munteanu & Marcu [30] attempt to identify pairs of parallel sentences from large collections of comparable news documents in several languages. They first align corpora of Arabic and English news documents by building a query from each Arabic document based on translating every word in the Arabic document to English using a bilingual dictionary. The highest ranked 100 English documents were retrieved for each Arabic document. Documents published outside of a specified time window from the Arabic collection were filtered out. Each sentence in the Arabic document and those in each of the English documents were paired, and certain features were evaluated in each sentence pair to identify which sentences were parallel.

Argaw & Asker [4] aligned news articles in Amharic and English published on the same date and occurring in the same place. This information is available in the metadata of the articles. No lexical resources were used to translate the words, instead [4] performed transliteration on the titles and calculated the edit distance between words in the titles. Pairs of documents which scored above a certain threshold were considered as comparable. Other approaches which have been used to align documents are based on overlapping named entities [21] and clustering documents [40].

Fung & Cheung [18] aligned non-parallel corpora by using parallel sentences from bilingual corpora to retrieve new documents. These documents were likely to have different topics and therefore not found by standard keyword searching on the topic, named entities or dates. However, as they contained similar sentences they tend to share similar terminologies which could be used as a source of parallel data.

3 Retrieval of Comparable Documents

In this section, we describe new techniques that have been used to gather comparable documents from the Web. First, we identified different Web sources considered as promising sources of comparable documents, such as news, Wikipedia and Twitter. The characteristics of each source were explored in detail to enable retrieval methods to be created effectively. We then developed different retrieval techniques to gather comparable documents from each source. Techniques which we use to collect articles from news sites are described in Section 3.1. In Section 3.2, we focus on techniques to retrieve comparable documents from Wikipedia. Techniques developed to extract data from Twitter are described in Section 3.3.

3.1 News

News articles are continuously being published on the Web from news agencies across the globe in a variety of languages. These news stories can be highly similar, or even parallel across languages, because they are produced by the same agency written from the same newswire feed (e.g. Associated Press, Press Association or Reuters) or simply because they are reporting the same topic or story. The availability of large amounts of similar texts across languages makes the news domain an extremely promising area in which to perform comparable document mining. In this section we give an overview of our methods to automatically retrieve news data for the construction of comparable corpora. We show how collections with aligned document pairs can be produced and that these methods are useful for under resourced languages. We also show how it is possible to dramatically decrease the number of documents that must be compared using information retrieval and by using additional processing how it is possible to increase the accuracy of the method.

Although there exists an abundance of news articles on the Web, and many of these may be comparable in some way, identifying particular news articles that are the most similar can, in practice, be problematic. For example, often news stories are running stories rather than one-off events: they describe ongoing events that can proceed over the course of days, weeks, or even years with many updates that have little difference in their focus or content. Take, for instance, the March 2010 news article coverage of the Icelandic volcano eruptions. News articles concentrated on a range of subjects: when eruptions occurred, the drifting cloud of ash, the environmental impacts, and the assorted disruptions to air travel. Many of the articles during this period are very similar and contain pieces of overlapping information, so they could be considered weakly comparable. However, a smaller subset of these documents may actually be nearly identical in their focus and have a much stronger level of comparability. It is this special subset of similar documents that we aim to identify and match as they contain large sections of information that match across languages and are thus the most useful for improving SMT systems.

We approach this problem of identifying comparable news documents as a type of Cross-Language Information Retrieval (CLIR) task where the goal is to find, for each news article in a certain language, news articles that report the same information in other languages. Some articles may not have matching information in one or more languages and for these articles we would either like to identify articles that are close matches, if they exist, or else judge that the article does not have a match. This setup requires that we perform the CLIR task for a set of seed documents in one language (e.g. English) and find all matches in the target language for each of these documents. This can be computationally intensive, but it critically does not require us to perform the much more expensive and impractical computation of comparing every document in one language with every other document in another language, as long as we choose our method of performing CLIR carefully.

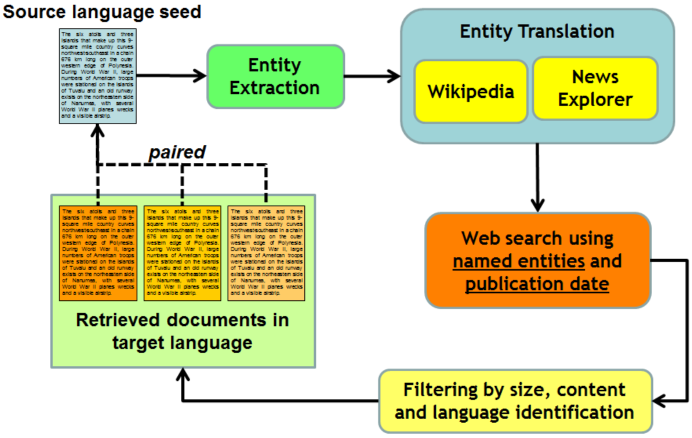
Cross-language information retrieval often involves translation of the query into the target language before standard information retrieval in the target language. This requires translation of every word (or phrase) in the query either using a dictionary or using a machine translation system. The use of a dictionary only allows for words (or phrases) to be translated independently and often gives results that are unsatisfactory, since it is difficult to pick the correct translation for a word due to ambiguity in both languages and limited context given the short length of the query. Machine translation can be more accurate and simpler, but is computationally expensive and requires parallel corpora for training the system. Both of these methods are not very suitable for languages where few electronic resources are available, which is one of the focus areas for our method of gathering comparable corpora. We instead adopt an approach that uses only limited language resources by making use of the unique properties of news articles.

