CINDOR TREC-9 English-Chinese Evaluation

Miguel E. Ruiz, Steve Rowe,
Maurice Forrester, Páraic Sheridan

MNIS-TextWise Labs
401 South Salina Street
Syracuse, NY 13202

mruiz@textwise.com
paraic@textwise.com

Abstract

MNIS-TextWise Labs participated in the TREC-9 Chinese Cross-Language Information Retrieval track. The focus of our research for this participation has been on rapidly adding Chinese capabilities to CINDOR using tools for automatically generating a Chinese Conceptual Interlingua from existing lexical resources. For the TREC-9 evaluation we also built a version of our system which loosely integrates the CINDOR Conceptual Language Analysis process with the SMART retrieval system. This was motivated by the conclusions of our TREC-8 experiments which pointed to sub-standard retrieval based on the underlying retrieval algorithm. This integrated system has further allowed us to experiment with a range of approaches for cross-language retrieval, some specific to Chinese, which we have used in combination for our official TREC submissions. For evaluation, we submitted a monolingual Chinese run and a cross-language English-Chinese run. Analysis of results to date allow us to conclude that the automatically generated Conceptual Interlingua helps to improve performance in both cross-language and monolingual retrieval.

1. Introduction

The CINDOR (Conceptual Interlingua Document Retrieval) project at MNIS-TextWise Labs is pursuing a 'Conceptual Interlingua' approach to cross-language information retrieval, based on a conceptual lexical resource modeled around WordNet [Miller 1990]. The current version of CINDOR supports cross-language retrieval in any combinations of English, Spanish, French, Italian, German, and Japanese (and now Chinese). For our TREC-9 participation we concentrated our efforts in rapidly adding Chinese capabilities to CINDOR and building tools that allow automatic generation of a (Chinese) Conceptual Interlingua. Our approach is based on automatically mapping Chinese terminology into English WordNet concepts using existing bilingual dictionaries and corpora.

This paper presents an overview of each stage of our research leading up to the submission of TREC-9 runs. It includes a brief introduction to the CINDOR approach to cross-language retrieval in Section 2, followed by a description of our techniques for mapping existing lexical resources into the Conceptual Interlingua in Section 3. Section 4 gives an overview of the CINDOR system used in our experiments, incorporating the

SMART retrieval engine for weighting and retrieval, and the various cross-language retrieval techniques that we combined in our final experiments. We conclude in Section 5 with an overview of the results obtained in our TREC-9 submission based on the brief analysis we have conducted so far.

2. The CINDOR System.

The CINDOR system is a cross-language text retrieval system capable of accepting a user's query stated in their native language and then seamlessly searching, relevance ranking, retrieving and displaying documents written in a variety of foreign languages. A general overview of the CINDOR system and approach to cross-language retrieval can be found in [Ruiz *et al* 2000].

At the core of the CINDOR approach to cross-language retrieval is the idea of a 'Conceptual Interlingua'; a hierarchically organized knowledge base of essentially language-independent concepts. This concept hierarchy is then linked to multiple terminological resources for different languages which realize the lexicalization of concepts in each of the languages of the system. Cross-language retrieval is enabled by mapping the terms of documents and user queries from different languages into the interlingual concept representation, which provides the vocabulary for indexing and matching of document and query content.

Our Conceptual Interlingua has been built around the Princeton WordNet [Miller 1990], which contains approximately 165,000 different word forms organized into some 92,000 concepts denoted by a group of synonyms, or '*synsets*'. We consider the synset hierarchy as the core of the Conceptual Interlingua, with the 165,000 English terms to be the starting terminology for English. This has been extended with terminology in French, Spanish, and Japanese mapping to about 20,000 synsets in each case. More recently, we have integrated the German and Italian versions of EuroWordNet in order to provide a basis of terminology coverage in those languages. A primary focus of our research for TREC-9 has been the automated extension of the Conceptual Interlingua to terminology in Chinese.

3. Chinese Conceptual Interlingua

A stated goal of our research agenda in conjunction with TREC-9 participation was to link Chinese terminology into the Conceptual Interlingua in a fully automated process in as little time as possible. Our resulting efforts spanned a two-month period with essentially one person concentrating on this effort. Our goal was to identify existing lexical or terminology resources in Chinese and one of the existing Conceptual Interlingua languages (English) and use a range of approaches to link concepts to Chinese terms. While we were aware of a Chinese resource similar to WordNet (HowNet), it was not clear that there was a simple mapping from HowNet to WordNet (plus it did not appear that we would ultimately be granted permission to use this resource commercially), so use of this resource was not considered. We therefore concentrated our work on a bilingual English-Chinese lexicon available through the Linguistic Data Consortium (LDC) [LDC 2000] and a parallel English-Chinese corpus of Hong Kong Laws.

