Hierarchical Agglomerative Clustering of English-Bulgarian Parallel Corpora

Rayner Alfred, Dimitar Kazakov, Mark Bartlett
Computer Science Department
University of York, York, UK
{ralfred,kazakov,bartlett}@cs.york.ac.uk

Elena Paskaleva
Bulgarian Academy of Sciences
Sofia, Bulgaria
hellen@lml.bas.bg

Abstract

Most multilingual parallel corpora have become an essential resource for work in multilingual natural language processing. In this article, we report on our work using the hierarchical agglomerative clustering (HAC) technique to cluster multilingual parallel text on web contents. A clustering algorithm taking constraints from parallel corpora potentially has several attractive features. Firstly, training samples in another language provide indirect evidence for a classification or clustering result. Secondly, constraints from both languages may help to eliminate some biased language-specific usages, resulting in classes of better quality. Finally, the alignment between pairs of clustered documents can be used to extract words from each language, which may then be used for other applications, as an example in this paper, we utilise these words for term reduction. We explain the findings that we obtain from the clustering of a significant parallel corpus for a low-density and high-density of paired language, English and Bulgarian. Preliminary results show that the HAC algorithm can effectively cluster bilingual parallel corpora separately and still produce the same extracted words that best describe these clusters for both English and Bulgarian corpora.

Keywords

Multilingual NLP, Evaluation, Corpus-based language processing, Bilingual parallel clustering

1. Introduction

Effective and efficient document clustering algorithms play an important role in providing intuitive navigation and browsing mechanisms by categorizing large amounts of information into a small number of meaningful clusters. In particular, clustering algorithms that build illustrative and meaningful hierarchies out of large document collections are ideal tools for their interactive visualization and exploration, as they provide data-views that are consistent, predictable and at different levels of granularity.

Research in clustering text documents has been performed intensively. However, there are few experiments that examine the impacts of clustering bilingual parallel corpora, possibly due to the problem of the availability of large corpora in translation, i.e. parallel corpora. Fortunately, we have obtained a large collection of over 20,000 bilingual parallel corpora of English-Bulgarian. Compared to a clustering algorithm based on a single language, a clustering algorithm taking constraints from parallel corpora potentially has several attractive advantages. Firstly, training samples in another language provide indirect evidence to verify a classification. Secondly, constraints from both languages may help to eliminate some biased language-specific usages (such as particular homonyms), resulting in classes of better quality. Finally, the alignment between pairs of clustered documents can be used to extract words from each language and can further be used for other applications, such as CLIR [5].

The aim of the experiments presented in this paper is to investigate the effect of applying a clustering technique to parallel multilingual texts. Specifically, the aim is to introduce the tools necessary for this task and highlight preliminary experimental results and problems which have become apparent. In this experiment, it is interesting to look at the differences of the tree structure derived from clustering English texts and Bulgarian texts. In this paper, we provide the results of clustering parallel corpora of English-Bulgarian texts. In addition, we also look at the similarities and differences of three main areas; English-Bulgarian cluster mappings, English-Bulgarian tree structures and the extracted terms for English-Bulgarian clusters. Additionally, the effect of term reduction on the cluster mappings is examined.

Chapter 2 covers some of the background about the vector space model representation of documents and the hierarchical agglomerative clustering method. Chapter 3 explains the experimental design set-up and the experimental results are outlined in chapter 4. Chapter 5 concludes this paper by suggesting what can be done to improve the hierarchical agglomerative clustering of bilingual parallel corpora of English-Bulgarian.

2. Background

2.1 Vector Space Model Representation

In this experiment, we use the vector space model [2], in which a document is represented as a vector in n-dimensional space (where n is the number of different words in the collection). Here, documents are categorized by the words they contain and their frequency. Before obtaining the weights for all the terms extracted from these documents, stemming and stopword removal is performed. Stopword removal eliminates unwanted terms (e.g., those from the closed vocabulary) and thus reduces the number of dimensions in the term-space. Once these two steps are completed, the frequency of each term across the corpus is counted and weighted using *term frequency – inverse document frequency* (tf-idf) [2], as described in equation (1).

