Learning Comparable Corpora from Latent Semantic Analysis Simplified Document Space

Ekaterina Stambolieva
euroscript Luxembourg S.à. r.l.
55, rue de Luxembourg, L-8077
Luxembourg
ekaterina.stambolieva@euroscript.lu

Abstract

Focusing on a systematic Latent Semantic Analysis (LSA) and Machine Learning (ML) approach, this research contributes to the development of a methodology for the automatic compilation of comparable collections of documents. Its originality lies within the delineation of relevant comparability characteristics of similar documents in line with an established definition of comparable corpora. These innovative characteristics are used to build a LSA vector-based representation of the texts. In accordance with this new reduced in dimensionality document space, an unsupervised machine learning algorithm gathers similar texts into comparable clusters. On a monolingual collection of less than 100 documents, the proposed approach assigns comparable documents to different comparable corpora with high confidence.

1 Introduction

The problem of collecting comparable corpora is challenging and yet enchanting. Many can benefit from the availability of such corpora as translation professionals, machine learning researchers and computational linguistics specialists. Yet there is not an even consent about the notion covered by the term comparable corpora. The degree of similarity between comparable corpora documents has not been formalized strictly and leaves space for different interpretations of similarity, contributing to abundant text collections of similar and semi-similar documents. The current research endeavors to contribute to an approach, which assembles a collection of comparable documents that are closely related to each other on the basis of a strict definition of comparable corpora. The proposed approach incorporates originally a Latent Semantic Analysis technique in order to match similar concepts instead of words thus contributing to better automatic learning of comparability between documents.

2 Comparable Corpora Definition

Maia (2003) discusses the characteristics of comparable corpora. Nevertheless, the adopted definition of comparable corpora in this study is given by McEnery (2003):

"Comparable corpora are corpora where series of monolingual corpora are collected for a range of languages, preferably using the same sampling and frame and with similar balance and representativeness, to enable the study of those languages in contrast."

McEnery (2003) characterizes comparable corpora as "corpora where series of monolingual corpora are collected for the range of languages". In the views of McEnery (2003), a monolingual corpus is a corpus that is not collected for a range of languages, but instead the documents selected are written in one language. In the context of the current research, a comparable corpus, a sublanguage corpus, can be constructed from documents in one language under the condition they are compliant with the preferred guidelines provided by McEnery (2003). These preferred guidelines are similar sampling frame, balance and representativeness.

A document feature corresponding to text sampling is explicated taking into consideration the domain and genre of the documents. Addi-

Proceedings of the 6th Workshop on Building and Using Comparable Corpora, pages 129–137,
Sofia, Bulgaria, August 8, 2013.

tionally, similar terminology vocabulary insures genre correspondence. Therefore, the same sampling scheme in collecting documents is evaluated considering domain and genre and viewed as document features.

Language is rapidly changing and evolving throughout the years (Crystal 2001). As a result, restricting the time period a document has been published increases the chances of it being comparable to another one written during the same time frame. When events are reported in the newspaper domain, their date of publication is strong similarity evidence and is used as a filter between weakly comparable and non-comparable text articles (Skadiņa et al. 2010a).

The question of how representativeness of a corpus is decided upon is answered in different ways depending on the specific corpus purpose. For the purposes of this research, a corpus is considered representative when corresponding texts are similar in size. As reported by Manning and Schűtze (1999), a balanced corpus is one, which is assembled "as to give each subtype of text a share of the corpus that is proportional to some predetermined criterion of importance".
Skadina et al. (2010b) present a good summary of the advantages of exploiting comparable corpora. It is discussed that "they can draw on much richer, more available and more diverse sources which are produced every day (e.g. multilingual news feeds) and are available on the Web in large quantities for many languages and domains." (Skadina et al. 2010b).

3 Related Work

The most closely-related to machine learning work that mines comparable corpora is that by Sharoff (2010). His research incorporates intelligent self-learning techniques to the compilation of comparable documents. Unlike other researchers that experiment with Cross-Lingual Information Retrieval (CLIR) techniques as in Tao and Zhai (2005), Sharoff (2010) estimates the document collection's internal subgroup system in search for structure. The possible structure and grouping of a set of documents is most easily defined by ranked words that are representative for the subsets in the collection. Sharoff's approach relies heavily on keywords and keyword estimation. One thing Sharoff (2010) does not elaborate on in details is the definition of a comparable corpus. A possible reason for that is that unsupervised machine learning approaches produce related sets of documents in an environment where the selection process is automated and not supervised by any linguistically-dependent rules.

