The use of machine translation tools for cross-lingual text mining

Blaž Fortuna BLAZ.FORTUNA@IJS.SI
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

John Shawe-Taylor JST@ECS.SOTON.AC.UK
University of Southampton, Southampton SO17 1BJ

Abstract

Eigen-analysis such as LSI or KCCA was already successfully applied to cross-lingual information retrieval. This approach has a weakness in that it needs an aligned training set of documents. In this paper we address this weakness and show that it can be successfully avoided through the use of machine translation. We show that the performance is similar on the domains where human generated training seta are available. However for other domains artificial training sets can be generated that significantly outperform human generated ones obtained from a different domain.

1. Introduction

The use of eigen-analysis in cross-lingual information retrieval was pioneered by Dumais et al. (Dumais et al., 1996). They used Latent Semantic Indexing to documents formed by concatenating the two versions of each document into a single file. The training set was therefore required to be a paired dataset, meaning a set of documents together with their translations into the second language.

This restriction also applied to the later application of kernel canonical correlation analysis to this task (Vinokourov et al., 2002). The difference in this approach is that the two versions of the documents are kept separate and projection directions for the two languages are sought that maximise the correlation between the projections of the training data. These directions are then used to create a 'semantic space' in which the cross-lingual analysis is performed.

Appearing in *Proceedings of the Workshop on Learning with Multiple Views*, 22^{nd} ICML, Bonn, Germany, 2005.

This approach was applied initially to the Hansard corpus of English/French paired documents from the Canadian parliament (Vinokourov et al., 2002). The semantic space derived in this way was further used to perform text classification on a separate corpus. Here the Reuters-21578 data was used.

The same approach has been used for more distinct languages in a paper studying cross-lingual information retrieval of Japanese patents (Li & Shawe-Taylor, 2005). Again this relied on using a paired dataset of Japanese patents as training data.

The approach to cross-lingual information retrieval and semantic representation has therefore proven reliable and effective in a number of different contexts. There is, however, an apparently unavoidable weakness to the approach in that a paired training set is required whose documents adequately cover the topics of interest. Indeed in the experiment that applied the semantic space learned with Hansard data to the Reuter's documents, the small overlap of the two vocabularies inevitably resulted in poorer performance.

This paper addresses this weakness by using machine translation to generate paired datasets that can be used to derive a semantic space using documents directly relevant to the target domain.

The paper is organised as follows. The next section discusses the questions raised by the use of automatic translation and outlines the experiments that will be presented to provide answers to these questions. Section 3 gives a brief summary of the KCCA approach to finding a semantic subspace mapping, while Section 4 presents the experimental results. We finish with some conclusions.

2. Using machine translation

The use of machine translation (MT) ensures that appropriate datasets can be generated but raises the

question of whether their quality will be sufficient to derive an accurate semantic space. Clearly we would expect that having a hand translated dataset will be preferable to using MT software. The first question this paper will address is the extent to which this is true.

Hence, the paper investigates how the quality of a machine translation generated dataset compares with a true paired corpus when one is available. This experiment is performed on the Hansard corpus with very encouraging results.

The advantage of using a machine generated paired dataset is that the topic of the articles will be identical to those on which the analysis is to be performed. In contrast the best available hand translated corpus might be for documents whose topics are only loosely related to those being studied. So we have a dilemma: do we use a machine translated corpus with a close topic match or a hand translated corpus with a weaker match. The second set of experiments reported in this paper will attempt to address this dilemma.

We consider a dataset for which paired training data is not available. Here we tackle a classification task and investigate the effectiveness of the semantic space generated from the translated paired corpus. We compare classification accuracy using this space with the space obtained from a paired dataset with a weaker overlap of topic with the documents being classified. For these experiments we used the now standard classification algorithm of support vector machines. Again the results obtained are very encouraging.

2.1. Related work

The MT was already used in the context of cross-language IR. D. W. Oard used it in the (D. W. Oard, 1998) as a method for translating the queries or the documents between bag-of-words spaces for different languages. A more similar approach to ours was used in the (M. L. Littman, S. T. Dumais and T. K. Landauer, 1998). They generated a separate LSI semantic space for each of the languages. For example, the semantic space was generated using the English documents from the training set and all non-English documents from the test set were than translated using MT and mapped into this semantic space. Our approach differs in that it only uses MT for the training period. In a practical setup this can be crucial since there is no need to call the time-expencive MT in the query loop. The aim of this paper is to show that MT can be used for obtaining a paired corpus for KCCA that is well matched to the target documents and not to to perform a general comparison of KCCA with other CLIR methods.