$$\text{tf-idf} = \text{tf}(t,d) \cdot \text{idf}(t) \quad (1)$$

$$\text{idf}(t) = \log\left(\frac{|D|}{\text{df}(t)}\right) \quad (2)$$

$$\text{sim}(di, dj) = \left(d_i d_j\right) \Big/ \left(\|d_i\| \|d_j\|\right) \quad (3)$$

$$\text{Precision P}(C,L) = \frac{|C \cap L|}{|C|}, \; C \in C_{ALL}, \; L \in L_{ALL} \quad (4)$$

$$\text{Purity} = \sum_{C \in C_{ALL}} \frac{|C|}{|D|} \cdot P(C,L) \quad (5)$$

$$\text{Precision (EBM)} = \frac{|C(E) \cap C(B)|}{|C(E)|} \quad (6)$$

$$\text{Precision (BEM)} = \frac{|C(B) \cap C(E)|}{|C(B)|} \quad (7)$$

Weights are assigned to give an indication of the importance of a word in characterizing a document as distinct from the rest of the corpus. In summary, each document is viewed as a vector whose dimensions correspond to words or terms extracted from the document. The component magnitudes of the vector are the tf-idf weights of the terms. In this model, tf-idf, as described in equation (1), is the product of term frequency tf(t,d), which is the number of times term t occurs in document d, and the inverse document frequency, equation (2), where |D| is the number of documents in the complete collection and df(t) is the number of documents in which term t occurs at least once. To account for documents of different lengths, the length of each document vector is normalized so that it is of unit length [4].

2.2 Hierarchical Agglomerative Clustering

In this work, we concentrate on hierarchical agglomerative clustering. Unlike partitional clustering algorithms that build a hierarchical solution from top to bottom, repeatedly splitting existing clusters, agglomerative algorithms build the solution by initially assigning each document to its own cluster and then repeatedly selecting and merging pairs of clusters, to obtain a single all-inclusive cluster, generating the cluster tree from leaves to root [3]. The main parameters in agglomerative algorithms are the metric used to compute the similarity of documents and the method used to determine the pair of clusters to be merged at each step.

In these experiments, the cosine distance, equation (3), is used to compute the similarity between two documents di and dj. This widely utilised document similarity measure becomes one if the documents are identical, and zero if they share no words. The two clusters to merge at each step are found using the average link method. In this scheme, the two clusters to merge are those in which the average similarity between the documents in one cluster and those in the other is least.

Table 1. Statistics of Document News and Features

Category (Num Docs)	Language	Total Words	Avg. Words	Different Terms
News briefs (1835)	English	279,758	152	8,456
	Bulgarian	288,784	157	15,396
Features (2172)	English	936,795	431	16,866
	Bulgarian	934,955	430	30,309

To measure the quality of clustering, we use purity [6], equation (4), where each cluster C from a clustering tree C_{ALL} of the set of documents D is compared with the expected assigned category labels L from the list of categories L_{ALL}. Precision is the probability of a document in cluster C being labelled L. Purity is the percent of correctly clustered documents, equation (5).

3. Experimental Design

In this experiment, there are two categories of parallel corpora (News Briefs and Features) in two different languages, English and Bulgarian. In both corpora, each English document E corresponds to a Bulgarian document B with the same content, see Table 1. It is worth noting that the Bulgarian texts have a higher number of terms after stemming and stopword removal.

The process of stemming English corpora is relatively simple due to the low inflectional variability of English. However, for morphologically richer languages, such as Bulgarian, where the impact of stemming is potentially greater, the process of building an accurate algorithm becomes a more challenging task [1]. In this experiment, the Bulgarian texts are stemmed by the BulStem algorithm [1]. English documents are stemmed by a simple affix removal algorithm.

Figure 1 illustrates the experimental design set up. The documents in each language are clustered separately according to their categories (News Briefs or Features) using hierarchical agglomerative clustering. The output of each run consists of three elements: a list of terms characterizing the cluster, the cluster members, and the cluster tree for each set of documents. The next section contains a detailed comparison of the results for the two languages looking at each of these elements.

4. Experimental Results

4.1 Mapping of English-Bulgarian cluster memberships

In a first experiment, every cluster in English is paired with the Bulgarian cluster with which it shares the most documents. Pairing the languages the other way is possible, but omitted here for reasons of space Two precision values of these pairs are then calculated, the precision of the English-Bulgarian mapping (EBM) and that of the Bulgarian-English mapping (BEM).

