Report on CLEF-2003 experiments: two ways of extracting multilingual resources from corpora

∓Nicola Cancedda, ∓Hervé Déjean, ∓Éric Gaussier, ∓Jean-Michel Renders, and ⋆Alexei Vinokourov

∓Xerox Research Centre Europe,
⋆Royal Holloway University of London
∓Firstname.Lastname@xrce.xerox.com,
⋆alexei@cs.rhul.ac.uk

Abstract. We present in this report two main approaches to cross-language information retrieval based on the exploitation of multilingual corpora to derive cross-lingual term-term correspondences. These two approaches are evaluated in the framework of the multilingual-4 (ML4) task.

1 Introduction

Most approaches to Cross-Language Information Retrieval (CLIR) rely on query translation based on existing machine-readable dictionaries and/or translation systems ([15, 2, 6, 12, 21], to name but a few), and face the problem of the adequacy of existing bilingual resources to the collection that is searched. However, when this collection is multilingual, one may benefit from automatically extracted bilingual lexicons, which can display a better coverage and allow for more accurate translations of queries. This perspective is mentioned in [6], even though the authors failed to derive accurate bilingual lexicons from their collection. It is indirectly exploited in [18] where the authors derive a probabilistic translation lexicon, based on IBM translation models 1 and 2 ([4]), from a corpus of parallel texts different from the searched collection.

We want to experiment here with two methods to exploit parallel corpora for CLIR purposes. The first one relies on the inference of a bilingual semantic representation via cross-language canonical correlation analysis, whereas the second one, more traditional, relies on the extraction of bilingual lexicons from parallel corpora.

However, the CLEF-2003 multilingual collection is not parallel, but comparable, that is to say that rather than being translations of one another, documents cover the same topics, in the same domains. Nevertheless, up to now, extraction methods developed on comparable corpora, unlike methods for parallel corpora, have not provided results good enough to be directly used in CLIR, as is argued in [19]. This indicates that a compromise between the use of parallel and comparable corpora has to be found, so as to derive query translation modules that display both accuracy and coverage properties. In addition to the above-mentioned methods, we will thus report on experiments aimed at combining bilingual lexicons extracted from parallel and comparable corpora. In our case, the parallel corpus retained is the JOC[1], whereas the comparable one is the

[1] Used in the Arcade evaluation task, www.lpl.univ-aix.fr/projects/arcade

collection itself. The implicit goal behind these experiments is to develop state-of-the-art query translation modules, fully adapted to the collection to be searched.

2 Linguistic preprocessing

As a preprocessing step, we tag and lemmatize corpora, queries and bilingual resources. Only lexical words (nouns, verbs, adverbs, adjectives) are indexed and only single word entries in our resources are used. Our (lexicon-based) lemmatizer provides a partial segmentation for the German compounds. Additionally, we segment German words which were not decomposed by the lemmatizer using the following patterns:

Pattern	Segmentation
A([ˆaeuioy])sB	A([ˆaeuioy]) B
A-B	A B

German spelling (umlaut and eszett) is also normalized.