We make use of a useful feature of news articles: the high frequency of proper nouns, such as people, places and organizations. News articles often contain a large number of such entities and identifying them and classifying them as to their type is a well-established problem called Named Entity Recognition (NER) and can be performed with high accuracy [20] [47] [2]. We make use of these entities to query for documents that may be comparable. For this purpose it is not necessary to perform the full Named Entity Recognition task, as the class of these entities is not needed. It is only important to extract these entities from every news article and not label them with their semantic category (e.g. person, place, organization, etc.).

We also exploit the fact that news articles typically have dates associated with them: possibly the date the story was written, the date of some event, or the date it was published to the Web. We propose to make use of a search method that uses the date articles are published to the Web. Publication date is available for every document and does not have the resolution issues associated with dates found in the article, which can refer to past or future events. Making use of these dates allows us to significantly reduce the search space for documents. Instead of performing a CLIR search over all similar news documents on the Web, we search only the news documents published at the same time or within a few days of each other. It seems very likely that limiting the search increases the precision of the search because similar articles will have been written at similar times. It has been shown in previous work that even with these restrictions it is still possible to obtain a large collection of results [31][16]. The process we developed to retrieve news documents is represented in Figure 1 and details involved in each process are described in more details in the following sections.

3.1.1 Collecting Source Language Seeds To build this large collection of comparable text we first gathered collections of news articles in English that we believed likely to have comparable documents in some other project languages. We make use of News Explorer to identify these popular news articles. News Explorer (http://emm.newsexplorer.eu/NewsExplorer/), developed at the Eu-

Fig. 1. Retrieval Process of News Documents


Source language seed
Entity Extraction
Entity Translation
Wikipedia
News Explorer
paired
Web search using named entities and publication date
Retrieved documents in target language
Filtering by size, content and language identification

ropean Commission's Joint Research Centre (JRC), is a tool for multilingual news gathering, analysis, and exploration. News Explorer monitors news from approximately 1,500 news portals worldwide. We gathered popular news articles by collecting the top 15 news stories in newsexplorer.com for every day from January 2008 to November 2010. These stories have the most news coverage for that day (across languages) according to NewsExplorer. Other English collections were gathered from more focused sources in order to target specific languages, for example, a collection of all news stories from the Romanian Times (www.romaniantimes.com) newspaper which covers local news relevant to Romania, but is written in English. These articles were then post processed to extract only the main news story. We disregard any very short articles whose cleaned text is less than 1,600 characters long. For each article we store the original HTML page, the date it was published, and the cleaned main text from the web page. In total, 29 million words of cleaned English text was gathered. The exact sizes of the corpora and their sources are shown in Table 1.

Table 1. Size of English Documents Seeds

Corpora Source	Size in Words (Excluding Markup)
Athens News	4,984,796
Baltic Times	3,307,234
Croatian Times	355,809
NewsExplorer	7,589,978
Nine O'Clock (Romanian)	13,150,159
Romanian Times	196,426
TOTAL	29,584,402

3.1.2 Named Entity Extraction Our method works by taking every news document in one language (the source language) and extracting all proper nouns and the date the article was published. We then use these entities to perform a search over all documents in the target language published within two days of the source document. To increase the recall of results we do not use all entities that occur in the document, but instead choose only the top n most frequent entities. This procedure works well when searching languages with similar writing systems, as these named entities are often written exactly the same across languages with no translation. Some entities are written slightly differently across languages even with the same writing system, for example, Băsescu, the president of Romania, is often written in English without the breve as Basescu. In languages with very different writing systems, often either translation or transliteration occurs. Google has a transliteration tool available online (http://www.google.com/transliterate/) with an API available to programmers, but in tests this tool did not perform particularly well on proper names. Instead we can achieve much better translation of named entities by making use of resources specifically generated for entities that have had the benefit of human involvement.

3.1.3 Named Entity Translation We then implemented methods to make use of this collection of English documents by searching the web for all documents in other languages that are likely to be comparable with each document in the English collection. The basic method we developed was to identify all proper nouns (entities) in each English document and use the most frequent of these to search for comparable documents written in another language that contained the same entities and were published within 5 days of each other. Specifically, to find news articles comparable to an English article, D, in another language, say L2, we first identify the 5 most frequent entities in D and then attempt to translate them into L2 using one of 4 resources: Wikipedia, Google translate, Bing translate, and NewsExplorer. We make use of the translations recourses in the order listed and only use then next resource if no translation was found (or the resource was unresponsive). Wikipedia is used by checking if the English entity has a wikipedia entry with this title in English and if so we check to see if there is a link to a Wikipedia article in L2 and use its title. We use these translated entities to perform a focused search for News articles in the L2 language.