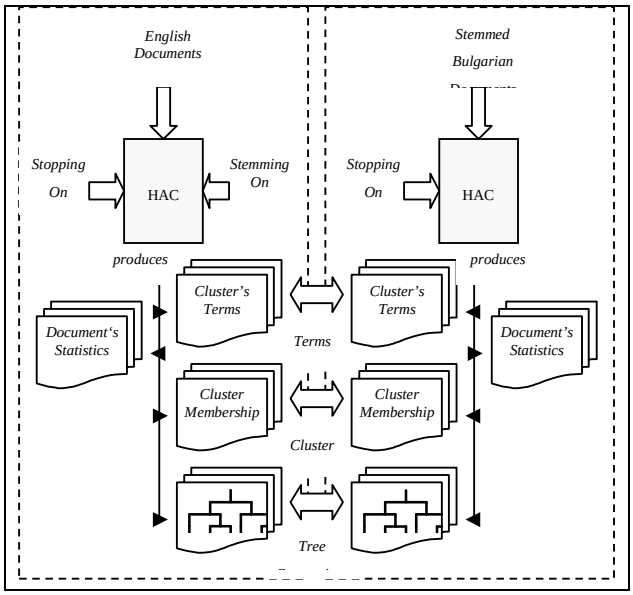

Figure 1. Experimental set up for parallel clustering task

Figures 2-7 show the precisions for the EBM and BEM for the cluster pairings obtained with varying numbers of clusters, k (k = 10, 20, 40) and for both News Briefs and Features. For example, in Figure 2, cluster 1 from English texts is best matched with cluster 1 from the Bulgarian texts with the English-to-Bulgarian mapping (EBM) precision equals to 76.11 and Bulgarian-to-English mapping (BEM) precision equals to 95.19.

It is also possible to study the purity of the mappings. Table 2 indicates the purity of the English-Bulgarian document mapping for various values of k.

A final point of interest is the extent to which the mapping EBM matches BEM. Table 3 shows that alignment between the two clusters is 100% when k = 10 for both categories of document. However, as the number of clusters increases, there are more clusters that are unaligned between the mappings. This is probably due to the fact that Bulgarian documents have a greater number of distinct terms. As the Bulgarian language has more word forms to describe English phrases, this may affect the computation of weights for the terms during the clustering process.

Table 2. Degree of Purity for Cluster Mapping for English-Bulgarian Documents

Category	k=5	k=10	k=15	k=20	k=40
News briefs	0.82	0.63	0.67	0.65	0.59
Features	N/A	0.77	N/A	0.61	0.54

Table 3. Percentage Cluster Alignment

Category	k = 10	k =20	k = 40
News briefs	100.0%	85.0%	82.5%
Features	100.0%	90.0%	80.0%