3. Summary of KCCA

Canonical Correlation Analysis (CCA) is a method of correlating two multidimensional variables. It makes use of two different views of the same semantic object (eg. the same text document written in two different languages) to extract representation of the underlying semantics.

Input to CCA is a paired dataset $S = \{(u_i, v_i); u_i \in U, v_i \in V\}$, where U and V are two different views of the data – each pair contains two views of the same document. The goal of CCA is to find two linear mappings into a common semantic space W from the spaces U and V. All documents from U and V can be mapped into W to obtain a view- or in our case language-independent representation.

The criterion used to choose the mapping is the correlation between the projections of the two views across the training data in each dimension. This criterion leads to a generalised eigenvalue problem whose eigenvectors give the desired mappings.

CCA can be kernelized so it can be applied to feature vectors only implicitly available through a kernel function. There is a danger that spurious correlations could be found in high dimensional spaces and so the method has to be regularised by constraining the norms of the projection weight vectors. A parameter τ controls the degree of regularisation introduced. The kernelized version is called Kernel Canonical Correlation Analysis (KCCA).

Example Let the space V be the vector-space model for English and U the vector-space model for French text documents. A paired dataset is then a set of pairs of English documents together with their French translation. The output of KCCA on this dataset is a semantic space where each dimension shares similar English and French meaning. By mapping English or French documents into this space, a language independent-representation is obtained. In this way standard machine learning algorithms can be used on multi-lingual datasets.

4. Experiments

In the following experiments, two issues regarding artificially generated corpora are discussed. First we compared it to a human generated corpus in domains where a human generated corpus is already available. The goal of this part is to check if the artificial corpus

can deliver comparable results. For the second part of the experiments we chose a domain and a problem for which human generated corpora were not available. We wanted to show, that by using documents from this domain an artificial corpus can be generated which outperforms human generated corpora obtained from different domains. Due to the datasets available we chose an information retrieval task for the first part of experiments and a text classification task for the second part.

4.1. Information Retrieval

The first part of experiments was done on the Hansards corpus (Germann, 2001). This corpus contains around 1.3 million pairs of aligned text chunks from the official records of the 36th Canadian Parliament. The raw text was split into sentences with Adwait Ratnaparkhi's MXTERMINATOR and aligned with I. Dan Melamed's GSA tool. The corpus is split into two parts, House Debates (around 83% of text chunks) and Senate Debates. These parts are than split into a training part and two testing parts. For our experiments we used the House Debates part from which we used only the training part and first testing part. The text chunks were split into 'paragraphs' based on '* * *' delimiters and these paragraphs were treated as separate documents. We only used documents that had the same number of lines in both their English and French version.

The training part was used as a human generated aligned corpus for learning semantic space with KCCA. In order to generate an artificial aligned corpus we first split the training documents into two halves. From the first half we kept only the English documents and only the French documents from the second half. In this way we obtained two independent sets of documents, one for each language. We then used *Google Language Tools* [1] to translate each document into its opposite language and generate an artificial aligned corpus. Some statistics on the corpora used in this experiment can be found in Table 1.

Table 1. Hansards aligned corpora

	Train	Artificial	Test1
Text chunks	495,022	495,022	42,011
Documents	9,918	9,918	896
En. words	38,252	39,395	16,136
Fr. words	52,391	55,425	21,001

From each corpus we learned with KCCA a language independent semantic space with 400, 800 or 1200 dimensions on a subset of 1500 documents.

[1] `http://www.google.com/language_tools`

Table 2. Top1 and Top10 results for the queries with 5 keywords are on left side and with 10 keywords are on the right side

n	1 [%]	10 [%]	1 [%]	10 [%]
En - En	96	100	99	100
Fr - Fr	97	100	100	100

The documents for these subsets were selected randomly and all results were averaged over five runs with different seeds for the random number generator. We ran experiments for the regularization parameter τ set to 0.2, 0.5 and 0.8, but because results for different parameters were not much different only results for $\tau = 0.5$ are presented. The threshold for the Partial Gram-Schmidt algorithm (or equivalently incomplete Cholesky decomposition of the kernel matrices) was set to 0.4.

