Disambiguation and Unknown Term Translation in Cross Language Information Retrieval

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

[1] School of Computer Science, 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 2007 Chinese-English *ad hoc* bilingual track. 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 with no provision for out-of-vocabulary terms.

1 Introduction

Our participation in the CLEF 2007 *ad hoc* bilingual track was motivated by a desire to test and integrate two newly developed cross-language information retrieval (CLIR) techniques. The first of these techniques addresses the correct translation of *ambiguous* terms. A typical bilingual dictionary will provide a set of alternative translations for every term occurring within a given query. Choosing the correct translation for each term is a difficult procedure 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], [3], [4]. Extending this approach, we have developed a disambiguation strategy which employs a novel graph-based analysis of co-occurrence information to determine the most appropriate translations.

The second cross language IR technique we have developed addresses the *coverage problem*. Certain types of words are not commonly found in parallel texts or dictionaries, 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 [5]. We have adopted a much simpler approach to this problem which centres upon the application of linguistic and punctuative patterns to mixed language text [6].

Overall, the purpose of this paper is to evaluate a retrieval system which *combines* these two specific techniques in order to examine the effect of operating them concurrently.

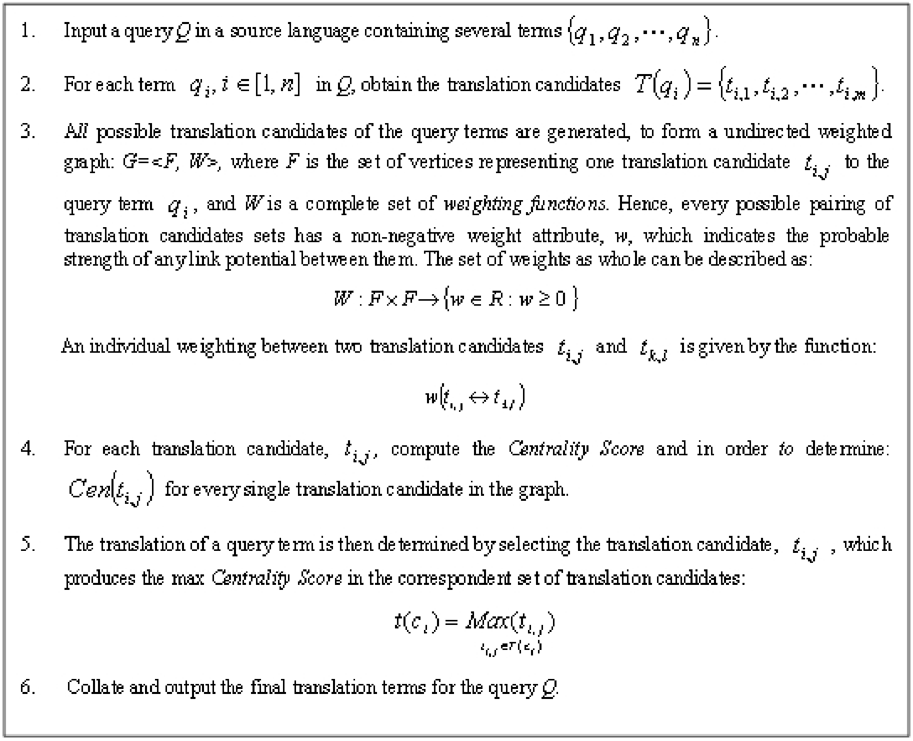
C. Peters et al. (Eds.): CLEF 2007, LNCS 5152, pp. 64–71, 2008.

2 Methodology

2.1 Resolution of Translation Ambiguities

The rationale behind the use of co-occurrence data to resolve translation ambiguities is relatively simple. 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, 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 [2], [4].

We approached this problem from a different direction. First of all, we viewed the co-occurrence of possible translation terms within a given corpus as *a graph*.