What is written by Goeuriot et al. (2009) is also an influential and relevant material to the current research. Their paper is on the compilation of comparable corpora in a specialized domain with a focus on English and Japanese. The article is significant for the reason the authors investigate ways of building comparable corpora using machine learning classification algorithms, namely Support Vector Machine and C4.5. The experimental setup in the work of Goeuriot et al. (2009) relies on manually labeled data, which is then fed to the machine learning algorithm core. The paper by Goeuriot et al. (2009) is directed towards building a tool to automatically compile comparable corpora in a predefined set of documents and languages. The text comparability characteristics extracted, which allow comparison between the documents, are external and internal to the textual data. Goeuriot et al. (2009) emphasize on selecting ways to automatic recognition of useful features similar texts have and experiment with these features to test and predict their reliability. The comparability of the documents defined by them is on three levels - type of discourse, topic and domain, focusing on locutive, ellocutive and allocutive act labels.

Bekavac et al. (2004) discuss the grounds of a methodology describing similarity comparison of under-resourced monolingual corpora. Contrary to other methodologies that exploit seed words or seed texts as a basis for search, the researchers have at their disposal two monolingual documents sets from which they aim to mine comparable documents. The advantage of their approach is that it is applicable to texts collection written in one language for the reason that they are easily mined and compiled from the available textual resources nowadays. The concept behind their research is to align comparable documents that are found in pre-collected different monolingual corpora. Content features are used to test the degree to which two texts are similar to each other in the sense of sharing the same information and common words. These features, composition features, need to be representative for the texts. The composition features, extracted from the data, monitor the size, the format and the time span of the documents.

Clustering based on semantic keyword extraction is performed by Finkelstein et al. (2001). This approach is relevant to the current research as it suggests a different methodology of feeding texts to machine learning algorithms. The re-

searchers aim to generate new content based on input user queries by using context – "a body of words surrounding a user-selected phrase" (Finkelstein et al. 2001). They emphasise on the significance of using context when developing Natural Language Processing (NLP) applications. The keyword extraction algorithm presented relies on a precisely-designed clustering algorithm, different than k-means, to recursively clean clustering results and present refined statistical output.

With regards to evaluation metrics of comparable corpora, one of the main focuses of the ACCURAT Project (Skadina et al. 2010b) is to design metrics of comparability estimation between texts. The ACCURAT researchers (Skadina et al. 2010b) concentrate on the development of comparable corpora criteria for different texts and different types of parallelism between the texts. Saralegi et al. (2008) suggest measures based on distribution of topics or time with regards to publication dates. Kilgariff (2001) aims to measure the level of comparability between two collections of documents. He focuses additionally on the shortcoming of known corpus similarity metrics. He discusses evaluation methods for corpus comparability measures, which are based on Spearman rank correlation coefficient, perplexity and cross-entropy, χ^2 and others. To his knowledge, the χ^2 test performs the best when comparing two sets of documents. It is important to note that the approach adopted by Kilgariff (2001) relies on words and n-gram sequence features. Not only does he regard the texts as bag-of-words, but also he incorporates n-gram characteristics in his evaluation metric analysis.

Mining word similarity techniques are discussed in the work of Deerwester et al. (1990); Baeza-Yates and Ribeiro-Netto (1999); and Dagan, Lee and Pereira (1999). Deerwester et al. (1990) incorporate LSA as a technique to identify word relatedness. LSA "identifies a number of most prominent dimensions in the data, which are assumed to correspond to 'latent concepts'." (Radinsky et al. 2011). Radinsky et al. (2011) indicate that LSA vector space models are "difficult to interpret". Consequently, the current research focuses not only on the incorporation of LSA to mapping content, but also of the employment of a machine learning technique to group projected into the two-dimensional space documents into similar clusters. Baeza-Yates and Ribeiro-Netto (1999), as Sharoff (2009) and Goeuriot et al. (2010), consider texts as bag-of-words as the least complex word similarity approaches can be incorporated. Mapping distributional similarity, Lee (1999) opts for similar word co-occurrence probability estimation improvement. Dagan et al. (1999) also aim for better estimation of word co-occurrence likelihood not based on empirical methods, but instead relying on distributional similarity for the generation of language models. WordNet-based and distributional-similarity comparisons of word similarity are presented in Agirre et al. (2009). They suggest different views of word relatedness comparison – bag-of-words, context windows and syntactic dependency approaches. They describe their findings as yielding best results on known test sets. What is important to be remarked is that their methodology requires minor fine-tuning in order to give good results on cross-lingual word similarity.