For the information retrieval task, the entire first testing part of the Hansards corpus was projected into the language independent semantic space learned from the human generated corpus or from the artificial corpus. Each query was treated as a text document and its TFIDF vector was projected into the KCCA semantic space. Testing documents were than retrieved using nearest neighbour with cosine distance to the query.

In the first experiment each English document was used as a query and only its mate document in French was considered relevant for that query (Vinokourov et al., 2002). The same was done with French documents as queries and English documents as test documents. We measured the number of times that the relevant document appeared in the set of the top n retrieved documents (Top n). The Top1 results for both corpora are on average 96-98%, with results for human generated corpus generally scoring around 2% higher. The Top10 results were 100% for the both corpora.

For the next experiment we extracted 5 or 10 keywords from each document, according to their TFIDF weights, and used them for a query. Only the document from which the query was extracted and its mate document were regarded as relevant. We first tested queries in the original bag-of-words space and these results can serve as a baseline for the experiments done in the KCCA semantic spaces. Results are shown in Table 2. All queries were then tested in a similar way as before, the only difference is that this time we also measured the accuracy for cases where the language of the query and the relevant document were the same. Results for the queries with 5 keywords are presented

Table 3. Top1 and Top10 results for the queries with 5 keywords for the human generated corpus (top) and artificial corpus (bottom). The numbers are Top1/Top10 in percent.

	EN – EN	EN – FR	FR – EN	FR – FR
dim	1/10	1/10	1/10	1/10
400	76/98	59/93	60/92	74/98
800	83/99	64/95	65/94	81/99
1200	87/99	66/96	65/95	84/99
400	76/97	49/89	50/87	72/97
800	84/99	55/91	56/89	80/99
1200	86/99	58/91	59/90	83/99

in Table 3. and for the queries with 10 keywords in Table 4.

It is interesting to note that, for cases where the query was in the same language as the documents we searched over, the results are equal or slightly better for the artificial corpus than for the human generated one. This shows that, from both corpora, KCCA finds a similar semantic basis in vector-space models of English and French documents. However, the results for the artificial corpus are not as good as for the human generated corpus when it comes to cross-lingual queries. For queries with only 5 keywords, Top1 results for the artificial corpus are on average around 8% lower than for the human generated corpus while for queries with 10 keywords this drops to around 7%. Note that this difference stays constant when the dimensionality of semantic space increases. The difference between artificial and human generated corpora, when measuring the recall for the top 10 retrieved documents, drops to around 5% for queries with 5 keywords and to only 2% for queries with 10 keywords. The results for the cross-language parts of the experiments are lower for the artificial corpus than for the human generated corpus. The difference is not significant and a language independent semantic space learned on an artificial aligned corpus can still be successfully used in practice.

Table 4. Top1 and Top10 results for the queries with 10 keywords for the human generated corpus (top) and artificial corpus (bottom). The numbers are Top1/Top10 in percent.

	EN – EN	EN – FR	FR – EN	FR – FR
dim	1/10	1/10	1/10	1/10
400	93/99	79/99	78/97	90/100
800	96/100	82/99	81/98	94/100
1200	97/100	82/99	81/98	96/100
400	94/100	70/96	69/96	91/100
800	97/100	75/98	75/97	95/100
1200	97/100	77/98	75/97	96/100

4.2. Classification

The second part of the experiments was done on the Reuters multilingual corpora (Reuters, 2004) (mul, 2004), which contain articles in English, French, German, Russian, Japanese and other languages. Only articles in English, French and German were used for this experiment. Articles for each language were collected independently and no human generated aligned corpus was available for this domain. All articles are annotated with categories.

