CLEF 2001 Bilingual Task
Simple Dictionary-Based Query Translation

Marine Carpuat and Pascale Fung
HKUST
Human Language Technology Center
Department of Electrical and Electronic Engineering
University of Science and Technology
Clear Water Bay, Hong Kong
{eemarine,pascale}@ee.ust.hk

Abstract

In this paper, we describe our approach to the French English Bilingual Task in CLEF 2001. A simple dictionary-based method is used to translate the French query into a bag of weighted English words, the English query, which is submitted to the SMART retrieval engine. Despite the simplicity of the method, the results happen to be reasonable.

1 Introduction

For our first participation to CLEF, our goal was to evaluate what level of performance can be expected from a simple CLIR system, which does not make use of sophisticated resources. We decide to work on a fully automatic method, based on freely available resources. We participated to the French-English bilingual task and present the results obtained using a simple dictionary-based query translation method.

2 Method

2.1 Dictionary-Based Keyword Translation

The easiest and most common way to cross the language boundary for CLIR is to translate the original query in the target language and to input this new query to a monolingual IR Sytem. In this work, French queries are simply translated into English by dictionary look-up. Our "bilingual dictionary" is in fact a bilingual word list generated from a free online Bilingual French-English dictionary[3]. The lists consists of around 200,000 entries, but many different forms are listed for each French word (singular/plural, male/female, tenses). The number of terms actually represented is therefore much lower. Multiple translations are all included according to the online dictionary.
This naive construction of the bilingual lexicon allows us to perform very little pre-processing on the French queries. We only focus on keyword selection. In order to translate meaningful terms, we try to remove as many stopwords or function words as possible. We chose a basic stoplist [5] and augmented it.

CLEF2000 French topics are used to identify function words that are used to express the information need but do not carry information themselves. These words typically occur in a large number of queries with a low frequency in each query, whereas content-bearing words occur in fewer queries with a high frequency in each query. Potential function words are therefore easily selected with tf.idf information.
The English query is a simple bag of words, taking into account all the possible translation of the keywords given by the dictionary. If no translation candidate can be found in the dictionary, the French words are directly copied and pasted in the English query. For closely related languages such as French and English, keeping the original word in the source language is helpful, since the missing translations happen to be proper nouns, English words included in French topics or cognates.

2.2 Text Retrieval Engine

The Text Retrieval Engine used in this work is the well-known IR sytem Smart[2][6]. It is well adapted to index and query a large corpus such as the CLEF Los Angeles Times corpus. Smart implements the vector space model, in which queries and documents are represented by vectors containing tf.idf weights. English queries consist of all the translation candidates found for the French topics keywords, weighted by the frequency of the original French keyword and the translation probability when multiple candidayes are found. We assume these translation probabilities are uniform.

3 Results

Our experiment was conducted using the title and description fields of the topics only. We obtain average results, which are reasonable given the simplicity of the method used. The majority of unique translation candidates given by our lexicon happen to be correct. When multiple translations are proposed, co-occurence with the translation of other keywords performs a simple disambiguation. Not surprisingly however, when the number of translation candidates grows larger and when the terms are more ambiguous, the query results drop below the average.

Recall	Precision
0.00	0.5634
0.10	0.4996
0.20	0.4388
0.30	0.3630
0.40	0.3040
0.50	0.3234
0.60	0.2010
0.70	0.1743
0.80	0.1554
0.90	0.1302
1.00	0.0932

Table 1: Interpolated Recall-Precision Averages

Improvements could be achieved by adding a statistical disambiguation method[1] and making use of a more sophisticated translation weighting scheme.

4 Conclusion

In summary, our simple dictionary-based keyword translation method performs reasonably well with a simple lexicon. Of course, query expansion and disambiguation methods are much needed to improve these results. But this shows that even a simple lexicon based on the most frequent translations is a reasonable basis for Bilingual Information Retrieval.

References

[1] Ballesteros, L. and Croft, B. (1997). *Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval.* In: Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR) 20th International Conference on Research and Development in Information Retrieval; 1997 July 25-31; Philadelphia, PA. New York, NY: ACM, 1997. 84-91.

[2] Buckley, C. 1985 *Implementation of the smart information retrieval system*, Technical Report 85-686, Computer Science Department, Cornell University, Ithaca, New York, May 1985.

[3] English-French bilingual dictionary. Available: http://sun-recomgen.univ-rennes1.fr/FR-Eng.html

[4] Grefenstette, G. (1998). The Problem of cross-language information retrieval. In G.Grefenstette (Ed.) Cross Language Information retrieval. pp 1-9. Kluwer Academic Publishers.

[5] Veronis, J. Un Antidictionnaire. Available: http://www.up.univ-mrs.fr/~veronis

[6] Smart. Available: ftp://ftp.cs.cornell.edu/pub/smart.