3 Canonical correlation analysis for cross-lingual retrieval

In this work we automatically model a semantic correspondence between terms of different languages, in the spirit of cross-lingual latent semantic indexing (CL-LSI) [17]. In CL-LSI, using a parallel corpus, after merging each pair into a single 'document', one can interpret frequent co-occurrence of two terms in the same document as an indication of cross-language correlation. In this framework, a common vector-space, including words from both languages, is created and then the training set is analysed in this space using SVD. This problem can be regarded either as an unsupervised problem with paired documents, or as a supervised monolingual problem with very complex labels (i.e. the label of an English document could be its French counterpart). In either way, the data can be readily obtained without an explicit labeling effort, and furthermore there is no loss of information in compressing the meaning of a document into a discrete label. As an alternative to CL-LSI, we employ Canonical Correlation Analysis (CCA) [1] [16] to learn a representation of text that captures aspects of its meaning. Given a paired bilingual corpus, this method defines two embedding spaces for the documents of the corpus, one for each language, and an obvious one-to-one correspondence between points in the two spaces. CCA then finds projections in the two embedding spaces for which the resulting projected values are highly correlated. In other words, it looks for particular combinations of words that appear to have the same co-occurrence patterns in the two languages. Our hypothesis is that finding such correlations across a paired bilingual corpus will locate the underlying semantics, since we assume that the two languages are 'conditionally independent', or that the only thing they have in common is their meaning. The directions would carry information about the *concepts* that stood behind the process of generation of the text and, although expressed differently in different languages, are, nevertheless, semantically equivalent. This representation is then used for the retrieval task, providing a better performance than LSI on some tested corpora [24]. Such directions are then used to calculate the coordinates of the documents in a 'language independent' way. Of course, particular statistical care is needed for excluding 'spurious' correlations. We have shown that the correlations we find are not the effect of chance, and that the resulting representation significantly improves performance of retrieval systems [24]. Indeed, the correlation between English and French documents can be explained by means of relations between the generative processes of the two versions of the documents, that we assume to be conditionally independent given the *topic* or *content*. Under such assumptions, hence, such correlations detect similarities in content between the two documents, and can be exploited to derive a semantic representation of the text. The CCA machinery is briefly given below.

3.1 Canonical Correlation Analysis

For us, the multivariate random variables to which CCA is applied correspond to document-vectors (in the bag of words representation) in English and French, and there is a one-to-one relation between them corresponding to documents that are translations of each other. We will now consider sets of words that are correlated between the two languages (sets of words in the two languages that have a correlated pattern of appearance in the corpus). We will assume that such sets approximate the notion of 'concepts' in each language, and that such concepts are the translation of each other. Rather than considering plain sets, we will consider terms to have a degree of membership to a given set. In other words, the term t_i will be assigned a weight α_i for each concept we consider, and every concept will correspond to a vector $\alpha_x \in \Re^n$ in English, and a vector $\alpha_y \in \Re^m$ in French. We will use that weight α_i to form linear combinations of terms, so that they can define a direction in the term space.

Suppose as for CL-LSI we are given *aligned* texts in, for simplicity, two languages, i.e. every text in one language $x_i \in \Re^n$ is a translation of text $y_i \in \Re^m$ in another language. In practice, each text can correspond to a complete document, a paragraph, or a sentence. The finer the textual units, the more accurate the correlation statistics. Our hypothesis is that having aligned texts $S_x = (x_1, ..., x_\ell) \subseteq \Re^n$ and $S_y = (y_1, ..., y_\ell) \subseteq \Re^m$ we can learn (semantic) directions $\hat{w}_x$ and $\hat{w}_y$ where we use the notation $\hat{w} = \frac{w}{||w||}$ so that the projections $\hat{w}_x' x$ and $\hat{w}_y' y$ of input data images from the different languages would be maximally correlated. These new random variables are univariate, and linear combinations of the previous ones. We consider optimizing this quantity with respect to the choice of $\hat{w}_1 \in \Re^n$ and $\hat{w}_2 \in \Re^m$. This leads to the following objective functions and optimization problems:

$$\rho = \max_{w_x, w_y} \operatorname{corr}(\hat{w}_x' x, \hat{w}_y' y)$$

This optimization problem can be transformed into a generalized eigenvalue problem as follows. One is looking for the maximum correlation directions:

$$\begin{aligned} &\text{maximize } \rho = \frac{\operatorname{cov}(x,z)}{\operatorname{var} x \cdot \operatorname{var} z} = \frac{E[xz]}{\sqrt{E[xx]E[zz]}} = \frac{E[\mathbf{w}_x' \mathbf{C}_{xz}' \mathbf{w}_z]}{\sqrt{E[\mathbf{w}_x' \mathbf{C}_{xx}' \mathbf{w}_x] E[\mathbf{w}_z' \mathbf{C}_{zz} \mathbf{w}_z]}} \\ &\text{subject to } ||\mathbf{w}_x|| = ||\mathbf{w}_z|| = 1 \end{aligned}$$

where we are using the covariance matrix:

$$C = \begin{pmatrix} C_{xx} & C_{xz} \\ C_{zx} & C_{zz} \end{pmatrix} = E\left(\begin{pmatrix} \mathbf{x} \\ \mathbf{z} \end{pmatrix}\begin{pmatrix} \mathbf{x} \\ \mathbf{z} \end{pmatrix}'\right)$$