The basic approach to linking Chinese terminology to our Conceptual Interlingua is to find the most likely Chinese translations for each English term. The LDC bilingual lexicon contains translations for 110,834 English terms (including single terms and phrases). Generating pairs for all possible translations from English to Chinese from this lexicon generates 224,427 English-Chinese translation pairs.

While a simple approach would involve linking each English term in the Conceptual Interlingua to every one of the Chinese translations from the lexicon, we have investigated the use of a process of *'lexical triangulation'* in order to find evidence to support the choice of the most likely translation(s) for each term when multiple translations are possible. In the first instance, we take advantage of the WordNet synsets that are retained in the Conceptual Interlingua. More often than not, a synset contains more than one synonymous term, each of which may have multiple translations from the bilingual lexicon. By applying various intersections to the set of Chinese translations, we can limit the Chinese terms to those more likely to be translations of the sense of the English terms as used in that particular concept (synset). If we take all of the Chinese translations of all synonyms in a synset and rank them by frequency of occurrence, then several criteria for selection can be applied:

1. Strict intersection: Only the Chinese terms that are translations of all synonyms associated with the synset are selected.

2. Threshold method: Only the Chinese translations with frequency above a certain threshold are selected. We have set this threshold at 50% of the number of English terms associated with the synset.

3. Relaxed threshold: All the Chinese terms with frequency above a threshold, and all those terms that have frequency greater than 1. We selected the same threshold value of 50% of the number of English terms associated with the synset. Observe that for synsets that have 4 or less English terms associated, this option is the same as option 2. However for synsets that have more than 5 words associated, this option tends to generate a larger number of terms.

An example of this process of lexical triangulation is presented in Figure 1 below. The English concept (*kidnap*, verb) has four English terms associated with it. Given the Chinese translations of each of these terms from the bilingual lexicon, there is one term (诱拐) that is a possible translation for three of the four English synonyms. If we use criteria 1 above, the method will generate no entries in the Chinese terminology of the

Conceptual Interlingua for the verb *kidnap*. If we use criteria 2 or 3 above however, the Chinese term 诱拐 will be linked to the concept of the verb *kidnap*.

English Synset:
15 / verb /abduct/ kidnap/ nobble/ snatch/

Bilingual Dictionary Entries:

Abduct	/诱拐/绑架使外转/
Kidnap	/绑架/拐骗/诱拐/拐走/绑票/拐/
Nobble	/诈骗/
Snatch	/抓/搔/夺/匆忙地做/带走/好不容易救出/ 逼近来抓住/诱拐/碎片/一口/夺取/扒/

Figure 1: Translation of terms of the concept *kidnap*

When we applied this process over the entire LDC lexicon to link Chinese terms to the Conceptual Interlingua, the 'strict intersection' criteria yielded 13,337 Chinese terms linked to about 12,000 concepts. The 'threshold' method generated coverage comparable to our European languages (terms linked to about 25% of concepts), while the 'relaxed' method resulted in terms linked to about 63% of concepts, but is expected to contain more noise.

We were then further able to extend our process of lexical triangulation through the use of a parallel English-Chinese corpus as an additional source of translation evidence. Using a bilingual lexicon as a bootstrapping device, one can examine the sentence contexts of a parallel corpus to identify translations of English terms for which no translation is given in the bilingual lexicon. Given an English word W_E for which a translation is sought, a context can be identified (either the sentence in which the word occurs or a fixed window of surrounding words) in the English corpus. We then use the bilingual lexicon to find translations of as many of the words in the context of W_E as possible and align these with words in the context from the aligned Chinese corpus. Words remaining in the Chinese context, after stopwords have been removed, are candidate translations for W_E. For any given English word, candidate selection can be performed over multiple occurrences of the word in the corpus and candidates then ranked by frequency of occurrence. Further, this can be expanded to encompass Chinese terms as candidates for translation of all synonym terms within a concept, as outlined in the process above.

While investigating our corpus processing approach however, we discovered that many translation candidates identified from the corpus intersected with translation candidates from the bilingual lexicon, but which were not selected by any of our three selection criteria described above. We therefore adopted a modified translation selection approach as follows:

1. Generate candidate Chinese terms for a concept based on translation of all English synonyms from the bilingual dictionary (note: the first step was to translate the LDC lexicon from GB to Big-5 encoding to match the parallel corpus)
2. Generate frequency-ranked candidate Chinese translations for an English term from aligned contexts in the parallel corpus
3. Accept Chinese terms which meet the ‘threshold’ criteria above
4. Accept additional Chinese terms which are candidates in both the lexicon and corpus sets (corpus-attested translations).