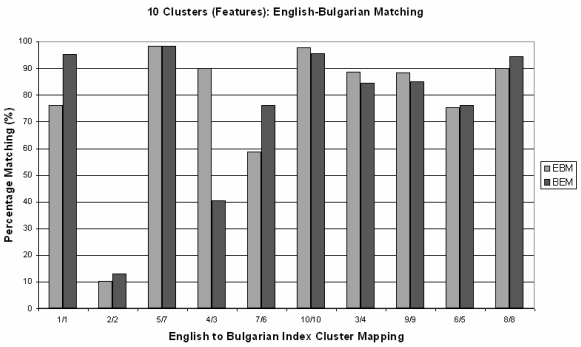

Figure 2. 10 Clusters (Features): English to Bulgarian Index Cluster Mapping

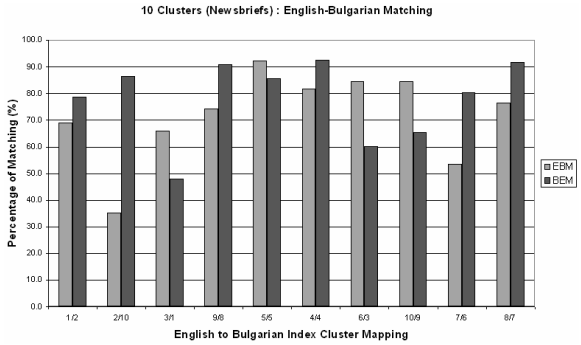

Figure 3. 10 Clusters (News Briefs): English to Bulgarian Index Cluster Mapping

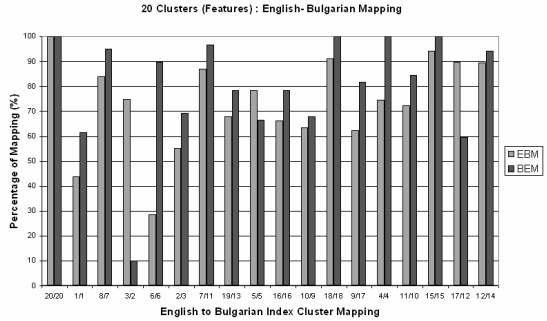

Figure 4. 20 Clusters (Features): English to Bulgarian Index Cluster Mapping

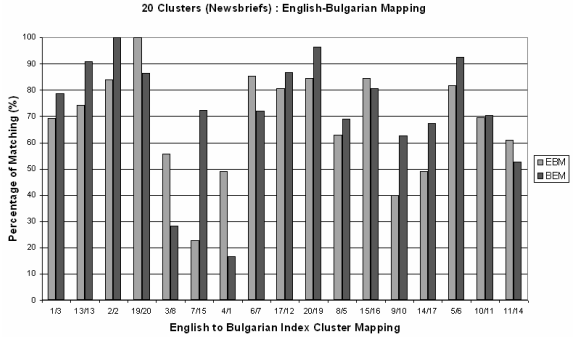

Figure 5. 20 Clusters (News Briefs): English to Bulgarian Index Cluster Mapping

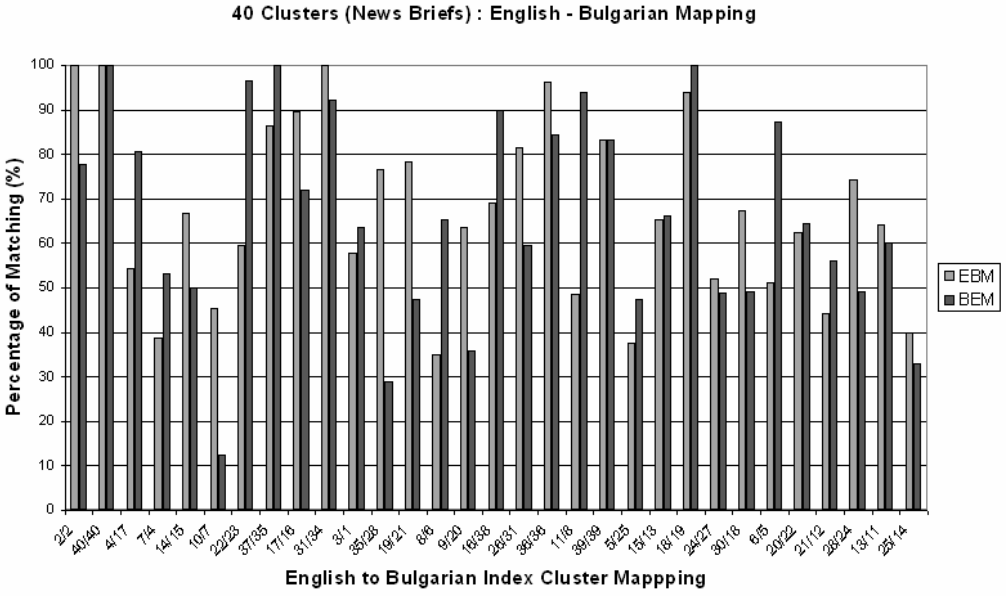

Figure 6. 40 Clusters (News briefs): English to Bulgarian Index Cluster Mapping

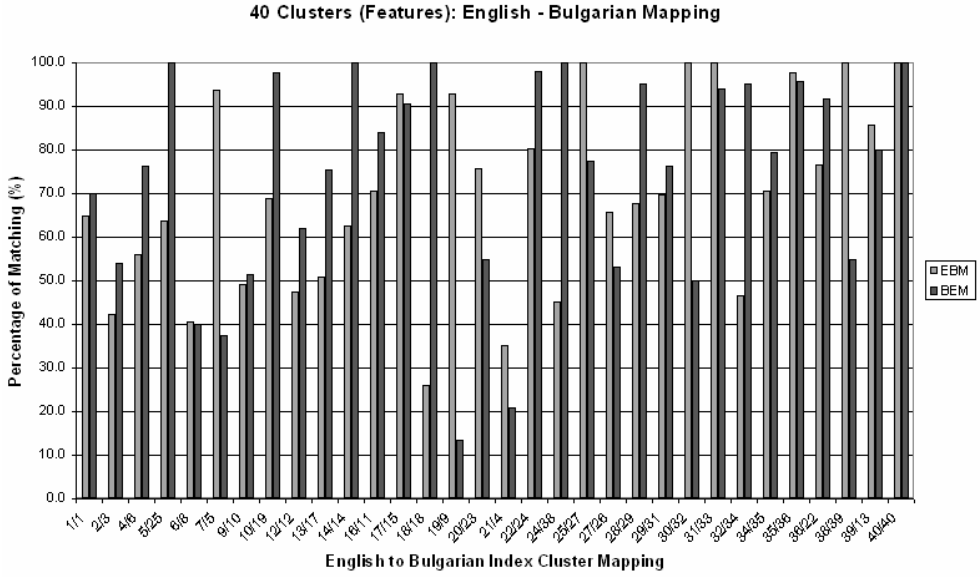

Figure 7. 40 Clusters (Features): English to Bulgarian Index Cluster Mapping

4.2 Comparison of HAC Structure

The trees' structures are then constructed by first aligning the clusters produced from both sets of documents and observed the differences in the tree's structure. When k = 10, Figure 8 and Figure 10 illustrate that the