4 Methodology

The novelty of our approach is the incorporation of the Latent Semantic Analysis technique, which matches concepts, or information units, from one document to another instead of approximating word similarity. LSA expects and constructs a new vector-based representation of the documents to be compared. A concept holds not only textual, but also morphological information about each word present in the texts. By employing LSA, the document space is projected into the two-dimensional space in correspondence with the latent relationships between the words in the texts. In the two-dimensional space, clusters of similar documents are compiled together using a simple, but powerful unsupervised machine learning algorithm, k-means clustering. Clustering evaluation metrics such as precision, recall and purity are employed towards automatic evaluation and analysis of the resulting comparable corpora.

In order to compile comparable corpora with the current settings, a set of pre-collected documents is needed. From this set of documents, two to five comparable corpora are identified and texts with similar topics, domains and features are assigned to relevant comparable corpora.

LSA has its known limitations. It acknowledges documents as bags-of-words and mines

the latent relationships between the words in the bags-of-words. Working with information units overcomes this limitation of LSA. The information units contain additional linguistic information about the syntactic and morphological relationships between words, therefore forming concepts of these words. The order of the words, or the information units, is not imperative, therefore it is not controlled by the methodology.

LSA allows words to have only one meaning thus restricting the robustness of the natural languages. This limitation is tackled by suggesting different word sense candidates for words and constructing a separate information unit for each promoted word sense.

5 Data Feature Selection

The innovation of the discussed research approach lays in its basic concept of perceiving texts as bags of interrelated concepts. The surface-form words found in the texts are enriched with linguistic information that furnishes better matching procedure of the concepts lying within the texts for comparison.

Unlike previous work, which regards documents as bags-of-words (Sharoff 2009, Goeuriot et al. 2010) the methodology treats documents as collections of concepts, each concept containing comparable textual information. The concepts are represented by information units. The process of recognizing such units happens at document level, where each document is viewed as a separate text with its own context. Each information unit is defined as the inseparable pair of lemma and its context-dependent part-of-speech (POS) tag. A lemmatization technique is applied to transform the texts into linguistically-simplified versions of the originals, where each word (infected or not) is substituted by its corresponding lexeme.

As stated before, the information units incorporate POS output. A POS tagger is used to process the texts before linguistically-simplifying it using lemmatization techniques. The idea of enriching the words by POS information is not new to the research of Natural Language Processing, but it is new for the research of compiling comparable corpora. By identifying the POS information of a sentence, lexical ambiguity is reduced. The accompanying POS tag to each lemma assists the disambiguation of the information units. For example, *run* as being the action of walking fast has a verb POS tag opposed to *run* as the period of some event happening has a noun POS tag. In this example, the POS tag provides the needed information for disambiguating the two different meanings of a word. In the current research scenario, the POS tagging module[1] emulates the results of a basic Word Sense Disambiguation technique.

Furthermore, the input set of documents is transformed into a set of lists of information units as described, where a single list of units corresponds to a single document. When compared, the units are matched for correspondence both based on the lemma's lexical category in the sentence and its base form.

Another feature, which helps build context related concepts, is the identification of Noun Phrases (NP) in the texts. Noun Phrase recognition is imperative since it further develops the simple word sense disambiguation method. Some words to have a different meaning when occurring in a chain of words such as a noun phrase. Unlike the proposed by Su and Babych (2012) approach to NP recognition, NPs are identified following linguistically-derived rules, which represent common constructions of the language under consideration. When a NP is identified, it is listed as a new information unit with a corresponding *NP* POS tag. All POS annotations as well as lemma information of its constituent words are removed from the documents' list of information units.

6 Experiments

6.1 Experimental Corpus

A pre-collected corpus of documents, part of the NPs for Events (NP4E) corpus (Hasler et al. 2006), is used for experimenting. The NP4E corpus is collected for the special purpose of extracting coreference resolution in English. Nevertheless, the structure and the organization of the corpus are suitable for the needs of acquisition of a test corpus for the current study. The NP4E corpus contains five different groups of news articles based on topic gathered from the Reuters. The news articles are collected in the time frame

[1] TreeTagger http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

of two years – 1996 and 1997 (Rose, Stevenson and Whitehead 2002). Four of the five NP4E news article groups are used to compile an experimental corpus containing roughly 40000 words or 520 words per text. The chosen experimental collection consists of sub-corpora that have documents comparable to the others in their sub-corpora based on domain. The domain of these comparable corpora is terrorism, and the four distinct topics are connected with terrorism, bombing and suicide respectively for events in Israel, Tajikistan, China and Peru. In total, the experimental corpus consists of 77 newswire articles. The distribution of the documents in this selected corpus is 20 on Israel topic, 19 on Tajikistan topic, 19 for China topic and 19 on Peru topic. These sub-corpora are referred to as Israel (I), Tajikistan (T), China (C) and Peru (P) onwards.