3.1.4 Web Search We make use of Google News to search once for every document in our source language collection using the most common entities from each source document and limiting the search to within two days of the original source document. We propose to use the top five most frequent entities, but this parameter can be tuned depending on the number of entities that occur in the document and the level of recall required. The use of Google News search allows us to dramatically reduce the search space for every document and with this reduced set of documents we can perform a more exhaustive comparison to determine if they are likely to be comparable. We thus take the documents returned by this search procedure to be a set of candidate comparable document pairs and run additional tests to further limit this set.

An alternative to using Google News search is to use a fixed set of news sites either by crawling and indexing all these sites (as mentioned previously), or simply using this list of sites to restrict a Web search engine, like Google or Bing to these domains. This would require having a fixed set of news sites and then sending a separate search request to every site, whereas Google News searches all sites at once. The use of Google News does not require this list of fixed sites and the sites indexed by Google are likely to be much broader and more comprehensive than a fixed list would provide and are also continuously updated. This allows for a large amount of current news documents to be searched easily without requiring a fixed list of sites.

To perform the language specific focused web search we currently make use of several available search engine services that allow us to search for news documents by date as well as more general search engines that allow us to search by date. Just as in entity translation we rely on multiple services but use them only as backups for cases when the primary service is unresponsive or fails to find any results. The services we use in order are: Google Search, Bing Search, and Google News Search. We query each service for documents that contain all of the the top 5 translated entities and were published within 5 days of the English document and constrain the search to only the desired language (e.g. Croatian) and if no results are found in any service we perform the searches again using only the top 3 entities and then the top 2 entities.

3.1.5 Filtering This procedure will undoubtedly return some documents that are not relevant, so we perform some additional checks before storing a document. We first extract the main story from the web page and ensure that the cleaned text is greater than 1,100 characters long. We also check to make sure the document is not too large and does not contain an unusually large percentage of numbers, punctuation, or uppercase letters, all of which are signs that the documents will probably not be useful for comparability. We then send the document to Bing's language identification service and check that the language returned by the service is indeed the target language in with we were searching for the document (e.g. Croatian). The top 5 results returned for each search query that meet these minimum requirements are saved as likely comparable documents.

3.1.6 Results Our retrieval method has a number of advantages over previous methods described in the related work. First, it does not require us to crawl all news sites in the Web. Instead, only a small set used as initial seeds is required; the rest of the documents are found using a search engine. Second, our method requires no specific lexical resources for named entity translation. We use all data which are publicly accessible, and therefore our methods can be implemented for any other language pairs.

By implementing this retrieval method, we managed to gather a large aligned comparable corpus. The size of comparable texts gathered for each language is shown in Table 2.

Table 2. Size of Automatically Retrieved Comparable Text

Language	Size of Comparable Text Gathered (words)
Croatian	3,802,495
Estonian	11,409,322
German	27,422,578
Greek	13,438,848
Latvian	16,634,981
Lithuanian	28,750,162
Romanian	43,841,777
Slovenian	9,169,704
TOTAL	154,469,867

3.2 Wikipedia

Wikipedia (http://www.wikipedia.com/) is the world's largest multilingual encyclopaedia with over 16 million articles. Wikipedia covers all languages of the ACCURAT project and therefore is seen as a promising source of comparable documents. The number of Wikipedia documents has increased dramatically over the past few years and it covers a wide variety of topics. Even though Wikipedia documents can be edited by anyone, the contents are moderated and this has ensured Wikipedia documents to have reasonably higher quality than other Web documents.

An advantage of using Wikipedia as a source of comparable documents is the interlanguage links feature: a link connecting documents about the same topic but written in different languages. Several works were conducted to extract information from these interlanguage links, such as the creation of bilingual lexicon [1] by performing title extraction of the linked documents. Even though not all documents in Wikipedia have interlanguage links, i.e. documents which are only written in one language, the number of articles written in more than one language is considerably high and ranges between 46% to 84% as shown in Table 3. This data is based on Wikipedia dump in September 2010.

Wikipedia articles linked across languages are about the same entity, process, event, or topic, however there is a disagreement regarding how similar the information contained in these documents are. Mohammadi & GhasemAghee [29] found that many of the articles linked across languages contain very similar information and are thus ideal as a source for gathering comparable corpora. On the other hand, Adafre & Rijke [1] found that these connected documents do not necessarily talk about the same topics as they vary in length and may

Table 3. Percentage of documents with language links

Language	All pages	Page with language links	Average number of links
EN	3,110,586	1,425,938 (45.84%)	4.84
DE	1,036,144	636,111 (61.39%)	7.76
RO	141,284	106,321 (75.25%)	17.24
LT	102,407	67,925 (66.33%)	22.28
SL	85,709	58,489 (68.24%)	21.02
HR	81,366	60,770 (74.69%)	23.47
ET	72,231	49,440 (68.45%)	25.47
EL	49,275	37,337 (75.77%)	29.62
LV	26,227	22,095 (84.25%)	33.36

include additional, removal, or completely different information. Attempting to align every pair of linked documents in Wikipedia to assess their comparability is a computationally expensive process and also unnecessarily wasteful if the articles are very different. Therefore, when gathering texts in Wikipedia using the link structure it is necessary to verify that the text itself is actually comparable.