The task addressed in this experiment was how to make use of the existing corpus of annotated documents from one language, for example English, for doing classification in some other language, for example French. This can be done with the use of KCCA for construction of a language independent semantic space in which annotated English documents can be used to train a classifier that can also be applied to the French documents [4]. The problem with this approach is that the expensive task of annotating French documents is replaced with the even more expensive task of generating the aligned corpus needed for KCCA. This can be elegantly avoided through the use of MT tools. Another possibility is to use an aligned corpus from some other domain, for example the Hansards corpus used in the previous experiments. However, documents from that corpus belong to different domain and may not cover all the semantics that appear in the news articles. On the other hand the artificial corpus is constructed from the same set of documents that will be used for training the classifiers.

For this experiment we picked 5000 documents for each of the three languages from the Reuters corpus. Subsets of these documents formed the training datasets for the classifiers. These same documents were also used for generating artificial aligned corpora in the same way as in the first part of the experiments; Google Language Tools were used to translate English documents to French and German and the other way around. In this way we generated English-French and English-German aligned corpora. We used the training part of the Hansards corpus as English-French human generated aligned corpora. Some statistics on the corpora used in this experiment can be found in Table 5.

KCCA was used for learning a language independent semantic space from these aligned corpora. The pa-

Table 5. English-French and English-German aligned corpora from the Reuters corpus.

	En-Fr	En-Gr
Paragraphs	119,181	104,639
Documents	10,000	10,000
English words	57,035	53,004
French words	66,925	—
German words	—	121,193

rameters used for learning were the same as for the information retrieval task. The only difference is that only subsets of 1000 documents were used. These subsets were selected randomly and the results presented were averaged over 5 runs. A linear SVM was used as a classification algorithm with cost parameter C set to 1.

The classification experiment was run in the following way. All the classifiers were trained on subsets of 3000 documents from the training set and the results were averaged over 5 runs. This means that the presented results are averaged over 25 runs. The classifiers trained in the original vector-space models are used as a baseline to which the ones trained in the KCCA semantic space can be compared. The documents from the English training set were projected into KCCA semantic space and a classifier was trained on them. The same was done with the French and German documents. The classifiers were tested on a subset of 50,000 documents from the Reuters corpora. The testing documents were also projected to KCCA semantic space for classifiers living in that space. We measured average precision: baseline results are shown in Table 6.

Table 6. Average precision for classifiers for categories CCAT, MCAT, ECAT and GCAT

	CCAT	MCAT	ECAT	GCAT
English	85 %	80 %	62 %	86 %
French	83 %	85 %	63 %	94 %
German	85 %	86 %	62 %	91 %

The results for the human generated and for the artificial English-French corpus are presented in Table 7 and in Figure 1. The results obtained with the artificial corpus were in all cases significantly better than the results obtained with the Hansard corpus and are close to the baseline results for single language classification. Note that the results for the artificial corpus only slightly improve when the dimensionality of semantic space increases from 400 to 800 while the results for the human generated corpus increase by 10 or more percent. This shows that the first 400 dimensions learned from the artificial corpus are much richer at capturing the semantics of news articles than the ones learned from Hansard corpus.

Table 7. Average precision [%] for classifiers learned in KCCA semantic space learned Hansards/artificial corpus (Hansard/artificial). The results are for the semantic space wiht 400 (top) and 800 (bottom) dimensions.

	CCAT	MCAT	ECAT	GCAT
En-En	59/79	40/76	25/51	51/78
En-Fr	41/78	21/81	18/54	75/89
Fr-En	55/80	30/76	22/50	40/77
Fr-Fr	40/78	24/82	19/54	77/89
En-En	67/80	61/82	38/54	67/79
En-Fr	47/79	32/82	27/55	80/90
Fr-En	60/80	43/76	30/52	51/78
Fr-Fr	53/79	59/83	38/56	85/89

Results for classification based on English-German artificial aligned corpus are shown in Table 8. Surprisingly in some cases the cross-lingual classifications do better than a straight German classifier. The results are not as close to the base line (Table 6) as the results from English-French artificial corpus. We suspect that this is due to the different structure of German which is evident in Table 5; the number of different words in the German articles is twice as high as in the English or French documents. One workaround would be to use more advanced preprocessing before using the bag