6.2 Experimental Set-up

The experimental set-up is structured as a chain of two simple procedures. They are respectively an experimental setup data selection and experimental setup clustering distribution.

6.2.1 Data Selection Frame

The data selection frame describes how document features are selected. The documents are afterwards preprocessed in order to extract all underlying text features and binary vectors are constructed to represent each separate document. The document features on focus consist of all identified information units enriched with the noun phrases that were recognized in the texts. The binary vectors then are used as an input to the LSA algorithm.

6.2.2 Cluster Distribution

The number of resulting clusters, or comparable corpora, should be set in advance for unsupervised machine learning algorithms. An experiments with *k*, *k* is in the range of 2 to 5, are conducted. Testing with number of clusters greater or equal to two comes logical. In the case of expecting two resulting clusters, the methodology groups all similar documents in one comparable corpus, and withdraws the non-similar documents to the second collection. When *k* is chosen to be 2 or 3, the resulting comparable corpora tend to be weakly-comparable (Skadiņa et al. 2010a) for the reason the algorithms are forced to gather documents with four distinct topics into only two or three comparable collections. It is interesting to analyze the research methodology's performance in the case four output comparable corpora are expected, meaning when the learning algorithm is asked to suggest four comparable sets of documents.

To evaluate clustering performance in terms of forcing the system to split the document collection into more comparable corpora than present, *k* equals to 5 is also used in the experiments. Consequently, the number of clusters varies between 2 and 5.

6.2.3 Evaluation Metrics

Three metrics are chosen to evaluate results - the standard precision and recall, and additionally - purity. Precision shows how many documents in the resulting collections are identified correctly as comparable to the majority of documents on a specific topic in the cluster. For example, when 16 out of 19 documents are recognized to be comparable to each other, the precision of this clustering result is 0.84. Recall shows how many false negatives are identified as comparable to a certain topic-related collection of texts. The false negatives are the documents on a different topic, which the machine learning algorithm falsely lists to be comparable to documents on another topic. When 21 documents are grouped in one similarity cluster, 19 of them being on a related topic, 3 of them being on another topic, the recall of the learning performance is 0.86.

Purity is an evaluation metric used to estimate the purity of the resulting clusters (Figure 1.). A cluster is recognized as pure when it contains a number of documents with the same label (meaning they are listed to be comparable to each other by a human evaluator) and as less as possible documents that have a different label from the dominant label (Manning et al. 2008):

$$Purity = \frac{(nom_{cluster\ 1} + .. + nom_{cluster\ k})}{no_{clusters}}$$

Figure 1. Purity score formula

where $nom_{cluster\ i}$ is the number of the majority class members in each resulting cluster *i*, and $no_{clustrers}$ is the number of resulting clusters, or *k*. As Manning et al. (2008) warn "High purity is easy to achieve when the number of clusters is large - in particular, purity is 1 if each document gets its own cluster". The number of clusters for the current research is not big. Nonetheless, the results are evaluated based on two other metrics.

The other metrics for measuring the comparability between documents that are chosen for exploitation in the current research, are Mutual Infromation (MI) and Normalized Mutual Infroamtion (NMI). The formula for NMI is as follows and shown in Figure 2.:

$$NMI(\Omega, C) = \frac{MI(\Omega, C)}{(H(\Omega) + H(C))/2}$$

Figure 2. NMI score formula

MI is explained in details in Kalgariff (2001) and (Manning et al. 2008). Manning et al. (2008) discuss additionally the formula for the entropy H, and NMI. Ω is the group of clusters addressed in the experiments, and *C* is the group of labels – namely the different characteristics of the comparable corpora.

In the current scenario, no human evaluation is performed. Rather than that the corpus is predesigned in a way to contain four different comparable corpora that need not to be manually labeled