3.2.1 Identifying Comparable Documents In this section we describe our attempt to maximize the likelihood of finding comparable documents. We implement two simple filters, which are document's minimum size and length's difference. First filter, we eliminate pairs in which any of the documents is smaller than 2 KB. Afterwards, we compare the length between the two documents and disregard document pairs if the length difference is greater than 20%. We implement this feature to focus the retrieval on documents of similar length, assuming that they will have similar structure and content, and therefore have higher probability to be comparable than other documents.

3.2.2 Results An initial analysis find that most of the Wikipedia documents have a diverse size within different languages even though they describe the same topic. For example, the article about "Europe" in English has 12 main sections with the content reaching just under 10,000 words; the corresponding Latvian article, however, has just 6 main sections with a total word count of around 3,000 words. On average, articles from the corresponding under resourced languages have significantly smaller sizes compared to the English version. By finding articles of similar length only, we disregard over 80% of the initial pairs, as shown in Table 4.

Furthermore, by performing manual assessment on some Wikipedia documents, we also found that some under resourced language documents were created by translating a main paragraph of English documents. This pair of documents may have a large difference in lengths, but these translated fragments are particularly useful for machine translation. These documents would not have been retrieved had a length difference filter been implemented as a filter. We

Table 4. Bilingual documents of similar size for each ACCURAT language pairs

Language Pairs	All Pairs	After Size Filtering	After Doc Length Filtering
EL-EN	32,015 pairs	23,206 pairs	3,993 pairs
RO-EL	14,339 pairs	8,725 pairs	1,815 pairs
RO-EN	84,862 pairs	27,234 pairs	3,243 pairs
LT-EN	51,011 pairs	26,804 pairs	2,886 pairs
LV-EN	18,480 pairs	11,893 pairs	932 pairs
SL-EN	44,923 pairs	23,313 pairs	4,893 pairs
HR-EN	43,984 pairs	26,520 pairs	3,628 pairs
ET-EN	37,043 pairs	15,870 pairs	1,118 pairs
DE-EN	418,327 pairs	254,793 pairs	56,734 pairs
RO-DE	49,155 pairs	16,644 pairs	2,298 pairs
RO-LT	19,794 pairs	10,552 pairs	2,488 pairs
LT-LV	10,762 pairs	7,135 pairs	1,810 pairs

therefore propose a different retrieval method as a future work as described in the Section 3.2.3.

3.2.3 Future Work in Wikipedia As discussed in the previous section, the proposed filters disregard a large number of document pairs in Wikipedia. We therefore propose another method to filter out comparable documents by identifying comparability in a finer granularity, i.e. sentence level, which is defined below.

1. First, we crawl documents which have interlanguage links for all the ACCURAT language pairs and disregard those with a size smaller than our specified threshold.
2. Documents from the original source language were then translated into English using available MT system. In our case, we make use of Google Translate.
3. We then split the documents using a simple sentence splitter and filter out sentences which are not useful for the corpus, such as sentences which contain a large number of named entities or numbers.
4. Sentences which passed these filters were then paired to each sentence from the English document.
5. We count similarity score on each of these sentence pairs using Jaccard similarity measure and choose the highest scoring sentence for each of the sentences in the smaller documents.
6. If the Jaccard similarity score is higher than the defined threshold, we save this sentence pair and its score as a possible alignment. Otherwise, we disregard that sentence pair. We repeat step 5 and 6 until all sentences in the smaller documents have been paired to the highest scoring sentence of the other document.
7. We represent the comparability score of these documents by the average of these aligned sentences' scores.

By using these methods, we do not only filter documents based on the topic, but also whether there are any parts in the documents which can be aligned. This information is crucial for the next process: the phrase extraction. Even though the simple similarity scoring does not guarantee a perfect alignment of the sentences, we manage to filter out sentences which do not share any overlapping terminology. This feature works well in identifying similar sentences, unfortunately it also requires all documents to be translated to English using a fast and high quality machine translation, which is mostly not available for under resourced languages.

To avoid this problem, we implement a retrieval method which uses limited linguistic resources. This method is an adaptation of Adafre & Rijke [1] who find parallel sentences in bilingual Wikipedia documents by their anchor text information. This method requires no linguistic resource apart from the information already available in Wikipedia. First, a bilingual lexicon is generated by extracting all Wikipedia titles which are connected by interlanguage links. We then translate all the anchor texts in source language into English using this lexicon. This lexicon will also be used to identify other parts in the documents existing in the lexicon, and translate them to English. Similarity scores are then calculated in the same manner between the English document and its translated source document. When tested in a Dutch-English environment, this method performs with high precision [1]. On the other hand, under resourced languages have considerably fewer documents which may limit the number of anchor texts inside the document. Furthermore, some words in Dutch and English share similar terms, which could assist the retrieval process, while this is not necessarily the case for documents from these under resourced languages.

