Ambiguity and Unknown Term Translation in CLIR

Dong Zhou[1], Mark Truran[2], and Tim Brailsford[1]

1. School of Computer Science and IT, University of Nottingham, United Kingdom
2. School of Computing, University of Teesside, United Kingdom
dxz@cs.nott.ac.uk, M.A.Truran@tees.ac.uk, tjb@cs.nott.ac.uk

Abstract. In this paper we present a report on our participation in the CLEF Chinese-English *ad hoc* bilingual track, and we discuss a disambiguation strategy which employs a modified co-occurrence model to determine the most appropriate translation for a given query. This strategy is used alongside a pattern-based translation extraction method which addresses the 'unknown term' translation problem. Experimental results demonstrate that a combination of these two techniques substantially improves retrieval effectiveness when compared to various baseline systems that employ basic co-occurrence measures or make no provision for out-of-vocabulary terms.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing ~ *dictionaries, linguistic processing, thesauruses*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.7 [**Pattern Recognition**]: Applications ~ *text processing*.

General Terms

Algorithm, Experimentation, Performance

Keywords

Disambiguation, Co-Occurrence, Unknown Term Detection, Patterns

1. Introduction

Our participation in the current CLEF *ad hoc* bilingual track is motivated by a desire to test two newly developed CLIR techniques. The first of these concerns the resolution of ***translation ambiguity***, which is a classic problem of cross language information retrieval. Translation ambiguity is a difficulty that will inevitably occur when attempting to translate a multi-term query using a bilingual dictionary. This problem stems from choice, because a typical bilingual dictionary will provide a set of alternative translations for each term within the given query. Choosing the correct translation of each term is a difficult procedure, but it is also critical to the efficiency of any related retrieval functions. Previous solutions to this problem have employed co-occurrence information extracted from document collections to aid the process of resolving translation-based ambiguities[1, 2]. In the following experiment we use a disambiguation strategy which extends this basic approach. Our technique uses a novel graph-based analysis to determine the most appropriate translation for a given query.

The second technique we wish to test addresses the ***coverage problem***. This refers to the limited linguistic scope of parallel texts and dictionaries. Certain types of words are not commonly found in either of these types of resources, and it is these out-of-vocabulary (OOV) terms that will cause difficulties during automatic translation. Previous work on the problem of unknown terms has tended to concentrate upon complex statistical solutions[3, 4]. In this experiment we will be using a new approach to OOV terms which extracts translation candidates from mixed language text using linguistic and punctuative patterns [7].

The purpose of this paper is to examine the effect of combining these two techniques in the hope that operating them concurrently, would improve the efficacy of a cross language retrieval engine.

2. Methodology

2.1 Resolution of Translation Ambiguities

The rationale behind the use of co-occurrence data to resolve translation ambiguities is that for any query containing multiple terms which must be translated, the correct translations of individual query terms will tend to co-occur as part of a given sub-language, while the incorrect translations of individual query terms will not. Ideally, for each query term under consideration, we would like to choose the best translation that is consistent with the translations selected for all remaining query terms. However, this process of inter-term optimization has proved computationally complex for even the shortest of queries. A common workaround, used by several researchers working on this particular problem[5], involves use of an alternative resource-intensive algorithm, but this too has problems. In particular, it has been noted that the selection of translation terms is isolated and does not differentiate correct translations from incorrect ones[5].

We approached this problem from a different direction. The co-occurrence of possible translation terms within a given corpus may be viewd as a graph. Each translation candidate of a source query term may then be represented by a single node in that graph. Edges drawn between these nodes are then weighted according to a particular co-occurrence measurement. We use a graph-based analysis (inspired by research into hypermedia retrieval [6]) to determine the importance of a single node using global information recursively drawn from the entire graph. The importance of a node is then used to guide query term translation.

2.2 Resolution of Unknown Terms

Our approach to the resolution of unknown terms is documented in detail elsewhere [7]. Stated succinctly, translations of unknown terms are obtained from a computationally inexpensive pattern-based processing of mixed language text.

3. Experiment

3.1 Experimental Setup

In our experiment we used the English LA Times 2002 collection[1]. All of the documents were indexed using the Lemur toolkit[2]. Prior to indexing, Porter’s stemmer was used to remove stop words from the

[1] http://www.clef-campaign.org/
[2] http://www.lemurproject.org

English documents[8]. A Chinese-English dictionary is adopted in our experiment from the web[3].

In order to investigate the effectiveness of our various techniques, we performed a simple retrieval experiment with several key permutations. These variations are as follows:

MONO (***monolingual***): This part of the experiment involved retrieving documents using manually translated versions of English queries. The performance of a monolingual retrieval system such as this has always been considered as an unreachable 'upper-bound' of CLIR as the process of automatic translation is inherently noisy.

ALLTRANS (***all translations***): Here we retrieved documents from the two test collections using all the translations provided by the respective dictionaries for each query term.

FIRSTONE (***first translations***): This involved retrieving documents from the test collections using only the first translation suggested for each query term by the bilingual dictionaries. Due to the way in which these bilingual dictionaries are constructed, the first translation for any word generally equates to the most frequent translation for that term according to the World Wide Web.

COM (***co-occurrence translation***): In this part of the experiment, the translations for each query term were selected using the basic co-occurrence algorithm described in [2]. We used the target document collection to calculate the co-occurrence scorings.

GCONW (***weighted graph analysis***): Here we retrieved documents from the collections using query translations suggested by our analysis of a weighted co-occurrence graph. Edges of the graph were weighted using co-occurrence scores derived using [2].