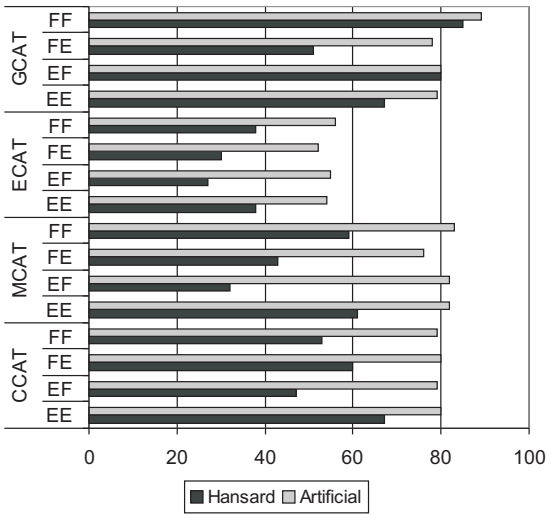

Figure 1. Average precision [%] for the classification of English and French documents in the 800 dimensional semantic space.

Table 8. Average precision [%] for classifiers learned in KCCA semantic space learned on artificially generated English-German aligned corpus

	CCAT	MCAT	ECAT	GCAT
En-En	75	77	49	81
En-Gr	72	82	46	87
Gr-En	70	75	43	78
Gr-Gr	67	83	44	86
En-En	76	78	52	82
En-Gr	73	82	47	88
Gr-En	71	75	46	79
Gr-Gr	68	83	47	86

of words or a use of different document representation like the string kernel.

5. Conclusion

The paper has addressed a pressing practical problem in the application of KCCA to cross-lingual information retrieval and language-independent semantic space induction in general, namely how to find an appropriate paired dataset.

Frequently we will only have access to a hand translated training corpus that is loosely related to the document corpus that is being analysed. The paper proposes a method of addressing this problem by using automatic machine translation tools to generate an 'artificial' paired corpus directly from the document corpus itself.

This raises two questions that are analysed in the paper. Firstly, how much worse is a semantic space derived from an artificial corpus than from a hand translated one, and secondly can the topic match offset any degradation resulting from the machine translation.

The first experiment showed that the degradation in performance does exist when we move to MT, but in a testing cross-lingual information retrieval task the reduction in recall was below 10%. This result certainly suggests that the advantage of exact topic match could well result in an increase in the quality of the semantic space obtained for a corpus with no hand translations available.

Our second experiment confirms this conjecture by demonstrating that the MT method improves cross-lingual classification results for the multi-lingual Reuters corpus when compared with using the semantic space induced from the hand translated Hansard corpus.

For these experiments the results are even more encouraging. They show a very significant advantage for the MT approach. Furthermore, the difference between the classification results using the semantic space and those obtained for single language classification using the bag of words feature space is not very large. This suggests that the method could be used to provide a general language independent classifier that can be used to classify documents from either language. This could potentially make it possible to use the topic labelling from one language to generate labels for newswire documents from the second language without the need for trained staff with appropriate language skills to perform the classification.

References

Reuters (2004). RCV2 - The Reuters Multi-lingual corpus.

Dumais, S. T., Landauer, T. K., & Littman, M. L. (1996). Automatic cross-linguistic information retrieval using latent semantic indexing. *Working Notes of the Workshop on Cross-Linguistic Information Retrieval.*

Germann, U. (2001). Aligned Hansards of the 36th Parliament of Canada. http://www.isi.edu/natural-language/download/hansard/. Release 2001-1a.

D. W. Oard (1998). A comparative study of query and document translation for cross-language information retrieval. *In proceedings of the 3rd Conference of the Association for Machine Translation in the Americas, pages 472-483.*

M. L. Littman, S. T. Dumais and T. K. Landauer (1998). Automatic Cross-Language Information Retrieval using Latent Semantic Indexing. *In cross-Language Information Retrieval, Kluwer..*

Li, Y., & Shawe-Taylor, J. (2005). Using kcca for japanese-english cross-language information retrieval and classification. *to appear in Journal of Intelligent Information Systems.*

Reuters (2004). RCV1-v2/LYRL2004: The LYRL2004 Distribution of the RCV1-v2 Text Categorization Test Collection. http://jmlr.csail.mit.edu/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm.

Vinokourov, A., Shawe-Taylor, J., & Cristianini, N. (2002). Inferring a semantic representation of text via cross-language correlation analysis. *Advances of Neural Information Processing Systems 15.*