This combined approach generated a Chinese Conceptual Interlingua with over 63,000 terms linked in to 38,000 concepts (40% coverage of WordNet) and this is what we used in our TREC-9 experiments.

4. CINDOR Chinese Retrieval

Given the Conceptual Interlingua approach encompassed in the CINDOR system, lexical-conceptual analysis of documents and queries is an integral part of the indexing process. Specifically, when dealing with Chinese, this necessitates tokenization/segmentation of input text as opposed to using character n-grams (bi-grams) as are often used for Chinese retrieval. We use the ‘mansegment’ segmentation module available through the Linguistic Data Consortium. An advantage of this module over other segmenters available is that it is capable (with different configuration) of segmenting Chinese text written with either traditional or simplified character sets. Chinese terms identified through segmentation are then matched against the Chinese Conceptual Interlingua terminology for mapping into concepts which are used in indexing.

It was the clear conclusion of our TREC-8 experimentation, supported by some follow-on investigation that we conducted, that retrieval performance of the CINDOR system was being negatively impacted by the use of a simplified *tf×idf* retrieval mechanism in the underlying search engine that was performing well below the standard of other retrieval engines participating in the TREC evaluation. In order to address this issue, we have loosely integrated CINDOR’s Conceptual Interlingua processing with the SMART retrieval system [Salton 1971]. We used the Cornell ftp version of SMART augmented with recent term weighting schemes (pivoted length normalization and BM25) and modified to handle UTF-8 encoded text. Our use of SMART therefore enabled various experiments with respect to CINDOR retrieval in the Chinese cross-language track.

Our first investigation concerned the use of multiple indexing vocabularies of Chinese text in CINDOR (known as *ctypes* in SMART terminology). Given the process of analyzing Chinese text in CINDOR, we had access to three possible indexing vocabularies:

- Terms: output from the Chinese segmenter
- Concepts: assigned from the Conceptual Interlingua for Chinese terms
- Bi-grams: derived directly from the Chinese documents (queries)

We therefore compiled three vector representations of each Chinese document, corresponding to each vocabulary. Similarity between a query and documents was then computed using a linear combination of the vector similarities of each vocabulary:

$$\mathrm{Sim}(\mathbf{d},\mathbf{q}) = \lambda * \mathrm{Sim}_{\mathrm{Terms}}(\mathbf{d},\mathbf{q}) + \theta * \mathrm{Sim}_{\mathrm{Concepts}}(\mathbf{d},\mathbf{q}) + \rho * \mathrm{Sim}_{\mathrm{Bigrams}}(\mathbf{d},\mathbf{q})$$

Where λ, θ, and ρ are coefficients that weight the contribution of each vocabulary, **d** is the document vector and **q** is the query vector. We used the well known pivoted length normalization (Lnu.ltu) weighting scheme [Singhal *et al* 1996]. This scheme weights the documents using logarithmic average term frequency and unique term pivoted length normalization, which corresponds to the formula:

$$\frac{\dfrac{1+\log(tf)}{1+\log(\text{average } tf)}}{(1.0-slope)\times pivot + slope \times \#\text{ of unique terms}}$$

where *tf* is the term frequency, *slope* and *pivot* are parameters of the pivoted length normalization scheme. For our runs we use a *slope*=0.25 and the *pivot* is set to the average document length of the collection.

We experimented with pseudo-relevance feedback using Rocchio's formula to rank the terms in an initial retrieved set to expand the query for a feedback loop:

$$w_{new} = \alpha w_{orig} + \beta \frac{\sum_{d_i \in R} w_{d_i}}{|R|} + \gamma \frac{\sum_{d_i \notin R} w_{d_i}}{n-|R|}$$

where w_{orig} is the weight of the term in the original query; w_{d_i} is the weight of the term in document d_i; R is the set of relevant documents; $|R|$ is the number of relevant documents; n is the number of documents considered (in retrieval feedback this is usually set to the number of documents presented to the user); and α, β, and γ are constant coefficients that control the contribution of each factor. Terms with negative weights are discarded. The terms are ranked by the computed weight w_{new} and the top m terms are used to expand the query.

Our pseudo-relevance feedback method uses the original query to obtain the top 1000 retrieved documents. We assume that the top N documents are relevant and that the bottom 100 documents are not relevant. The query is then expanded with the top m

ranked terms according to Rocchio's formula. Since we are using three index vocabularies, the pseudo-relevance feedback process adds *m* expansion terms to each vocabulary (vector).