GCONUW (**un*weighted graph analysis***): As above, we retrieved documents from the collections using query translations suggested by our analysis of the co-occurrence graph, only this time we used an *unweighted graph*.

GCONW+OOV (***weighted graph analysis with unknown term translation***): As GCONW, except that query terms that were not recognized were sent to the unknown term translation system.

GCONUW+OOV (**un*weighted graph analysis with unknown term translation***): As above, using unweighted scheme this time.

3.2 Experimental Results

The results of this experiment are provided in TABLES 1 and 2. Document retrieval with no disambiguation of the candidate translations (ALLTRANS) was consistently the lowest performer in terms of mean average precision. This result was not surprising and merely confirms the need for an efficient process for resolving translation ambiguities. The improvement in performance when switching from ALLTRANS to the FIRSTONE method was variable across the two test collections. When the translation for each query term was selected using a basic co-occurrence model (COM)[2], retrieval effectiveness always outperformed ALLTRANS and FIRSTONE. Graph based analysis outperformed the basic co-occurrence model in short queries but not in long queries, this is probably due to the dictionary we used. The combined model (with OOV term translation) scored highest in terms of mean average precision when compared to non-monolingual systems.

4. Conclusions

[3] http://www.ldc.upenn.edu/

In this paper we have described our contribution to the CLEF Chinese-English *ad hoc* track. We have used a modified co-occurrence model for the resolution of translation ambiguity, and this technique has been combined with a pattern-based method for the translation of OOV terms. The combination of these two methodologies fared well in our experiment, outperforming various baseline systems, and the results that we have obtained thus far suggest that these techniques are far more effective combined than on their own.

The Use of the CLEF document collections during this experiment has led to some interesting observations. There seems to be a distinct difference between the collection and the TREC alternatives commonly used by researchers in this field. Historically, the use of co-occurrence information to aid disambiguation has led to disappointing results on TREC retrieval runs[5]. Future work is currently being planned that will involve a side by side examination of the TREC and CLEF document sets in relation to the problems of translation ambiguity.

TABLE 1. Short query results (*title*) in CLEF

	MAP	**R-Prec**	**P@10**	**% of MONO**	**IMPR over ALLTRANS**	**IMPR over FIRSTONE**	**IMPR over COM**
MONO	0.4078	0.4019	0.486	N/A	N/A	N/A	N/A
ALLTRANS	0.2567	0.2558	0.304	62.95%	N/A	N/A	N/A
FIRSTONE	0.2638	0.2555	0.284	64.69%	2.77%	N/A	N/A
COM	0.2645	0.2617	0.306	64.86%	3.04%	0.27%	N/A
GCONW	0.2645	0.2617	0.306	64.86%	3.04%	0.27%	0.00%
GCONW+OOV	0.3337	0.3258	0.384	81.83%	30.00%	26.50%	26.16%
GCONUW	0.2711	0.2619	0.294	66.48%	5.61%	2.77%	2.50%
GCONUW+OOV	0.342	0.3296	0.368	83.86%	33.23%	29.64%	29.30%

TABLE 2. Long query results (*title+description*) in CLEF

	MAP	**R-Prec**	**P@10**	**% of MONO**	**IMPR over ALLTRANS**	**IMPR over FIRSTONE**	**IMPR over COM**
MONO	0.3753	0.3806	0.43	N/A	N/A	N/A	N/A
ALLTRANS	0.2671	0.2778	0.346	71.17%	N/A	N/A	N/A
FIRSTONE	0.2516	0.2595	0.286	67.04%	-5.80%	N/A	N/A
COM	0.2748	0.2784	0.322	73.22%	2.88%	9.22%	N/A
GCONW	0.2748	0.2784	0.322	73.22%	2.88%	9.22%	0.00%
GCONW+OOV	0.3456	0.3489	0.4	92.09%	29.39%	37.36%	25.76%
GCONUW	0.2606	0.2714	0.286	69.44%	-2.43%	3.58%	-5.17%
GCONUW+OOV	0.3279	0.3302	0.358	87.37%	22.76%	30.33%	19.32%

5. References:

1. Ballesteros, L. and W.B. Croft, *Resolving ambiguity for cross-language retrieval*, in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 1998, ACM Press: Melbourne, Australia. p. 64-71.
2. Jang, M.-G., S.H. Myaeng, and S.Y. Park, *Using mutual information to resolve query translation ambiguities and query term weighting*, in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. 1999, Association for Computational Linguistics: College Park, Maryland. p. 223-229.
3. Cheng, P.-J., et al., *Translating unknown queries with web corpora for cross-language information retrieval*, in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 2004, ACM Press: Sheffield, United Kingdom. p. 146-153.
4. Zhang, Y. and P. Vines, *Using the web for automated translation extraction in cross-language information retrieval*, in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 2004, ACM Press: Sheffield, United Kingdom. p. 162-169.
5. Gao, J. and J.-Y. Nie, *A study of statistical models for query translation: finding a good unit of translation*, in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 2006, ACM Press: Seattle, Washington, USA. p. 194-201.
6. Brin, S. and L. Page, *The anatomy of a large-scale hypertextual Web search engine.* Comput. Netw. ISDN Syst., 1998. **30**(1-7): p. 107-117.
7. Zhou, D., Truran, M., Brailsford, T. and Ashman, H, *NTCIR-6 Experiments using Pattern Matched Translation Extraction*, in *the sixth NTCIR workshop meeting*. 2007, NII: Tokyo, Japan. p. 145-151.
8. Porter, M.F., *An algorithm for suffix stripping*. Program, 1980. **14**: p. 130-137.