tree structures for both the English and Bulgarian documents are similar. However, when k = 20, the tree's structures are quite different due to the existence of unassigned clusters. Again, it may be caused by the different terms selected and used to cluster them. Since documents are clustered based on the weight of relevance of the terms, Bulgarian documents have more different terms considered during the clustering process.

4.3 Comparison of Terms Extracted from English and Bulgarian Clusters

Most terms that describe all the clusters for both sets of English and Bulgarian documents have similar meaning as illustrated in Tables 4 and 5 (Parts 1 and 2). However, as the number of clusters increases, some slight differences in terms of words/terms that describe the characteristics of the clusters occur.

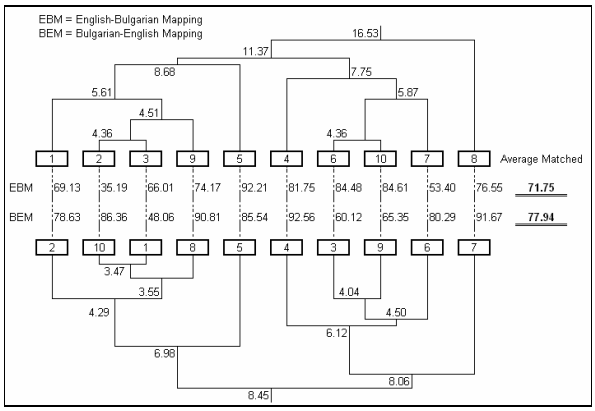

Figure 8. 10 Clusters of EBM (News briefs)

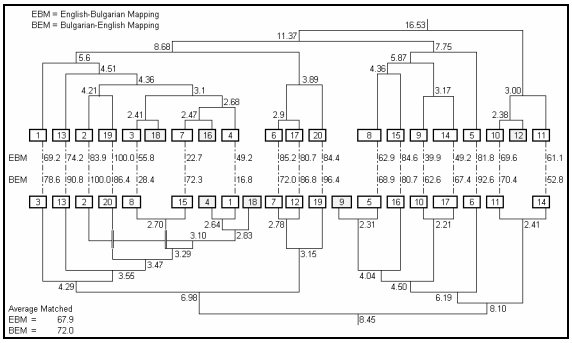

Figure 9. 20 Clusters of EBM (News briefs)