The solutions of this problem can be obtained by solving a related generalized eigenproblem

$$A\mathbf{w} = \lambda B\mathbf{w} \tag{1}$$

and the solution $\mathbf{w}$ directly provides the directions $\mathbf{w}_x$ and $\mathbf{w}_z$ of maximum correlation:

$$A = \begin{pmatrix} 0 & C_{xz} \\ C_{zx} & 0 \end{pmatrix}$$
$$B = \begin{pmatrix} C_{xx} & 0 \\ 0 & C_{zz} \end{pmatrix}$$
$$\mathbf{w} = \begin{pmatrix} \mu_x \mathbf{w}_x \\ \mu_z \mathbf{w}_z \end{pmatrix}$$

Note that if λ is an eigenvalue, so is $-\lambda$ thus the spectrum is $\{\lambda_1, -\lambda_1, ..., \lambda_N, -\lambda_N\}$.

3.2 Application of CCA to cross-lingual retrieval task

The kernel CCA procedure identifies a set of projections from both languages into a common semantic space. This provides a natural framework for performing cross-language information retrieval. We first select a number d of semantic dimensions, $1 \leq d \leq N$, with largest correlation values ρ. To process an incoming query q we expand q into the vector representation for its language $\widetilde{q}$ and project it onto the d canonical $\mathcal{F}$-correlation components: $[q] = A^T Z^T \widetilde{q}$ using the appropriate vector for that language, where A is a $N \times d$ matrix whose columns are the first solutions of (1) for the given language sorted by eigenvalue in descending order. Notice that in this case we use the standard dot product to perform the projection, but non-linear projections can also be obtained by replacing the dot product with a non-linear kernel ([]).

3.3 Learning on paired data

The whole training collection consists of 1.3 million pairs of aligned text chunks (sentences or smaller fragments) from the 36^{th} Canadian Parliament proceedings. We used only first 1000 documents. The raw text was split into sentences with Adwait Ratnaparkhi's `MXTERMINATOR` and the sentences were aligned with I. Dan Melamed's GSA tool (for details on the collection and also for the source see [11]).

The text was split into 'paragraphs' based on '***' delimiters and these 'paragraphs' were treated as separate documents. After removing stop-words in both French and English parts and rare words (i.e. appearing less than three times) we obtained 5159×1000 term-by-document 'English' matrix and 5611×1000 'French' matrix (we also removed a few documents that appeared to be problematic when split into paragraphs).

Table 1. Upper bound of the coverage of lexicons extracted from different sources

Elra	Oxford-Hachette	Hansard	JOC	ML4
0.78	0.78	0.80	0.90	0.98

3.4 Experimental Results

The test corpus and queries were processed by Xerox Research Centre Europe as explained in Section 2. The results were unimpressive due to the fact that we restricted ourselves only to the French part - Le Monde - due to the lack of time. Possibly there were also bugs in software and we are working to reveal them, renewed results may appear in the final version.

4 Query translation

We want to assess in this section the usefulness of bilingual lexicons extracted from collections. In order to illustrate the potential gain this approach could yield, we conducted the following simple experiment. We first collected all the English terms from the queries associated to the ML4 task, from years 2000 to 2003. We then tried to evaluate whether or not we were able to translate those terms with manually built, existing dictionaries, and whether or not we were able to translate them with bilingual lexicons automatically derived from multilingual collections. To this end, we retained two multilingual dictionaries, the ELRA dictionary[2], and the Oxford-Hachette[3]. For corpora, we retained a part of the Hansard[4], the JOC corpus (already mentioned in footnote 1, comprising ca. 3.5 millions English tokens), and the ML4 collection itself. For each term present in the set of English queries, we checked whether it was present in the lexicons associated with the above resources. The percentage of English terms found in the lexicons is summarized in table 1.

