XRCE's Participation to CLEF 2008 Ad-Hoc Track

Stephane Clinchant and Jean-Michel Renders
Xerox Research Centre Europe, 6 ch. de Maupertuis, 38240 Meylan, France
FirstName.LastName@xrce.xerox.com

Abstract

Our participation to CLEF2008 (Ad-Hoc Track, TEL Subtask) was an opportunity to develop and assess methods that tackle multilinguilality in a principled – while rather simple – way. It was also an opportunity to demonstrate the effectiveness of the dictionary adaptation method we designed last year in the case of the domain-specific track. Unfortunately, it turned out that several mistakes we accumulated in our implementation impacted significantly and negatively the performance of our submitted runs. We nevertheless decided to experiment extra runs, that we designed to (partially) compensate for the errors made in the official runs and whose performance are reported in this working note. These results are quite satisfying, as they reach (or exceed) the level of the other best participants for the bilingual tasks.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

Keywords

Cross Lingual Information Retrieval, Lexicon Extraction, Query Translation and Disambiguation, Dictionary Adaptation

1 Introduction

This article describes our participation to the Ad-Hoc Track (TEL Subtask). Our very first motivation was to try to tackle multilinguality in a principled way: this is the object of the next section. Then, we explain the general methodological steps that we followed in our runs. A specific section is devoted to the analysis of the performances and the mistakes of our official runs. Indeed, it appeared after the publication of the results that we accumulated several "bugs" (or errors) that significantly impacted the performance of our methods, so that these are not directly comparable with the other ones. Still, in order to be constructive, we took some actions in order to compensate for these errors after the submission and we present in the last section of this note some new results achieved by runs taking inspiration from the dictionary adaptation algorithm that we proposed last year [2].

2 Dealing with Multilingual Documents

The framework of our retrieval experiments is the Language Model approach to Information Retrieval [4]. The TEL collections are clearly multilingual: a document can be described by French

words in a field and in German in an other field. Following the language modelling approach, we decide not to split a document into parts according to the language: a document is a sequence of tokens, which may be of any language; accordingly, a single language model is associated to the document, which is a probability distribution over the words (actually lemma's) of three concatenated vocabularies (English, French and German). In the following, this concatenation of vocabularies will be called the "meta-language". Thus, the feature space of different languages is aggregated into a single description space. This way, we do not build different indexes for a collection (according to the identified languages) but a single index is built containing all the languages.

However, building a single index to cope with multilinguality is just halfway to the solution, as the query is in general expressed only in one language. Indeed, since collections are multilingual, a query word need to be translated into the "meta-language", including its original language. This is done by building probabilistic meta-dictionaries (from a single source language to the meta-language). To be more concrete, here is a simplified excerpt of a probabilistic meta-dictionary we used:

```
roman(English) Latein(German) 0.02
roman(English) roman(English) 0.8
roman(English) antiqua(German) 0.01
roman(English) lateinisch(German) 0.02
roman(English) roemisch(German) 0.05
roman(English) romain(French) 0.1
Gauguin(English) Gauguin(English) 0.8
Gauguin(English) Gauguin(German) 0.1
Gauguin(English) Gauguin(French) 0.1
```

This probabilistic dictionary is built as a combination of a monolingual resource (thesaurus) and bilingual lexicons extracted from parallel corpora (in our case, the JRC-AC corpus[1]) and completed by approximate string matching equivalences (for lemmas not covered by the JRC-AC corpus). An important issue is how to weight the different translation probabilities when we merge the monolingual thesauri and the pair-wise bilingual dictionaries. We have chosen to merge them linearly. We believe that those linear weights should depend on the target collection and the task given. A natural choice, that we propose, is to give more weight to the official language of the target collection (French for BNF, German for ONB and English for BL). Formally, suppose that we are targeting the BL collection (whose official language is English), then the value $P(E_j|E_i)$ that represents the fact that English word E_j will be used as substitute (synonym) for E_i, will be weighted by α (typically, α=0.8); the value $P(F_j|E_i)$ that represents the fact that French word F_j will be used as substitute (translation) for E_i, will be weighted by $1 - \alpha/2$ and similarly for the entry $P(G_j|E_i)$. Note that, as $P(E_j|E_i)$, $P(F_j|E_i)$ and $P(G_j|E_i)$ individually sum up to 1 (over j) for a given E_i, the new probabilities also sum up to 1.

Once the meta-dictionary is built from these standard monolingual and bilingual resources, we propose to adapt it for a specific (query, target collection) pair, following the method we presented last year [2]. This amounts to filter out irrelevant, spurious meta-translations, as well as increasing the probabilities of more coherent word translations or synonyms.