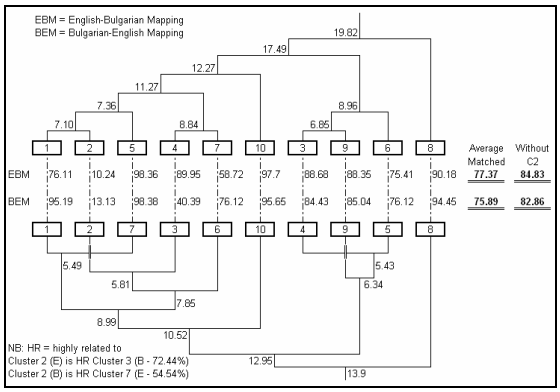

Figure 10. 10 Clusters of EBM (Features)

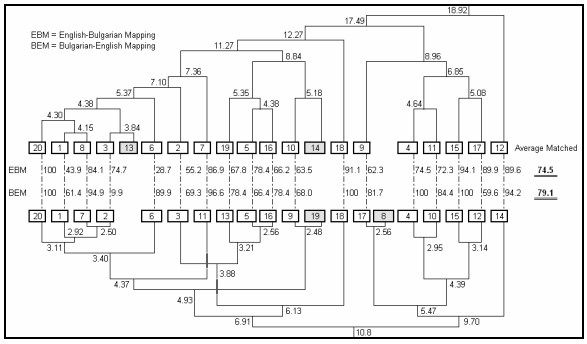

Figure 11. 20 Clusters of EBM (Features)

Table 4. Extracted Terms Comparison between English-Bulgarian Matched Cluster (k = 10)

1	2	3	4	5	6	7	8	9	10
macedonia	olymp	eu	kosovo	turkei	tribun	serbia	bih	bulgarian	croatia
macedonian	bird	albania	provinc	turkish	crime	serbian	rs	bulgaria	croatian
tv	flu	albanian	statu	erdogan	war	montenegro	ashdown	mediapool	gotovina
a1	game	romania	unmik	eu	milosev	mladic	novin	sofia	hina
skopj	test	minist	serb	ankara	trial	belgrad	nezavisn	btv	zagreb
vesnik	medal	countri	pristina	cypru	court	tanjug	repres	iraq	list
utrinski	greek	cent	albanian	cypriot	prosecutor	b92	pb	bta	sanad
makfax	athen	european	belgrad	anadolu	hagu	minist	high	parvanov	vecernji
crvenkovski	greec	nato	jessen	agenc	bosnian	zoran	republika	minist	ant
mia	bronz	bih	petersen	greek	serb	kostunica	srpska	trud	hrt
2	**10**	**1**	**4**	**5**	**3**	**6**	**7**	**8**	**9**
македони	грип	алба	косов	турци	трибунал	гора	рс	българск	хърват
македонск	птичи	ес	провинци	турск	престъпл	сърбия	бих	българи	хърватск
а1	птици	парти	статут	ердоган	милошевич	Черна	ващаун	ирак	готовин
цървенковск	вирус	румъни	прищин	ес	оон	младич	представител	софия	хина
скопие	H5N1	нато	юнмик	анкар	военни	сърбия-Черн	сръбск	бтв	лист
тв	лебед	минист	косовск	кипър	сръбск	белград	независн	медиапул	загреб
бучковск	птичия	други	йесен-петерсен	анадолск	обвин	сръбск	новин	първанов	санадер
утринск	случаи	правителств	оон	агенци	г	б92	пбс	бнт	ес
макфакс	мъртв	новин	сръбск	кипърск	хага	танюг	републи	бта	месич
трайковск	щам	македони	белград		понте	ес	върхов	минист	вечер

Table 5 (Part 1). Extracted Terms Comparison between English-Bulgarian Matched Cluster (k = 20)