6.3 Evaluation

Results are obtained after conducting different set-up experiments. One set-up focuses on evaluating comparable corpus collection having as an input part of the experimental corpus. This part contains documents on two out of the four different topics. The two-topic collections are compiled by combining all combinations possible of two topic-based sets together from the four distinct topic sub-corpora. In this experimental scenario, the total of different corpora for evaluation is 6 (according to the combination's formula $\binom{4}{2}$) - Peru and China, Peru and Tajikistan, Peru and Israel, Tajikistan and China, Tajikistan and Israel, China and Israel. Table 1 shows the results of running LSA with k-means clustering on the discussed sub-groups. As seen on Table 1. the learning algorithm performance is excellent when the number of comparable corpora that are expected is greater than two. When three or more comparable clusters are elected, each similar by topic document is grouped with all other documents that are comparable to it in the same resulting comparable corpus. In the case of expecting three comparable corpora with Precision and Recall equal to 1.0, one of these corpora contains all documents of two different sub-corpora and the rest contain all documents of one of the predefined experimental sub-corpora. In the case of expecting five comparable corpora with Precision and Recall equal to 1.0, one sub-corpus is split into two comparable clusters, these clusters containing documents on the same topic. What is interesting in this experimental set-up are the results the learning algorithm obtains when it aims to produce only two comparable clusters. For three of the test sets - China and Israel, Peru and China and Tajikistan and Israel, grouping of documents on different topics into the same similar collection is seen. The lowest results obtained are for the test set Tajikistan and Israel, where 3 of the 19 documents on an Israel topic are grouped together with the texts on the Tajikistan topic. The reason behind this automatic learning confusion originates from the fact the Tajikistan and Israel topic documents contain many similar concepts, which make good clustering harder to achieve.

The purity of the resulting corpora is very high, above 0.9, indicating that comparable documents are identified correctly with high relevance. The only exception is the results on the Tajikistan and Israel test set with purity 0.56. This exception occurs because of poor clustering results, which have been discussed.

Sub-corpus	Topic	Precision				Recall				Purity
		2Cl	**3Cl**	**4Cl**	**5Cl**	**2Cl**	**3Cl**	**4Cl**	**5Cl**	
P	Peru	0.84	1	1	1	1	1	1	1	0.921
C	China	1	1	1	1	0.86	1	1	1	
P	Peru	0.84	1	1	1	1	1	1	1	0.921
T	Tajikistan	1	1	1	1	0.86	1	1	1	
P	Peru	1	1	1	1	1	1	1	1	1.00
I	Israel	1	1	1	1	1	1	1	1	
T	Tajikistan	1	1	1	1	1	1	1	1	1.00
C	China	1	1	1	1	1	1	1	1	
T	Tajikistan	1	1	1	1	0.52	1	1	1	0.56
I	Israel	0.15	1	1	1	1	1	1	1	
C	China	0.86	1	1	1	1	1	1	1	0.923
I	Israel	1	1	1	1	0.85	1	1	1	

Table 1. Clustering results for test sets of combinations of two topic sub-corpora (*n*Cl pointing to the numbers of clusters identified)

Another set-up focuses on the analysis and evaluation of the results on clusters containing documents on three of the four different topics. The same way as the two-topic collections are constructed, combining three topic sub-corpora into one results in the development of the input for the LSA and k-means clustering algorithms. In this experimental scenario, a total of 4 distinct input collections are compiled -Tajikistan, Israel and China; Tajikistan, Israel and Peru; Peru, China and Israel; and Tajikistan, China and Peru.

The results of the learning comparable corpora from them are listed in Table 2. As it can be easily seen, the clustering performance is impeccable. Therefore, providing more documents, more data features, helps identifying better similar documents applying the proposed research approach.

Table 3. Shows the clustering results when all texts of the experimental corpus are suggested as an input. The algorithms once more do not have problems collecting the similar documents into comparable corpora with high precision and recall.

MI and NMI are computed only for the results presented in Table 1. The reasoning behind is that Table 2. And Table 3. show perfect clustering results of comparable corpora obtained on the whole set of input documents described in Section 6.1.

The results of the comparable texts grouping are estimated using a clustering quality trade-off metric, NMI. Table 4. shows the NMI results of the clustering performance on the two-topic collections described in the first experimental set-up at the beginning of Section 6.3.

Sub-corpus	Topic	Precision				Recall				Purity
		2Cl	**3Cl**	**4CL**	**5Cl**	**2Cl**	**3Cl**	**4Cl**	**5Cl**	
T	Tajikistan	1	1	1	1	1	1	1	1	
I	Israel	1	1	1	1	1	1	1	1	1.00
C	China	1	1	1	1	1	1	1	1	
T	Tajikistan	1	1	1	1	1	1	1	1	
I	Israel	1	1	1	1	1	1	1	1	1.00
P	Peru	1	1	1	1	1	1	1	1	
P	Peru	1	1	1	1	1	1	1	1	
C	China	1	1	1	1	1	1	1	1	1.00
I	Israel	1	1	1	1	1	1	1	1	
T	Tajikistan	1	1	1	1	1	1	1	1	
C	China	1	1	1	1	1	1	1	1	1.00
P	Peru	1	1	1	1	1	1	1	1	