3.3 Twitter

Another potentially interesting source of comparable documents is Twitter. Twitter allows users to post short text snippets (or tweets) which may contain news events, messages to other users, or comments about particular topics or links. Although the service only started in 2006, the use of Twitter has increased dramatically, with the number of users reaching 190 million in August 2010 [38]. As a resource of comparable documents, Twitter has several advantages. First, the messages are relatively short, which reduces issues surrounding alignment. Each tweet also contains the date of publication, enabling retrieval and alignment to be performed more accurately.

Around 10% of tweets provide url links, which are sometimes accompanied with a short text message describing the link content or comments about it. An example of this is:

"Obama Wants Kids To Make Video Games [Politics]: US President Barack Obama ... http://bit.ly/cPrY4m."

The link refers other users to the complete article of the news. Sometimes users also define their tweets by using # symbols, such as:

“#football HARGREAVES CLOSE TO RETURN - SURGEON: Owen Hargreaves’ surgeon believes the Manchester United star is cl... http://bit.ly/dsf352”

It is possible that the same links or topics are being tweeted in different languages. Our approach to use Twitter is to use URLs and topics as queries to retrieve comparable tweets in different language. We analysed two different retrieval methods: the first method uses English tweet as a seed to retrieve tweets from other languages, while the second method chooses a popular non English tweet and use it as a seed to find a comparable English tweet. The first method performed poorly due to the domination of English tweets in the search result, making it difficult for tweets from other languages to be retrieved. We manage to find interesting results by using the second method; an example of the retrieved tweets is shown in Table 5.

Table 5. Example of Comparable Tweets

Latvian Tweet	
Original Text	Arkartigi interesants raksts par pasaule ietekmigaka tech portala Techcrunch raditaju un ipašnieku http://bit.ly/9nFsGa
Translated by Google	Extremely interesting article on the world’s most powerful tech site Techcrunch creators and owners http://bit.ly/9nFsGa
Other Retrieved English Tweets	
Original Text	Interesting Inc article on Michael Arrington via @marshallk’s podcast. "bust the door down and clean the mess up later" http://bit.ly/9nFsGa
Original Text	really interesting profile in Inc. magazine about Mr. TechCrunch himself - Michael Arrington http://bit.ly/9nFsGa
Original Text	Interesting article by @arrington on how he works. Except you can have more than 2 monitors with a Mac - http://bit.ly/9nFsGa

Our initial analysis found that there exist bilingual comparable tweets. However, there are several problems in retrieving tweets of under resourced languages. First of all, based on [42], over 50% of tweets are in English, while the rest are in Japanese, Malay, Indonesian, and major European languages, such as Spanish, French and German. This means that the use of Eastern European languages in Twitter is relatively unpopular, which limit the retrieved data for ACCURAT project languages. Second, it was more difficult to identify language in Twitter as language identifier does not perform reliably on short texts. We also find that the same URL is mostly tweeted in that particular language only, which again limits the number of bilingual tweets found. Nevertheless, with the popularity of Twitter which is increasing rapidly, there is a possibility that tweets from these countries will also increase in the future. We plan to do further analysis once we have a bigger dataset for these language pairs.

4 Evaluation

At the end of the retrieval process we will have gathered pairs of documents, sentences or fragments, which were considered to be comparable by our retrieval methods. The next step will be to assess the accuracy of these methods to retrieve comparable texts. In this phase, we focus on the comparability level of the retrieved documents rather than the effect of these documents to an MT system. To decrease the judgment effort, we plan to evaluate only a sample of the retrieved text pairs. We describe the methods to choose the text pairs to be judged in Section 4.1. An online tool is under development and will be used by assessors to define comparability level of the chosen text pairs. The online judgment tool is described in Section 4.2 and both the classifier and assessment tool will be used iteratively as described in Section 4.3.

4.1 Classifier

As an evaluation method, we built a classifier which was trained using previously assembled comparable corpora. These Initial Comparable Corpora, later referred to as ICC, contains comparable documents of at least one million words of each language. The search was performed using semi automatic retrieval methods and each of the retrieved document pairs was annotated with its corresponding comparability level: parallel, strongly comparable or weakly comparable. Since the training data does not include non comparable document pairs, we created non comparable pairs by pairing documents from different domain or genre.

4.1.1 Features / Criteria of Comparability We identified various features which are useful in identifying comparability level of a document pair as shown in Table 6. These features are divided into two categories: language-dependent features which need some translation methods or other linguistic knowledge, and language-independent features.