As one may have noticed, the figures we obtained are only upper bounds on the actual coverage of each resource, since the presence of a term in a dictionary does not imply that the proposed translation(s) are appropriate for the collection at hand. Furthermore, there is an important qualitative difference between manually built and automatically extracted lexicons, a difference that may well balance the advantage for corpus-based methods displayed in table 1. However, were we able to accurately extract bilingual lexicons from corpora, table 1 shows that we would have an important gain over using existing, general purpose dictionaries. Table 2 supports this latter fact and shows how the average precision evolves, on a sub-part of ML4, according to the lexicon used to translate queries.

The column (JOC+ML4) combines the lexicons extracted from the JOC and ML4 corpora, as detailed in section 4.3. The bilingual runs correspond to English queries

[2] Multilingual dictionary, available from ELRA, www.elra.info, comprising ca. 45000 English entries

[3] Bilingual English-French dictionary, comprising ca. 45000 English entries.

[4] In fact a sub-part of it, comprising ca. 20 millions English tokens

Table 2. Performance of different lexicons for query translation

Average precision	Elra	JOC	ML4	JOC+ML4
Bilingual	0.29	0.365	0.228	0.388
Multilingual (merge)	0.192	0.289	0.165	0.302
French (bilingual)	0.271	0.362	0.188	0.389
German (bilingual)	0.276	0.361	0.203	0.380
Spanish (bilingual)	0.304	0.411	0.221	0.431

translated in the corresponding target language (all these experiments, as well as the following ones, are based on the vector-space model). As one can note, the use of automatically derived bilingual lexicons significantly outperforms the use of existing dictionaries on this collection.

We are now going to review the methods we used for extracting bilingual lexicons from parallel and comparable corpora.

4.1 Bilingual lexicon extraction from parallel corpora

Recent research has demonstrated that statistical alignment models can be highly successful at extracting word correspondences from parallel corpora ([4, 5, 8, 9, 13, 14]) among others. All these works are based on the assumption that, once documents have been aligned at the sentence level, the more two words from different languages co-occur in aligned sentences, the more likely they are translations of each other. In the present paper, we rely on the word-to-word translation lexicon obtained from parallel corpora, following the method described in [10], which can be summarized as follows.

We first represent co-occurrences between words across translations by a matrix, the rows of which represent the source language words, the columns the target language words, and the elements of the matrix the expected alignment frequencies (EAFs) for the words appearing in the corresponding row and column. Empty words are added in both languages in order to deal with words with no equivalent in the other language.

The estimation of the expected alignment frequency is based on the Iterative Proportional Fitting Procedure (IPFP) presented in [3]. This iterative procedure updates the current estimate $n_{ij}^{(k)}$ of the EAF of source word i with target word j, using the following two-stage equations:

$$n_{ij}^{(k,1)} = \sum_{s,(i,j)\in s} n_{ij}^{(k-1,2)} \times \frac{s_i}{n_{i.}^{(k-1,2)}}$$
$$n_{ij}^{(k,2)} = \sum_{s,(i,j)\in s} n_{ij}^{(k,1)} \times \frac{s_j}{n_{.j}^{(k,1)}}$$

where $n_{i.}$ and $n_{.j}$ are the current estimates of the row and column marginals, s is a pair of aligned sentences containing words i and j, and s_i and s_j are the observed frequencies of words i and j in s. The initial estimates $n_{ij}^{(0,2)}$ are the observed frequencies of co-occurrences, obtained by considering each pair of aligned sentences and by incrementing the alignment frequencies accordingly. The sequence of updates will eventually converge and the EAFs are then normalized (by dividing each element n_{ij} by the row marginal $n_{i.}$), so as to yield probabilistic translation lexicons, in which each source word is associated with a target word through a score. In the remainder of the paper, we will use $P_1(t|s)$ to denote the probability of selecting target word t as translation for source word s, as given by this method.