Table 2. Clustering results for test sets of combinations of three topic sub-corpora

	Precision				Recall				Purity
	2cl	**3cl**	**4cL**	**5cl**	**2cl**	**3cl**	**4cl**	**5cl**	
T	1	1	1	1	1	1	1	1	
C	1	1	1	1	1	1	1	1	1.00
I	1	1	1	1	1	1	1	1	
P	1	1	1	1	1	1	1	1	

Table 3. Clustering results on the whole experimental corpus

	Mutual Information	**H(Ω)**	**H(C)**	**NMI**
	2CL	**2CL**	**2CL**	**2Cl**
Peru China	0.6866	0.9927	1	0.6916
Peru Tajikistan	0.6866	0.9927	1	0.6916
Peru Israel	1.0230	1.0074	1.0074	0.9522
Tajikistan China	1	1	1	1
Tajikistan Israel	0.0844	0.3912	1.0074	0.1262
China Israel	0.6855	0.9744	1.0074	0.6917

Table 4. MI and NMI scores results for test sets of combinations of two topic sub-corpora

Consequently, the results shown on Table 4. are obtained with respects to the precision, recall and purity scores presented in Table 1. The NMI score is evidence of the identified comparable corpora quality. As seen on Table 4., the lowest NMI score correspond to the clustering results on the Peru- and China- topic texts. As shown on Table 1., the proposed approach is not confident when grouping the Peru- and China- topic texts into comparable collections. The results of the NMI metric shown on Table 4. only confirm this conclusion. The best results obtained according to the NMI score are NMI is dependent on the mutual information and the entropy the texts to be clustered share. MI is a metric, which estimates how the amount of information presented in the documents affect the clustering output. When the MI score is low, as in the example of grouping the Tajikistan- and Israel- topic texts, the information contained in the documents does not contribute to highly-comparable clusters of corpora. When the MI score obtained is high, as

in the Tajikistan- and China- topic documents experiment, the information in these documents is a strong evidence of the text relatedness. Table 4. lists the intermediate calculations of the entropy based on the available labels H(C) and the resulting clusters H(Ω).

7 Remarks

The problems identified in the current methodology are classified into two different groups: text processing resources errors and clustering output errors. The processing resources are taken as off-the-shelf modules and the development focus of the study in not concentrating on improving their performance. The second type of errors is the clustering errors. Their size can be reduced by improving the performance of the text preprocessing resources. Additionally, enhanced clustering output evaluation metrics can reveal learning algorithm's weaknesses and suggest ways for improvement.

8 Future Work

More can be done in the future to improve the proposed methodology. One idea for further investigation is experimenting with larger collections of data. The results on the experimental corpus are promising, but the document collection is not big and contains less than 80 texts. It would be interesting to experiment with corpora that consist of hundreds of documents to test clustering performance. Additionally, a new experimental collection of documents is being compiled. It contains psycholinguistics texts both in Spanish and English. As the collection of this document set is still in progress, the results obtained on it are not presented in the current paper. These results will be reported in future work publications.

Furthermore, a new translation equivalent source can be added. In the case of compiling specialized collections of comparable documents, a specialized bilingual or multilingual dictionary can prove to be a valuable resource. An untested interesting experimental setup can be investigating the resulting clustering performance when more than 50% or more of the most relevant lemmas (with noun phrases) are selected as document features. A Named Entity Recognizer (NER) and a synonymy suggestion module have the possibility to serve as good text processing resources and further improve grouping outcomes. In connection with NER, it is interesting additionally to investigate if the test corpus contains local names, which make clustering better easier. Lastly, potential source for further development is the automatic recognition of diasystematic text features, such as diachronic, diatopic or diatechnic information.

Clustering results of comparable corpora are obtained when the document characteristics are filtered by best keyword estimation metric - TF.BM25, explained in Pérez-Iglesias et al. (2009). The results show decrease in good clustering performance. A future work aspect is to investigate the cause this lower performance.

9 Conclusion

An innovative approach to the problem of compilation of comparable corpora is described. The approach suggests guidelines to textual characteristics selection scheme. Additionally, the approach incorporates LSA and unsupervised ML techniques. Different evaluation metrics, such as precision, purity and normalized mutual information, are employed to estimate comparable corpus clustering results. These metrics show good results when evaluating comparable clusters from a predefined set of less than 100 documents. The methodology suggested is applied for monolingual selection of documents; nonetheless it is readily extendable to more languages.

References

Agirre, Eneko, Alfonseca, Enrique, Hall, Keith, Kravalova, Jana, Paşca, Marius and Soroa, Aitor.2009. A study of Similarity and Relatedness Using Distributional and WordNet-based approaches. In *NAACL '09*, pages 19-27.