1. Input a query *Q* in a source language containing several terms $\{q_1, q_2, \cdots, q_n\}$.
2. For each term $q_i, i \in [1, n]$ in *Q*, obtain the translation candidates $T(q_i) = \{t_{i,1}, t_{i,2}, \cdots, t_{i,m}\}$.
3. *All* possible translation candidates of the query terms are generated, to form a undirected weighted graph: *G=<F, W>*, where *F* is the set of vertices representing one translation candidate $t_{i,j}$ to the query term q_i, and *W* is a complete set of *weighting functions*. Hence, every possible pairing of translation candidates sets has a non-negative weight attribute, *w*, which indicates the probable strength of any link potential between them. The set of weights as whole can be described as:

$$W : F \times F \rightarrow \{ w \in R : w \geq 0 \}$$

An individual weighting between two translation candidates $t_{i,j}$ and $t_{k,l}$ is given by the function:

$$w(t_{i,j} \leftrightarrow t_{k,l})$$

4. For each translation candidate, $t_{i,j}$, compute the *Centrality Score* and in order *to* determine: $Cen(t_{i,j})$ for every single translation candidate in the graph.
5. The translation of a query term is then determined by selecting the translation candidate, $t_{i,j}$, which produces the max *Centrality Score* in the correspondent set of translation candidates:

$$t(c_i) = \underset{t_{i,j} \in T(c_i)}{Max}(t_{i,j})$$

6. Collate and output the final translation terms for the query *Q*.

Fig. 1. A graph-based algorithm for the disambiguation of query terms

In this graph, each translation candidate of a source query term is represented by a single node. Edges drawn between these nodes are weighted according to a particular co-occurrence measurement. We then apply graph-based analysis (inspired by research into hypermedia retrieval [7]) to determine the importance of a single node using global information recursively drawn from the entire graph. Subsequently, this measure of node importance is used to guide final query term translation.

The disambiguation algorithm we applied is summarized in Figure 1. The centrality score for a single vertex V_i in the graph is calculated in the following way: let $\{V_i\}_{in}$ be a set of nodes that point to V_i, and let $\{V_i\}_{out}$ be a set of nodes that V_i points at. Then, the centrality score of V_i is defined as follows:

$$Cen(V_i) = (1-d)/N + d \times \sum_{j \in \{V_i\}_{in}} \frac{w_{i,j}}{\sum_{V_k \in \{V_j\}_{out}} w_{j,k}} Cen(V_j) \tag{1}$$

Where d is a dampening factor which integrates the probability of jumping from one node to another at random (normally set to 0.85) and N is the total number of nodes in the graph.

Two variations of the weighting function $w(t_{i,j} \leftrightarrow t_{k,l})$ have been developed. They are called *StrengthWeighting (SW)* and *FixedWeighting (FW)* respectively. The SW function should be considered an undirected weighted graph calculation while FW function is the undirected, unweighted alternative:

StrengthWeighting: If the similarity score (co-occurrence measurement) between two terms is more than zero, then the weight between the two terms is equal to the similarity score. Otherwise the weight is set to zero.

$$w(t_{i,j} \longleftrightarrow t_{k,l}) = \begin{cases} sim(t_{i,j}, t_{k,l}) & sim(t_{i,j}, t_{k,l}) > 0 \\ 0 & otherwise \end{cases}$$

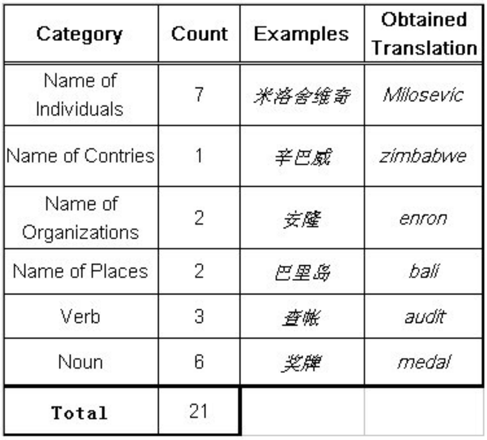
FixedWeighting: If the similarity score (co-occurrence measurement) between two terms is more than zero, then the weight between the two terms is equal to one. Otherwise the weight is set to zero.

$$w(t_{i,j} \longleftrightarrow t_{k,l}) = \begin{cases} 1 \; sim(t_{i,j}, t_{k,l}) > 0 \\ 0 \; otherwise \end{cases}$$

2.2 Resolution of Unknown Terms

Our approach to the resolution of unknown terms is documented in detail in [6]. Stated succinctly, translations of unknown terms are obtained from a computationally inexpensive pattern-based processing of mixed language text retrieved from the web. A high level summary of this web based translation technique for OOV terms is as follows:

- The OOV term is submitted to a web search engine, and the results are cached.
- The text of each resultant web page is analyzed and certain punctuative and linguistic patterns are detected semi-automatically.