1	2	3	4	5	6	7	8	9
macedonia	олимп	albanian	cent	kosovo	turkei	eu	tribun	serbia
macedonian	game	albania	gt	provinc	turkish	romania	crime	serbian
tv	medal	tirana	lt	statu	eu	romanian	war	montenegro
a1	greek	osc	bih	unmik	ankara	rompr	milosev	b92
skopj	athen	elec	bank	serb	erdogan	minist	trial	tanjug
vesnik	greec	moisiu	deficit	pristina	acces	wednesdai	court	djindjic
utrinski	bronz	ata	govern	albanian	istanbul	croatia	prosecutor	parti
makfax	won	tuesdai	imf	belgrad	membership	european	hagu	zoran
crvenkovski	men	alfr	undp	jessen	talk	countri	bosnian	belgrad
mia	stadium	countri	world	petersen	ntv	acces	serb	minist
3	**2**	**8**	**1**	**6**	**7**	**15**	**5**	**10**
македони	олимпийск	алба	сръбск	косов	турци	румъни	трибунал	гора
македонск	медал	нато	мли	провинци	турск	румънск	престъпл	Черна
а1	атин	македони	правителств	статут	ес	ромпрес	военни	сърбия
цървенковск	олимпиад	ес	бежан	прищин	анкар	ес	оон	сърбия-Черн
скопие	игрит	албанск	други	юнмик	ердоган	търчану	обвин	сръбск
тв	гърци	тиран	новин	косовск	преговор	попеску	г	белград
бучковск	спечел	минист	%	йесен-петер	членств	найн	караджич	б92
утринск	игри	комиси	евро	оон	юрдск	о'клок	понте	референдум
макфакс	бронзов	европейск	бих	сръбск	нтв	калин	дел	тадич
трайковск	категори	ек	представител	белград	люл	настас	хага	танюг

Table 5 (Part 2). Extracted Terms Comparison between English-Bulgarian Matched Cluster (k = 20)

10	11	13	14	15	17	19	20
bih	bih	bulgarian	mladic	croatia	turkei	bird	cypriot
ashdown	bosnian	bulgaria	serbia	croatian	iraq	flu	cypru
repres	sfor	mediapool	montenegro	gotovina	turkish	h5n1	turkish
rs	serb	sofia	serbian	hina	erdogan	dead	greek
high	karadz	btv	ratko	zagreb	akp	test	island
novin	rs	iraq	ljajic	list	anadolu	viru	plan
nezavisn	srebrenica	bta	belgrad	sanad	billion	swan	reunif
ohr	search	parvanov	tribun	vecernji	gul	case	annan
reform	srpska	minist	war	ant	recep	strain	turkei
pb	republika	trud	crime	hrt	tayyip	found	agenc
11	**14**	**13**	**17**	**16**	**12**	**20**	**19**
рс	рс	българск	младич	хърват	ирак	прип	кипърск
бих	сребрениц	българи	сърбия	хърватск	турск	птичи	кипър
ащдаун	сфор	ирак	гора	готовин	турци	птици	остров
представител	сръбск	софия	Черна	хина	пср	вирус	турци
независн	босненск	бтв	ратко	лист	ердоган	H5N1	плана
реформ	бих	медиапул	трибунал	загреб	анадолск	лебед	гърци
върхов	кланет	първанов	белград	санадер	американск	птичия	турск
новин	републи	бнт	ртс	ес	саш	случаи	оон
пбс	босненско-сръбск	бта	станкович	месич	север	мъртв	анан
парти	караджич	минист	сръбск	вечер	агенци	щам	денкташ

4.4 Term Reduction

Having seen in the previous experiment that the most representative words for each cluster are similar for each language, an interesting question is whether clustering using only these words improves the overall accuracy of alignment between the clusters in the two languages. The intuition behind this is that, as the words charactering each cluster are so similar, removing most of the other words from consideration may be more akin to filtering noise from the documents than to losing information.

The clustering is rerun as before, but with only a subset of terms used for the clustering. That is to say, before the tf-idf weights for each document are calculated, the documents are filtered to remove all but n of the terms from them. These n terms are determined by first obtaining 10 clusters for each language, and then extracting the n/10 terms which best characterise each cluster.

Four new clusters are thus created. In addition to the English and Bulgarian datasets used in the previous experiments, datasets containing only 100 and 500 terms for each of the languages are also clustered. For each cluster tree, 10 clusters are extracted, pairings are aligned and purities calculated as explained in section 4.1.

The results of comparing clusters in English and Bulgarian are shown in Table 6. These clearly indicate that as the number of terms used in either language falls, the alignment between the two clusterings produced also decreases. While term reduction in either language decreases the matching between the clusters, the effect is fairly minimal for English and far more pronounced for Bulgarian.