4.2 Bilingual lexicon extraction from comparable corpora

Bilingual lexicon extraction from non-parallel but comparable corpora has been studied by a number of researchers, [19, 23, 22, 20, 7] among others. Their work relies on the assumption that if two words are mutual translations, then their more frequent collocates (taken here in a very broad sense) are likely to be mutual translations as well. Based on this assumption, a standard approach consists in building context vectors, for each source and target word, which aim at capturing the most significant collocates. The target context vectors are then translated using a general bilingual dictionary, and compared with the source context vectors.

Our implementation of this strategy relies on the following steps:

1. For each word w, build a context vector by considering all the words occurring in a window centered on w, run through the corpus. Each word i in the context vector of w is then weighted with a measure of its association with w. However, in order to ensure we make adequate use of the prior knowledge provided by the general dictionary, we include w in its context vector. Lastly, we have used here a window of 5 words before and after w, and retained the mutual information as the measure of association.
2. The context vectors of the target words are then translated with our general bilingual dictionary, leaving the weights unchanged (when several translations are proposed by the dictionary, we consider all of them with the same weight)
3. The similarity of each source word s, for each target word t, is computed on the basis of the cosine measure
4. The similarities are then normalized to yield a probabilistic translation lexicon, $P_2(t|s)$.

4.3 Model combination

Because they contain different information, the comparable and parallel corpora yield different translations that need be combined in order to obtain a complete translated query. Such a combination should account for the fact that for some source words the information provided by the comparable corpus is more reliable than the one provided by the parallel one (as is the case when the source word is not present in the parallel corpus), whereas for some other source words the situation is reversed. However, because of time constraints, we were not able to adopt this strategy, and had to resort to a simpler linear combination, in which the final vector representing the query in target language is given by:

$$\overrightarrow{q_t} = (\alpha \times P_1 + (1-\alpha) \times P_2) \times \overrightarrow{q_s} \quad (2)$$

α is a scalar representing the weight associated with the translation provided by the parallel corpus. We optimized the value of α on the queries corresponding to years 2000 to 2002.

4.4 Multilingual merging

Our strategy to merge results from different languages relies on the fact that if we use "similar" translation matrices, and if the scoring method is identical for each language, then one can directly merge the results from different languages. Using similar translation matrices means that the length (as measured by the norm) of target queries should be identical (since they all issue from the same English query). In order to ensure this, we normalise each target query by its length ($\overrightarrow{q_t} \rightarrow \frac{\overrightarrow{q_t}}{\|\overrightarrow{q_t}\|}$). Furthermore, to get an equivalent, on the English collection, of the translation step used in the other languages, we consider the English sub-collection to constitue a comparable corpus on its own, from which we build a term-term co-occurrence matrix in exactly the same way as we built a translation matrix in section 4.2 (the source and target languages being identical here). This matrix is then used to expand English queries with most similar terms.

4.5 Weighting schemes

Table 3 shows the results we obtained with two different weighting schemes on the ML4 collection, for the 2003 queries. Note that queries are weighted prior to translation.

Table 3. Influence of weighting schemes

Weighting scheme	Average precision
Lnu/ntn	0.2118
Ltc/ntn	0.1860

The results display in table 3 are obtained by translating queries with the combination of the lexicons derived from JOC and ML4 as explained above. Despite the important difference the two weighting schemes have on the monolingual collections (cf. e.g. [21]), we see here that the bilingual retrieval, followed by the multilingual merge, flattens the difference to only ca. 2.5 points.