Baeza-Yates, Ricardo and Ribeiro-Neto, Betrhier. 1999. *Modern Infromation Retieval*, Addison Wesley.

Bekavac, Božo, Osenova, Petya, Simov, Kiril and Tadic, Marco. 2004. Making Monolingual Corpora Comparable: a Case Study of Bulgarian and Croatian. In *Proceedings of LREC2004*, pages 1187-1190, Lisbon.

Crystal, David. 2001. *Language and the Internet*. Cambidge University, Press. Cambidge.UK, pages 91-93.

Dagan, Igo, Lee, Lillian and Pereira, Fernando. 1999. Similarity-based models of word co-occurrence probabilities. *Machine Learning*. 34(1-3), pages 43-69.

Deerwester, Scott, Dumais, Susan, Furnas, George, Landauer, Thomas and Harshman, Richard. 1990. Indexing by latent semantic analysis. *Journal of*

the Americal Society for Information Science. 41(6), pages 391-407.

Finkelstein, Lev, Gabrilovich, Evgeniy, Matias, Yossi, Rivlin, Ehud, Solan, Zach, Wolfman, Gadi and Ruppin, Eytan. 2001. Placing Search in Context: The Concept Revisited. In *WWW'01*, pages 406-414.

Goeuriot, Lorraine, Emmanuel Morin and Béatrice Daille. 2009. Compilation of specialized comparable corpora in French and Japanese. In *Proceedings of the 2nd workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, August 06, 2009, Suntec, Singapore.

Hasler, Laura, Constantin Orasan and Karin Naumann. 2006. NPs for Events: Experiments in Conference Annotation. In *Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation (LREC2006),*pages 1167-1172, 24-26 May 2006, Genoa, Italy.

Ion, Radu, Dan Tufiş, Tiberiu Boroş, Ru Ceauşu and Dan Ştefănescu. 2010. On-line Compilation of Comparable Corpora and Their Evaluation. In *Proceedingds of the 7th International Conference of Formal Approaches to South Slavic and Balkan Languages (FASSBL7),* pages 29-34, Dubrovnic, Croatia.

Kilgarriff, Adam. 2001. Comparing corpora. *International Journal of Corpus Lingusitics*, 6(1), pages 97-133.

Lee, Lillian. 1999. Measures of distributional similarity. *Proceedings of ACL 1999*, pages 25-32.

Maia, Belinda. 2003. What are Comparable Corpora? *Electronic resource:* http://web.letras.up.pt/bhsmaia/belinda/pubs/CL2003%20workshop.doc.

Manning, Christopher D. and Hinrich Schűtze. 1999. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Manning, Christopher D., Prabhakan Raghavan, and Hinrich Schűtze. 2008. *Introduction to Information Retrieval*, Cambridge University Press, pages 356-358.

McEnery, Tony. 2003. Corpus Linguistics. In Ruslan Mitkov, editor, *The Handbook of Computational Linguistics*. Oxford University Press, Oxford, UK, pages 448-464.

Radinsky, Kira, Agichtein, Eugene, Gabrilovich, Evgeniy and Markovitch, Shaul. 2011. A word at a time: Computing Word Relatedness using Temporal Semantic Analysis. In *WWW'11*, pages 337-346.

Pérez-Iglesias, Joaquín, Pérez-Agüera, José, Fresno, Víctor and Feinstein, Yuval. 2009. Integrating the probabilistic model BM25/BM25F into Lucene. In *CoRR, abs/0911.5046*.

Rose, Tony, Mark Stevenson and Miles Whitehead. 2002. The Reuters Corpus Volume 1 – from Yesterday's News to Tomorrow's Language Resource. In *Proceedings of LREC2002*, pages 827-833.

Saragelı, Xabier., San Vincente, Inaki, Gurrutxaga. Antton 2002.Automatic Extraction of bilingual terms from comparable corpora in a popular science domain. In *Proceedings of the workshop on Comparable Corpora, LREC'08*.

Sharoff, Serge. 2010. Analysing similarities and differences between corpora. In *Proceedings of the 7th Conference of Language Technologies (Jezikovne Tehnologije),* pages 5-11, Ljubljiana. Slovenia.

Skadiņa, Inguna, Ahmet Aker, Voula Giouli, Dan Tufis, Robert Gaizauskas, Madara Mieriņa and Nikos Mastropavlos. 2010a. A Collection of Comparable Corpora for Under-Resourced Languages. In Inguna Skadiņa and Dan Tufis, editors, *Human Language Technologies. The Baltic Perspective. Proceedings of the 4th International Conference Baltic HLT 2010*, pages 161-168.