Category	Count	Examples	Obtained Translation
Name of Individuals	7	米洛舍维奇	*Milosevic*
Name of Contries	1	辛巴威	*zimbabwe*
Name of Organizations	2	安隆	*enron*
Name of Places	2	巴里岛	*bali*
Verb	3	查帐	*audit*
Noun	6	奖牌	*medal*
Total	21		

Fig. 2. A breakdown of OOV terms found in the CLEF 2007 query set

- Analysis of detected patterns enables the identification of one or more translation candidates for the OOV term.
- A final translation for the OOV term is extracted from the list of candidates using a combination of extraction frequencies and pattern based weightings.

Illustrative examples of OOV terms translated using this approach can be found in Figure 2 (see also section 3):

3 Experiment

In the following section we describe our contribution to the CLEF 2007 *ad hoc* bilingual track. The document corpus used in our experiment was the English LA Times 2002 collection (135,153 English language documents) [8]. The queries we used were provided by the CLEF 2007 organizing committee and consisted of 50 multiple field topics written in Chinese, complete with relevance judgments.

3.1 Overview of the Experimental Process

A description of the CLIR process we adopted during this experiment is as follows: In the first step of the process we employed a naive bilingual dictionary to obtain a semi-translated query set[1]. We then used the graph-based technique described above to resolve translation ambiguities within this set (with the co-occurrence scores obtained from the target document corpus). Finally, OOV terms occurring within the query set were passed to our pattern matcher to obtain final translation candidates. The fully translated queries were then passed to a information retrieval engine to retrieve the final document results list.

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

To prepare the corpus for the retrieval process, all of the documents were indexed using the Lemur toolkit[2]. Prior to indexing, Porter's stemmer [9] and a list of stop words[3] were applied to the English documents.

3.2 Experimental Setup

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 from the test collection using Chinese queries manually translated into English by the CLEF 2007 organising committee. 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 test collection using *all* of the translations provided by the bilingual dictionary for each query term.

FIRSTONE (first translations): This part of the experiment involved retrieving documents from the test collection using only the *first* translation suggested for each query term by the bilingual dictionary. Due to the way in which bilingual dictionaries are usually 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 [3]. We used the target document collection to calculate the co-occurrence scorings.

GCONW (weighted graph analysis): Here we retrieved documents from the document collection using query translations suggested by our analysis of a weighted co-occurrence graph (i.e. we used the SW weighting function). Edges of the graph were weighted using co-occurrence scores derived using [3].

GCONUW (unweighted graph analysis): As above, we retrieved documents from the collection using query translations suggested by our analysis of the co-occurrence graph, only this time we used an unweighted graph (i.e. we used the FW weighting function).

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

GCONUW+OOV (unweighted graph analysis with unknown term translation): As above, only this time we used the unweighted scheme.

3.3 Experimental Results and Discussion

The results of this experiment are provided in Tables 1 and 2. Document retrieval with no disambiguation of the candidate translations (ALLTRANS) was

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

[3] ftp://ftp.cs.cornell.edu/pub/smart/

Table 1. Short query results (*Title field only*)

	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 fields*)

	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%

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.

When the translation for each query term was selected using a basic co-occurrence model (COM) [3], retrieval effectiveness always outperformed ALLTRANS and FIRSTONE. Interestingly, this result is inconsistent with earlier work published by [4] observing the opposite trend in the context of a TREC retrieval experiment.

Graph based analysis always outperformed the basic co-occurrence model (COM) on short query runs in both the weighted and un-weighted variants. However, COM scored higher than GCONW and GCONUW on runs with longer queries. This is probably due to the bilingual dictionary we selected for the experiment. The Chinese-English dictionary provided by LDC contains very few translation alternatives, thereby creating limited scope for ambiguity.