To test the performance of our classifier, we focused on Greek-English (EL-EN) corpora by using Google Translate to translate all the Greek documents into English. Google Translate is not expected to perform well on these under resourced languages and domains, nevertheless it can still be considered of better quality compared to bilingual dictionaries or MT based on our parallel corpora. Thus, the results of our classification can then be considered a realistic upper bound. The extracted language-dependent features include relative word overlap (the number of common unique words in the source and target documents divided by the number of unique words in the source document), relative stem overlap (the same as the relative word overlap apart from the fact that words were first stemmed by Porters Stemmer [39], and cosine similarity in the document vector space [36][37]. In the vector space documents are represented as a t-dimensional vector where t is the number of words (or stems) in the entire corpus of source (English) and translated target (Greek) documents. The vectors can be binary (1 if a word/stem is present in a document and 0 otherwise),

Table 6. Features or criteria of comparability

Language-Dependent Features
1. Word/Stem overlap (relevant to source language)
2. Cosine similarities on word/stem occurrence
3. Cosine similarities on word/stem TF
4. Cosine similarities on stemmed word TF-IDF
5. Cosine similarities on stemmed bi-gram and tri-gram TF
Language-Independent Features
1. Out-link overlap
2. URL character overlap
3. URL number of slash difference
4. Image link word overlap
5. Image link filename overlap

include term (words or stems) frequencies (TF) or term frequencies weighted by the importance of a word in the corpus (TF-IDF). The cosine similarity of bi-gram and tri-gram TF vectors was also computed. The extracted language-independent features include out-link overlap (the number of common out-links in the two documents), image links overlap (the number of common image source URL in the two documents), URL level overlap (the difference of the number of slashes in the URL that correspond to the two documents (remember that documents were crawled from the Web and thus each document corresponds to a URL)) and URL character overlap.

After extracting these features for all the document pairs, we trained a classifier using ECOC (Error Correcting Output Codes) [14] in order to predict the comparability level of newly retrieved document pairs. When evaluated using 5-fold-cross-validation, our classifier showed promising results in identifying comparability levels of the document pairs, with precision scoring above 90% for each comparability level.

We used our classifier to evaluate the comparability levels of the automatically retrieved news corpora mentioned in Section 3.1. The results are found to be promising with most of retrieved documents judged as strongly comparable as shown in Table 7.

4.2 Assessment Tool

To verify the classifier performance, we are currently preparing the assessment tool for the assessors. The retrieval methods described in Section 3 will result in pairs of segments of different granularity. For example, retrieval of news articles using named entities will find documents of the same topic. On the other hand, retrieval of Twitter or other pages using the anchor methods will find comparable fragments or sentences instead. For the sake of simplicity, we refer to these different granularities as segments. Given a pair of segments, assessors will be

Table 7. Predicted Comparability Level of Retrieved Documents

Language	Non Comp.	Weakly C.	Strongly C.	Parallel
EL-EN	2.51%	0.01%	97.24%	0.24%
ET-EN	15.94%	0%	82.84%	1.21%
HR-EN	11.88%	0%	80.15%	7.96%
LV-EN	16.29%	0%	64.64%	19.07%
LT-EN	2.06%	0%	85.20%	12.74%
RO-EN	1.90%	0%	98.06%	0.05%
SL-EN	1.28%	73.38%	14.17%	11.17%

asked a series of questions to determine the comparability level. We will use this data to evaluate the classifier and improve its performance.

4.3 An Iterative Process

Having described the classification/ranking of document pairs by the retrieval techniques and the judging process that will be undertaken by the partners, here we illustrate an iterative process that could bootstrap both the quality of the classifier and the quality of the returned corpus.

- Ranking the documents: The aforementioned classifier, trained over the ICC can be used to classify document pairs returned by the retrieval methods described in previous sections. Each document pair can be then ranked based on the probability of being parallel, strongly comparable, weakly comparable or non comparable, respectively, with the documents with the highest probability of being parallel ranked on the top of the list and the ones with highest probability of being non-comparable ranked at the bottom.
- Sampling in a top-heavy manner: Given that assessing the comparability of all the returned document pairs requires extensive human effort, we propose to only assess the quality of a small sample of document pairs. The sampling method that could be used here is stratified sampling, with documents towards the top of the rankings (i.e. the ones that are more likely to be highly comparable) having higher probabilities of being sampled than the ones towards the bottom of the ranking. Measures, such as accuracy or precision at a certain cut-off along with information retrieval measures such as average precision, normalised discounted cumulative gain etc. can be statistically inferred.
- Re-train the classifier based on the results: Sampled document pairs are given to the human assessors to annotate them with respect to their comparability. After getting back the comparability grades of the deferent judged document pairs we can use these fresh data to re-train the classifier, re-classify/-rank document pairs - presumably with higher accuracy - and eventually bootstrap both the performance of the classifier and the quality of the produced comparable corpus.

– After the iteration process finished, we aim to have a fully trained classifier, which has high precision in predicting comparability levels between documents. This classifier will then be used as the criteria to evaluate all the retrieved comparable documents.

5 Conclusion

In this chapter, we discussed our work on developing retrieval methods to collect comparable documents from Web sources for under resourced languages. We identify three different Web sources and develop the appropriate retrieval methods to gather these comparable documents. News documents are retrieved using date and named entity which was translated using available online resources. We make use of the richness of interlanguage links in Wikipedia to retrieve comparable documents and perform several filters to filter out the irrelevant documents. Our method on retrieving tweets from Twitter involve the use of URL and topic as query. By using these methods, we managed to get a decent size of data from news and Wikipedia. Unfortunately, the use of under resourced languages in Twitter is less popular, and this has caused problems in retrieving high quality data.