In order to seek to explain this difference between the languages, it is possible to repeat the process of aligning and calculating purity, but using pars of clusters from the same language, based on datasets with different levels of term reduction. The results of this are summarised in Table 7.

This table demonstrates that, for both languages, as the number of terms considered decreases, the clusters formed deviate further and further from those for the unreduced documents. While the deviation for English is quite low (and may indeed be related to the noise reduction sought), for Bulgarian reducing the number of terms radically alters the clusters formed. As with earlier experiments, the high morphological variability of Bulgarian compared to English may again be the cause of the results observed.

Table 6. Number of aligned clusters and purity for reduced term clustering (k = 10)

		Bulgarian Terms		
		All	500	100
English Terms	All	10 74.9%	4 54.2%	3 53.0%
	500	9 72.9%	4 46.0%	3 51.5%
	100	9 70.3%	4 60.1%	2 75.5%

Table 7. Number of aligned clusters and purity for reduced term datasets against the unreduced datset (k =10)

	All	500	100
English	10 100%	10 80.1%	9 74.2%
Bulgarian	10 100%	4 53.0%	3 53.0%

5. Conclusions and Future Work

This paper has presented the idea of using hierarchical agglomerative clustering on bilingual parallel corpora. The aim has been to illustrate this technique and provide mathematical measures which can be utilised to quantify the similarity between the clusters in each language.

In the paper, we have clustered bilingual parallel corpora of English-Bulgarian. The differences of all the clusters were compared, based on the tree structures. We can conclude that with a smaller number of clusters, k, all of the clusters from English texts can be mapped into the clusters of Bulgarian texts, with higher degree of purity. In contrast, with a larger number of clusters, more clusters from English texts cannot be mapped into the clusters of Bulgarian texts, and the degree of purity is very low. In addition, the tree structures for both the English and Bulgarian texts are similar with respect to each other, when k is reasonable small ($k < 15$).

Attention then shifted to looking at the terms which best characterized each cluster. The top two or three terms extracted from both texts carried the same meaning. Building on this, the idea of using term reduction to prune documents down to these most characteristic terms was investigated. However it was found that the matching between clusters dropped as the number of terms was reduced, especially for the Bulgarian language.

A common factor of all the aspects of parallel clustering studied was the importance that may be attached to the higher degree of inflection in Bulgarian. From the very beginning, the significantly lower degree of compression that resulted from stemming Bulgarian was noted. This implies that there were a larger number of Bulgarian words which expressed the same meaning, but which were not identified as such. It is likely that this is one of the factors responsible for decreasing the alignment between the clusters for larger values of k. Additionally, it was clearly demonstrated that this phenomenon would preclude any improvements in alignment or purity being observed in the term reduction experiment.

The current paper has presented a novel technique for use in parallel corpora, and there are many aspects remaining to be studied. For future work, we plan to provide a dictionary of terms or words that will definitely improve the clustering task. A detailed comparison of other methods of clustering also could be done to determine the dependable of language-specifics on the methods used to cluster them. In short, stemming and stopwords removal are very vital stages in clustering bilingual parallel corpora and it has been shown in this experiment that different methods used for stemming and stopwords removal may lead to different cluster alignment.

6. References

[1] P. Nakov, BulStem: Design and Evaluation of Inflectional Stemmer for Bulgarian. In Proceedings of Workshop on Balkan Language Resources and Tools (1st Balkan Conference in Informatics), Thessaloniki, Greece, November, 2003.

[2] G. Salton and J. Michael, McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc., New York, NY, 1986.

[3] Y Zhao and G Karypis. Hierarchical clustering algorithms for document datasets. Data Mining and Knowledge Discovery, 10(2):141.168, 2005

[4] A. Hotho, S. Staab, and G. Stumme. Text clustering based on background knowledge. Technical Report No. 425, 2003.

[5] S. Dumais, T. Landauer, and M. Littman, Automatic cross-linguistic information retrieval using latent semantic indexing. In SIGIR '96 – Workshop on Cross-Linguistic Information Retrieval, pp. 16–23, 1996

[6] P. Pantel and D. Lin. Document clustering with committees. In *Proc. Of SIGIR'02, Tampere, Finland,* 2002.