The combined model (graph based ambiguity resolution plus OOV term translation) scored highest in terms of mean average precision when compared to the non-monolingual systems. As illustrated by the data, improvement over the COM baseline is more pronounced for *Title* runs. This seems to reflect the length of the queries. The *Title* fields in the CLEF query sets are very short, and correspondingly any OOV query terms not successfully translated will have a greater impact on retrieval effectiveness.

With respect to the monolingual system, a combination of our two new methods performed exceptionally well. For example, in the long query run, a combination of GCONW+OOV achieved 92.09% of monolingual performance. This means that our CLIR system as a whole achieved the second highest score in the whole CLEF 2007 *ad hoc* bilingual track (when compared with participating CLIR systems attempting retrieval of English documents using non-English query sets). This achievement is perhaps more remarkable when it is considered that our CLIR system does not yet employ *query expansion*, a technique renowned for improving retrieval effectiveness.

There were 21 unknown terms in the CLEF 2007 query set. Most of these terms were proper nouns or acronyms. Our system successfully translated 16 of the OOV terms, meaning its suggestions perfectly matched the manual CLEF 2007 translations. In our opinion, a translation hit rate of 76.2% in return for a meagre expenditure of resources emphatically validates the use of linguistics patterns in this context.

The OOV terms which were not successfully translated (23.8%) may have been *out of date*. Our method collects all translation candidates from the contemporaneous web. The query terms we worked with are several years old. It could be that the persons, organisations or acronyms which are referred to in that query set are no longer as prominent on the web as they once were. This would inevitably have a negative impact on our ability to generate appropriate translation candidates.

4 Conclusions

In this paper we have described our contribution to the CLEF 2007 Chinese-English *ad hoc* bilingual track. Our experiment involved the use of a modified co-occurrence model for the resolution of translation ambiguities, and a pattern-based method for the translation of OOV terms. The combination of these two techniques fared well, outperforming various baseline systems. The results that we have obtained thus far suggest that these techniques are far more effective when combined than in isolation.

Use of the CLEF 2007 document collection during this experiment has led to some interesting observations. There seems to be a distinct difference between this 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 [4]. 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.

Acknowledgments. Thanks to people in WebTech Group in the University of Nottingham for many useful discussions. The work here was partially funded by a scholarship from the University of Nottingham and the Hunan University.

References

1. Ballesteros, L., Croft, W.B.: Resolving ambiguity for cross-language retrieval. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia, pp. 64–71. ACM Press, New York (1998)
2. Gao, J., Nie, J.Y.: 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, Seattle, Washington, USA, pp. 194–201. ACM Press, New York (2006)
3. Jang, M.G., Myaeng, S.H., Park, S.Y.: 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, College Park, Maryland, pp. 223–229. Association for Computational Linguistics (1999)
4. Liu, Y., Jin, R., Chai, J.Y.: A maximum coherence model for dictionary-based cross-language information retrieval. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, Salvador, Brazil, pp. 536–543. ACM Press, New York (2005)
5. Zhang, Y., Vines, P.: 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, Sheffield, United Kingdom, pp. 162–169. ACM Press, New York (2004)
6. Zhou, D., Truran, M., Brailsford, T., Ashman, H.: Ntcir-6 experiments using pattern matched translation extraction. In: The sixth NTCIR workshop meeting, Tokyo, Japan, NII, pp. 145–151 (2007)
7. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Ashman, H., Thistlewaite, P. (eds.) Proceedings of the 7th International World Wide Web Conference, vol. 30(1-7), pp. 107–117 (1998); reprinted In: Ashman, H., Thistlewaite, P.(eds.): Comput. Netw. ISDN Syst. 30(1-7), 107–117 (1998) 297827
8. Di Nunzio, G., Ferro, N., Mandl, T., Peters, C.: Clef 2007 ad hoc track overview. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 13–32. Springer, Heidelberg (2008)
9. Porter, M.F.: An algorithm for suffix stripping. Program 14, 130–137 (1980)