We also developed methods to evaluate the retrieved documents by building a classifier to assess comparability levels of these automatically retrieved documents. Extraction of language dependent and language independent features was performed on the Initial Comparable Corpora (ICC) and these data were used as training data for the classifier. When evaluated using 5-fold-cross-validation, the classifier managed to identify comparability levels of document pairs with a high accuracy. We used the classifier to evaluate the news corpora and found that over 85% documents are found to be strongly comparable.

Our future work involves developing reliable independent features to extract comparable segments from Wikipedia and evaluate the results. We plan to compare the quality of retrieved documents found in different Web sources. We are also developing assessment tool to gather judgment from the assessors on a subset of document pairs. The judgment information will be used to evaluate the classifier's performance and improve the training data.

6 Acknowledgement

The project has received funding from the ACCURAT Project, European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no 248347.

References

1. Adafre, S. F. and de Rijke, M. 2006. Finding Similar Sentences across Multiple Languages in Wikipedia. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (2006), pp. 62-69.

2. Appelt, D. An Introduction to Information Extraction. Artificial Intelligence Communications, 12(3):161-172, 1999.
3. Ardo. 2005. Combine web crawler. Software package for general and focused Webcrawling. http://combine.it.lth.se/.
4. Argaw, A. A. and Asker, L. Web Mining for an Amharic - English Bilingual Corpus. In Proceedings of 1st International Conference on Web Information Systems and Technologies (WEBIST 2005), Miami, USA. May 2005.
5. Aslam, J., Pavlu, V. and Yilmaz, E. 2006. A Statistical Method for System Evaluation using Incomplete Judgments. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in Information Retrieval, Association for Computing Machinery, Inc.
6. Baroni, Marco.; Bernardini, Silvia. 2004. Bootstrapping Corpora and Terms from the Web. In Proceedings of LREC 2004.
7. Braschler, M. and Schäuble, P. 1998. Multilingual information retrieval based on document alignment techniques. In: Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries, pp. 183-197.
8. Carterette, B. 2007. Robust test collections for retrieval evaluation. In Proceedings of the 30th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Amsterdam, The Netherlands, July 23 - 27, 2007). SIGIR '07. ACM, New York, NY, 55-62.
9. Carterette, B., Allan, J., and Sitaraman, R. 2006. Minimal test collections for retrieval evaluation. In Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Seattle, Washington, USA, August 06 - 11, 2006). SIGIR '06. ACM, New York, NY, 268-275.
10. Cavnar, W. B., & Trenkle, J. M. 1994. N-gram-based text categorization. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval (pp. 161-175).
11. Chakrabarti, S. 2002. Mining the Web: Discovering Knowledge from Hypertext Data, Science & Technology Books.
12. Cho, J., Garcia-Moline, H., and Page, L. 1998. Efficient crawling through URL ordering. In WWW7: Proceedings of the seventh international conference on World Wide Web (pp. 161-172). Amsterdam: Elsevier Science Publishers B. V.
13. Crammer, K. and Singer, Y. 2002. On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research, 2(5): 265-292.
14. Dietterich, T. and Bakiri, G. 1995. Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research, 2(1): 263-286.
15. Dimalen, D, and Roxas, R. 2007. AutoCor: A Query Based Automatic Acquisition of Corpora of Closely-related Languages. In Proceedings of the 21st PACLIC, pp. 146-154.
16. Do, T., Le, V., Bigi, B., Besacier, L., and Castelli, E. (2009). Mining a comparable text corpus for a Vietnamese-French statistical machine translation system. In Proceedings of the Fourth Workshop on Statistical Machine Translation, pages 165-172. Association for Computational Linguistics.
17. Dorado, I. G. 2008. Focused Crawling: Algorithm Survey and New Approaches with a Manual Analysis.
18. Fung, P. and Cheung, P. 2004. Mining very non-parallel corpora: Parallel sentence and lexicon extraction vie bootstrapping and EM. In EMNLP 2004, pages 57-63.
19. Ghani, R., Jones, R., Mladenic, D. 2005. Building Minority Language Corpora by Learning to Generate Web Search Queries. KAIS Knowledge and Information Systems, 7 (1), 2005.