Skadiņa, Inguna, Vasiljeiv, Andrejs, Skadiņš, Raivis, Gaizauskas, Robert, Tufiş, Dan and Gornostay, Tatiana. 2010b. Analysis and Evaluation of Compoarable Corpora for Under Resourced Areas of Machine Translation. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora. Applications of Parallel and Comparable Corpora in Natural Language PEngineering and the Humanities*, pages 6-14.

Su, Fangzhoung and Bogdan Babych. 2012. Measuring Comparability of Documents in Non-Parallel Corpora for Efficient Extraction of (Semi-)Parallel Translation Equivalents. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra),* pages 10-19, Avignon, France.

Tao, Tao and Cheng Xiang Zhai. 2005. Mining Comparable Bilingual Text Corpora for Cross-Language Information Integration. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining,* pages 691-696.

Finding More Bilingual Webpages with High Credibility via Link Analysis

Chengzhi Zhang*
Nanjing University of Science and Technology
Nanjing, China

Xuchen Yao†
Johns Hopkins University
Baltimore, MD, USA

Chunyu Kit
City University of Hong Kong, Hong Kong SAR, China

Abstract

This paper presents an efficient approach to finding more bilingual webpage pairs with high credibility via link analysis, using little prior knowledge or heuristics. It extends from a previous algorithm that takes the number of bilingual URL pairs that a key (i.e., a URL pairing pattern) can match as the objective function to search for the best set of keys yielding the greatest number of webpage pairs within targeted bilingual websites. Enhanced algorithms are proposed to match more bilingual webpages following the credibility based on statistical analysis of the link relationship of the seed websites available. With about 12,800 seed websites as test set, the enhanced algorithms improve precision over baseline by more than 5%, from 94.06% to 99.40%, and hence find above 20% more true bilingual URL pairs, illustrating that significantly more bilingual webpages with high credibility can be mined with the help of the link analysis.

1 Introduction

Parallel corpora of bilingual text (bitext) are indispensable language resources for many data-driven tasks of natural language processing, such as statistical machine translation (Brown et al., 1990), cross-language information retrieval (Davis and Dunning, 1995; Oard, 1997), and bilingual lexical acquisition (Gale and Church, 1991; Melamed, 1997; Jiang et al., 2009), to name but a few. A general way to develop such corpora from web texts starts from exploring the structure of known bilingual websites, which are usually organized by their web masters in a way to facilitate both navigation and maintenance (Nie, 2010). The most common strategy is to create a parallel structure in terms of URL hierarchies, exploiting some known naming conventions for webpages of corresponding languages (Huang and Tilley, 2001; Nie, 2010). Following available structures and naming conventions, researchers have been exploring various means to mine parallel corpora from the web and a good number of such systems have demonstrated the feasibility and practicality in automatic acquisition of parallel corpora from bilingual and/or multilingual web sites, e.g., STRAND (Resnik, 1998; Resnik, 1999; Resnik and Smith, 2003), BITS (Ma and Liberman, 1999), PTMiner (Chen and Nie, 2000), PTI (Chen et al., 2004), WPDE (Zhang et al., 2006), the DOM tree alignment model (Shi et al., 2006), PagePairGetter (YE et al., 2008) and Bitextor (Esplà-Gomis and Forcada, 2010).

Most of these systems are run in three steps: first, bilingual websites are identified and crawled; second, pairs of parallel webpages are extracted; and finally, the extracted pairs are validated (Kit and Ng, 2007). Among them, prior knowledge about parallel webpages, mostly in the form of ad hoc heuristics for identifying webpage languages or pre-defined patterns for matching or computing similarity between webpages, is commonly used for webpage pair extraction (Chen and Nie, 2000; Resnik and Smith, 2003; Zhang et al., 2006; Shi et al., 2006; Yulia and Shuly, 2010; Tomás et al., 2008). Specifically, these systems exploit search engines and heuristics across webpage anchors to locate candidate bilingual websites and then identify webpage pairs based on pre-defined URL matching patterns. However, ad hoc heuristics cannot exhaust all possible patterns. Many webpages do not even have any language label in their anchors, not to mention many untrustworthy labels. Also, using a limited set of pre-

*Performed while a research associate at City University of Hong Kong.

†Performed while a visiting student at City University of Hong Kong.

Proceedings of the 6th Workshop on Building and Using Comparable Corpora, pages 138–143,
Sofia, Bulgaria, August 8, 2013.