3 Pre-processing and global approach

We have participated to all 'monolingual' and 'bilingual' tasks. None of the tasks were truly monolingual or bilingual, which motivated our method to cope with multilinguality.

For the 3 main languages (English, German, French), we used our home-made lemmatiser and word-segmenter (decompounder) for German. From the fields available for a document record, we only kept the title as well as the subject fields. Classical stopword removal was performed. As

[1] Available on *http://wt.jrc.it/lt/Acquis/*

Table 1: (Lost) Relevant Documents for each collection

Collection	# of relevant documents	# of relevant documents not indexed
BL	2533	240
BNF	1339	108
ONB	1637	69

monolingual resource, we used the Open Office thesauri[2]. As multilingual resources, we used a probabilistic dictionary, called ELRAC, that is a combination of a very standard one (ELRA) and a lexicon automatically extracted from the parallel JRC-AC (Acquis Communautaire) Corpus. Finally, we carried out our experiments relying on the Lemur Toolkit [1].

All our runs consisted in the following methodological steps:

- meta-translating the query with the multilingual meta-dictionary,
- adapting the meta-dictionary during a first pseudo-feedback step (details of this are given later),
- and finally applying another classical (monolingual) pseudo-feedback step.

4 Mistakes in the submitted runs

In this section, we present the analysis of the mistakes we did in our official runs.

The first one stemmed from a misunderstanding of what is considered as "bilingual" in the TEL task. When we preprocessed documents, we made the wrong hypothesis that *only* documents whose language is either French, English or German should be kept. As a consequence, we did not index documents whose title and content are indicated to belong to another language (Italian, Spanish, ...), even if they had a subject field in one of the three main languages. Te post analysis shows that we lost a significant number of relevant documents at indexing time, with respect to the given queries. Table 1 shows for each collection the count of relevant documents we lost at indexing time with respect to the total number of relevant documents.

The second error we made was to weight more the source language instead of the target language through the α parameter when building the meta-dictionary, i.e. we built one meta-dictionary per possible query (source) language giving more weight to this source language, instead of building one meta-dictionary per collection giving more weight to the official language of the collection.

Last, but not least, the third mistake we did, happened when we meta-translated the queries. Recall that we need to translate a query even in the 'monolingual' setting to address the fact that the collections are multilingual. We used a mixture model to achieve this effect:

$$P(w|q) = \beta P_0(w|q) + (1-\beta) \sum_{q_j \in q} P(w|q_j) P(q_j|q) \quad (1)$$

where $P(w|q_j)$ is given by our meta-dictionary and $P_0(w|q)$ is the initial language model of the query (obtained by maximum-likelihood estimation, with non-null values only for words of the source language). The β parameter controls the "weight of meta-translation" given to other languages and to a thesaurus (if any). In the scenario of 'monolingual' runs, we kept the β values high (between 0.8 and 0.9). The mistake we did in our 'bilingual' runs was to forget to change this β value to smaller values (between 0 and 0.2) in order to have a real effect of translation.

All these factors explain why our runs performed relatively poorly. In the last section (before conclusion), we briefly present some new runs and their results, that partially compensate for these errors. Before this, for the sake of completeness, we describe our dictionary adaptation method, that was already used last year (in the domain-specific track).

[2]Available on *http://wiki.services.openoffice.org/wiki/Dictionaries*

5 Dictionary Adaptation

We briefly recall the model underlying our dictionary adaptation method [2]. As already mentioned, the Language Modelling approach to information retrieval was adopted for our experiments. Crosslingual retrieval models translate the query into a query language model in the target language [3]. Then a monolingual search is performed, using a ranking criterion such as the Cross-Entropy:

$$CE(q_s|d_t) = \sum_{w_t,w_s} P(w_t|w_s)P(w_s|q_s)\log P(w_t|d_t) \tag{2}$$

The main idea of dictionary adaptation is to be able to adapt the entries of a dictionary to a query and a target corpus. Formally, let $q_s = (w_{s1}, \ldots, w_{sl})$ be the query in source language. Our input data are an initial source query language model $p(w_s|q_s)$ and a first dictionary $p(w_t|w_s)$. First of all, the source query is translated with all dictionaries entries. Then, we select the top n documents (pseudo-relevance feedback) and we model the set of feedback documents $\mathbf{F}$ with a generative model from which we learn a new dictionary θ_{st} : we see each document as the outcome of a multinomial random variable. First, the likelihood of the pseudo-feedback set can be written:

$$P(\mathbf{F}|\theta) = \prod_k \prod_{w_t} \left(\lambda(\sum_{w_s} \theta_{st} p(w_s|q_s)) + (1-\lambda)P(w_t|\mathcal{C})\right)^{c(w_t,d_k)} \tag{3}$$

As described in [2], the new dictionary θ_{st} can be learned by EM and a new query can be generated by using all entries in the adapted dictionary.