5 Conclusion

We have tested two main approaches to cross-language information retrieval based on the exploitation of multilingual corpora to derive cross-lingual term-term correspondences. The first approach makes use of parallel corpora to derive an interlingual semantic representation of documents, using canonical correlation analysis. The second approach aims at directly extracting bilingual lexicons, both from parallel and comparable corpora, to be used for query translation. Our experiments show that the second approach outperforms a standard approach using existing bilingual dictionaries for query translation. We plan in the future to pursue the promising road of bilingual lexicon extraction from comparable corpora.

Acknowledgements

This research was partially supported by the European Commission under the KerMIT Project No. IST-2001-25431.

References

1. F. R. Bach and M. I. Jordan. Kernel indepedendent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.
2. L. Ballesteros and B.W Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1997.
3. Y. Bishop, S. Fienberg, and P. Holland. *Discrete Multivariate Analysis*. MIT Press, 1975.
4. P. Brown, S. Della Pietra, V. Della Pietra, and R.L. Mercer. The mathematics of statistical machine learning translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.
5. I. Dagan and I. Itai. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4), 1994.
6. M.W. Davis and W.C. Ogden. Quilt: Implementing a large-scale cross-language text retrieval system. In *Proceedings of the 20th 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1997.
7. Pascale Fung. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In Jean Véronis, editor, *Parallel Text Processing*, 2000.
8. W. A. Gale and K. W. Church. A program for aligning sentences in bilingual corpora. In *Meeting of the Association for Computational Linguistics*, pages 177–184, 1991.
9. E. Gaussier. Flow network models for word alignment and terminology extraction from bilingual corpora. In *Proceedings of the joint 17th International Conference on Computational Linguistics and 26th Annual Meeting of the Association for Computational Linguistics*, pages 444–450, 1998.
10. É Gaussier, D. Hull, and S. Ait-Mokhtar. *Parallel Text Processing Alignment and Use of Translation Corpora*, chapter Term Alignment in Use: Machine-Aided Human Translation. Kluwer Academic Publishers, 2000.
11. Ulrich Germann. Aligned Hansards of the 36th Parliament of Canada. http://www.isi.edu/natural-language/download/hansard/, 2001. Release 2001-1a.
12. F.C. Gey, V. Jiang H.and Petras, and A. Chen. Cross-language retrieval for the clef collections - comparing multiple methods of retrieval. In *Proceedings of the Cross-Lingual Evaluation Forum, CLEF 2000.*, 2000.
13. D. Hiemstra. Using statistical methods to create a bilingual dictionary. Master's thesis, Universiteit Twente, 1996.
14. David Hull. Automating the construction of bilingual terminology lexicons. *Terminology*, 4(2), 1999.

15. David Hull and Gregory Grefenstette. Querying across languages: a dictionary-based apporach to multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996.
16. Pei Ling Lai and Colin Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(5):365–377, 2000.
17. M. Littman, S. Dumais, and K. Landauer. *Cross-Language Information Retrieval*, chapter Automatic Cross-Language Information Retrieval Using Latent Semantic Indexing. 1998.
18. J.-Y. Nie, M. Simard, Isabelle P., and Durand R. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*, 1999.
19. C. Peters and E. Picchi. Capturing the comparable: A system for querying comparable text corpora. In *JADT'95 - 3rd International Conference on Statistical Analysis of Textual Data*, pages 255–262, 1995.
20. R. Rapp. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the European Association for Computational Linguistics*, 1999.
21. Jacques Savoy. Report on clef-2002 experiments: Combining multiple sources of evidence. In *Proceedings of CLEF 2002*, 2002.
22. I Shahzad, K. Ohtake, S. Masuyama, and K. Yamamoto. Identifying translations of compound nouns using non-aligned corpora. In *Proceedings of the Workshop MAL'99*, pages pp. 108–113, 1999.
23. K. Tanaka and Hideya Iwasaki. Extraction of lexical translations from non-aligned corpora. In *International Conference on Computational Linguistics, COLING'96*, 1996.
24. Alexei Vinokourov, John Shawe-Taylor, and Nello Cristianini. Inferring a semantic representation of text via cross-language correlation analysis. In *Advances of Neural Information Processing Systems 15*, 2002.