20. Grishman, R. and Sundheim, B. Message understanding conference - 6: A brief history. In Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, June 1996.
21. Hassan, A., Fahmy, H., & Hassan, H. 2007. Improving named entity translation by exploiting comparable and parallel corpora. In Proceedings of the 2007 Conference on Recent Advances in Natural Language Processing (RANLP), AMML Workshop.
22. Joachim, T. 1999. Making large-scale support vector machine learning practical. Advances in kernel methods: support vector learning, pages 169-184.
23. Koehn, P. 2010. Statistical Machine Translation. Cambridge University Press.
24. Kohlschütter, C., Fankhauser, P. and Nejdl, W. 2010. Boilerplate Detection Using Shallow Text Features. In WSDM '10.
25. Kralisch, Anett; Mandl, Thomas. 2006. Barriers of Information Access across Languages on the Internet: Network and Language Effects. In: Proceedings Hawaii International Conference on System Sciences (HICSS-39) Track 3. p. 54b.
26. Lee, M. and Welsh, M. 2005. An empirical evaluation of models of text document similarity. In COGSCI 2005, pages 1254-1259.
27. Li, Y., McLean, D., Bandar, Z. A., O'Shea J. D. and Crockett, K. 2006. Sentence similarity b ased on semantic nets and corpus statistics. IEEE Trans. On Knowledge and Data Engineering, 18(8): 1138-1150.
28. Menczer, F., Pant, G., and Srinivasan, P. 2004. Topical Web Crawlers: Evaluating Adaptive Algorithms. ACM Trans. On Internet Technology, 4(4): 378-419.
29. Mohammadi, M. and GhasemAghaee, N. 2010. Building Bilingual Parallel Corpora based on Wikipedia. In Proceedings of Second International Conference on Computer Engineering and Applications, volume 2, pp. 264-268.
30. Munteanu, D. and Marcu, D. 2005. Improving Machine Translation Performance by Exploiting Comparable Corpora. Computational Linguistics, 31 (4), pp. 477-504, December.
31. Munteanu, D. S., Fraser, A. and Marcu, D. 2004. Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora. In HLT-NAACL, pages 265-272.
32. Och, F. J. and Ney, H. 2003. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29(1): 19-51.
33. Pinkerton, B. 1994. Finding what people want: Experiences with the Web Crawler. In Proceedings of the First World Wide Web Conference, Geneva, Switzerland.
34. Pouliquen, B, Steinberger, R, Ignat, C, Temnikova, I, Widiger, A, Zaghouani, W and Őiďka, J 2005. Multilingual person name recognition and transliteration. Journal CORELA - Cognition, Représentation, Langage. Numéros spéciaux, Le traitement lexicographique des noms propres.
35. Resnik, P. 1999. Mining the web for bilingual text. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, pages 527-534, Morristown, NJ, USA. Association for Computational Linguistics.
36. Salton, G. and McGill, M. J. 1983. Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, NY, USA.
37. Salton, G., Wong, A., and Yang, C. S. 1975. A vector space model for automatic indexing. Communications of the ACM, 18 (11): 613-620.
38. Schonfeld, Erick. 2010. Costolo: Twitter Now Has 190 Million Users Tweeting 65 Million Times A Day. http://techcrunch.com/2010/06/08/twitter-190-million-users/ Accessed on: 1 September 2010.
39. Sparck-Jones, K. and Willet, P. 1997. Readings in Information Retrieval. Morgan Kauffmann.

40. Steinberger, R., Pouliquen, B., & Ignat, C. (2005). Navigating multilingual news collections using automatically extracted information. Journal of Computing and Information Technology, 13(4), 257-264.
41. Talvensaari, T., Pirkola, A., Järvelin, K., Juhola, M., and Laurikkala, J. 2008. Focused web crawling in the acquisition of comparable corpora. Inf. Retr. 11, 5 (Oct. 2008), 427-445.
42. http://techcrunch.com/2010/02/24/twitter-languages/. Accessed on 1 April 2011.
43. Theobald, M., Siddharth, J. and Paepcke, A. 2008. SpotSigs Robust & Efficient Near Duplicate Detection in Large Web Collections. Stanford University Sigir 2008, Singapore.
44. Tsochantaridis, I., Hoffmann, T., Joachim, T., and Altun, Y. 2004. Support vector machine learning for interdependent and structured output spaces. In ICML '04: Proceedings of the twenty-first international conference on Machine learning, page 104, New York, NY, USA. ACM.
45. Utsuro, T., Horiuchi, T., Chiba, Y., & Hamamoto, T. 2002. Semi-automatic compilation of bilingual lexicon entries from cross-lingually relevant news articles on WWW news sites. In AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users (pp. 165-176). London: Springer-Verlag.
46. Uszkoreit, J., Ponte, J., Popat, A. and Dubiner, M. 2010. Large Scale Parallel Document Mining for Machine Translation. In Proceedings of the 23rd International Conference on Computational Linguistics. COLING 2010, pages 1101-1109, Beijing, August 2010.
47. Wakao, T. and Gaizauskas, R. and Wilks, Y. Evaluation of an Algorithm for the Recognition and Classification of Proper Names. In Proceedings of the 16th International Conference on Computational Linguistics. COLING '96. pages 418-423, 1996.
48. Vapnik, V. N. 1995. The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA.
49. Yilmaz, E., Aslam, J. 2006. Estimating average precision with incomplete and imperfect judgments. In Proceedings of the fifteenth ACM conference on Conference on information and knowledge management. CIKM '06.
50. Yilmaz, E., Kanoulas, E. and Aslam, J. 2008. A Simple and Efficient Sampling Method for Estimating AP and NDCG. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '08.
51. Zhang, Y., K. Wu, J. Gao, and Phil Vines. 2006. Automatic Acquisition of Chinese-English Parallel Corpus from the Web. In Proceedings of 28th European Conference on Information Retrieval. ECIR '06.