In all experiments reported in this note, the value of n was chosen as 50.

6 Unofficial Runs

We performed a set of extra runs, with the aim to be comparable with the results of other participants and to compensate for the effects of the mistakes and bugs we identified. In order to get rid of the issue of weighting more one language with respect to the other ones (selection of the α and β parameters) – things that we did in a completely erroneous way in our official runs –, we decided to make a simplifying assumption, namely that 'bilingual runs' are considered as really bilingual, with known source and target languages. In other words, we considered only the French part of BNF, the English part of BL and the German part of ONB and used purely bilingual dictionaries (which were subsequently adapted). A post-analysis on relevant documents shows that this assumption is not unreasonable:

```
For the BL collection:
number of relevant documents entirely in German : 24
number of relevant documents in English and German :  78
number of relevant documents entirely in French : 4
number of relevant documents completely in English : 2066
number of relevant documents in French and English : 122

For the BNF collection:
number of relevant documents entirely in German : 2
number of relevant documents in French and German :  11
number of relevant documents entirely in French : 1008
number of relevant documents completely in English : 12
number of relevant documents in French and English : 198

For the ONB collection:
number of relevant documents entirely in German : 1241
number of relevant documents in French and German :  29
```

Table 2: Dictionary Adaptation Experimental Results in Mean Average Precision - (1) refers to the unrestricted collection, while (2) refers to the indexed collection

Translation	Initial Dictionary	W/O adapt.(1)	W/ adapt.(1)	W/O adapt.(2)	W adapt.(2)
EN to BNF	English To French	22.00	25.75	24.06	28.58
DE to BNF	German To French	22.66	24.60	25.20	27.39
FR to BL	French To English	24.83	28.76	27.75	32.32
DE to BL	German To English	23.61	26.49	26.95	29.88
EN to ONB	English To German	20.78	23.00	23.00	25.28
FR to ONB	French To German	23.19	24.78	25.29	27.14

```
number of relevant documents entirely in French : 0
number of relevant documents completely in English : 37
number of relevant documents in German and English : 261
```

In order to compensate for the forgetting of documents in the index (documents whose title/content is not in French, German nor English), we simply removed non-indexed documents from the relevance assessment lists.

Table 2 shows the corrected runs using the dictionary adaptation using total translation ($\beta = 0$ in equation 1). The second column of the table shows the source and target languages we used for the runs. Our runs could achieve better results if we took into account the other languages and if we performed an additional step of classical pseudo-feedback, but this is left for further experiments. Results are given without and after adaptation. For completeness, we also give the results on the unrestricted relevance list (columns 3 and 4), while the MAP corresponding to the restricted collection (documents whose title/content is not in French, German nor English are removed from the relevance assessment lists) are given in columns 5 and 6.

Assuming that the documents we removed from the collection are completely random with respect to the queries and that there are no performance bias due to the nature of the removed documents, we can expect from the results given in columns 5 and 6 to be comparable with the performance of other participants. These results are very encouraging, as they first show clearly the beneficial effect of dictionary adaptation and by the fact that we achieve results more or less equivalent to the best results of the other participants (to be more precise, we are just behind the best one for the BL as target collection, and better than the first one for the ONB and BNF collections).

7 Conclusion

Our work was concerned about dealing with multilinguality in a principled way. Our goal was to get a single retrieval model and index for all the languages of one specific collection. However, this approach required to give weights to each language to merge dictionaries at retrieval time. While assigning such weights requires prior knowledge about the collections, the dictionary adaptation mechanism provides a partial solution to this problem, adapting weights to each query. This year, the accumulation of some mistakes rendered our official runs relatively inefficient. We presented the reasons of these mistakes and corrected partly some of them in a set of extra unofficial runs whose performances are among the best ones; they demonstrated that dictionary adaptation is effective for the TEL task and corpora. Further work will require re-processing the collections to keep the document we lost. We will also need to come back to a true multilingual setting by solving the issue of weighting differently the basic bilingual lexicons and monolingual thesauri, according to the target collection.

Aknowledgments

This work was partly supported by the IST Programme of the European Community, under the SMART project, FP6-IST-2005-033917.

References

[1] http://www.lemurproject.org/.

[2] S. Clinchant and J.-M. Renders. Xrce's participation to clef 2007 - domain specific track. In *Working Notes of CLEF 2007. Avalaible On-line on the CLEF Web Site*, 2007.

[3] W. Kraaij, J.-Y. Nie, and M. Simard. Embedding web-based statistical translation models in cross-language information retrieval. *Comput. Linguist.*, 29(3):381–419, 2003.

[4] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc to information retrieval. In *Proceedings of SIGIR'01*, pages 334–342. ACM, 2001.