Given this retrieval scenario, discovery of optimal settings for this retrieval model involves tuning the following parameters:

- λ, θ and ρ for the combination of each vocabulary vector in the final similarity score.
- α, β, γ, *N* and *m* for pseudo-relevance feedback using Rocchio's formula.

We first found the best parameter combination for pseudo-relevance feedback on the TREC-5 and TREC-6 Chinese track test collections, consisting of documents from the People's Daily newspaper and the Xinhua news agency, trying all combinations of the 5 parameters using only a single vocabulary (terms) for the monolingual Chinese queries with the following sets of possible values:

- *N* = 5, 10 ,15 and 20
- *M* = 5, 10 20, 50, 100, 150 and 200
- $\alpha = 8$
- $\beta = 32$ and 64
- $\gamma = 8$, 16 and 32

Our initial retrieval baseline for feedback was simple retrieval using the *Lnu.ltu* weighting scheme. Observe that the rationale for selecting the values for α, β and γ is that given a fixed value for the contribution of the original query terms (α=8), we explore relative weightings of the contribution of relevant and non-relevant documents (e.g. $2\times\alpha$).

We tried the 168 possible combinations with the above parameter values for each of the two sets of Chinese topics. Figure 2 below shows the variation of performance (average precision) for the set of TREC-6 queries and *N*=10 (48 runs). The highest performance is obtained for *N*=10, *m*=20, α=8, β=64 and γ=32 with an average precision of 0.5263. Therefore, the top 20 terms are selected for query expansion based on the assumption that the top 10 documents are relevant (using α=8, β=64 and γ=32 in the Rocchio formula).

Similarly, we found through testing on the TREC-5 and TREC-6 Chinese test collections the optimal set of parameters λ (terms), θ (concepts) and ρ (bi-grams) for weighting the relative contribution of each indexing vocabulary to the final document-query similarity value. We found that monolingual retrieval and cross-language retrieval have different optimal parameter settings. For monolingual retrieval the best performance was found to be λ=20, θ=1 and ρ=20 while for cross-language retrieval the best parameter settings are λ=4, θ=1 and ρ=4. These parameters indicate that, while not contributing as much as the Term or Bi-gram indexing vocabularies for retrieval, our Conceptual Interlingua concepts are relatively more important in a cross-language retrieval setting than in a monolingual search environment.

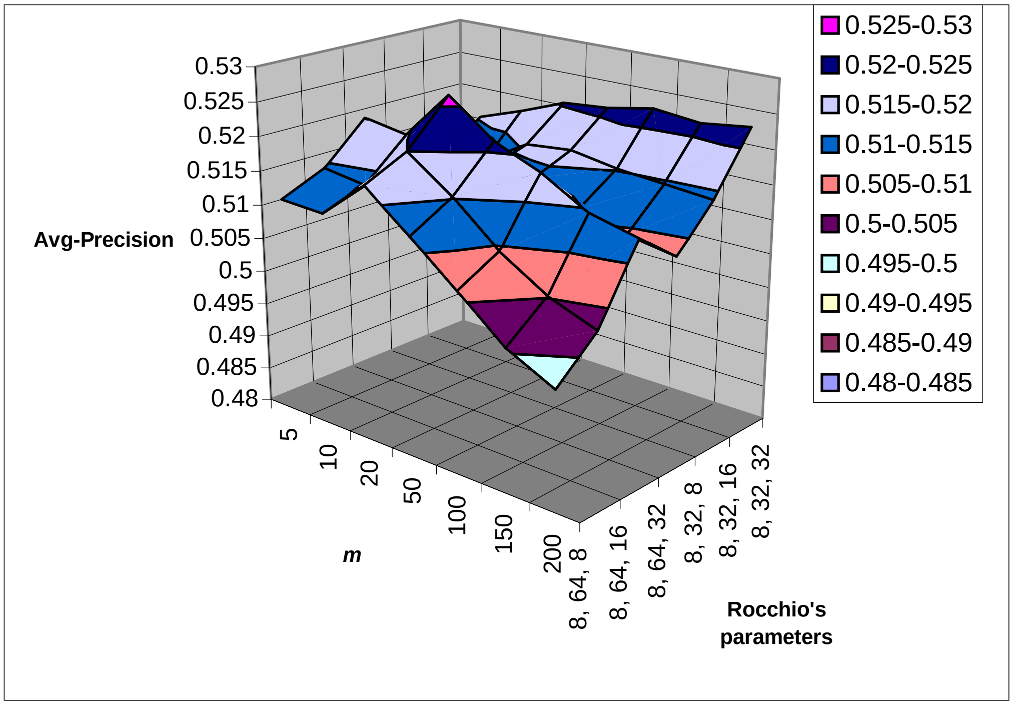

Figure 2. Variation of performance (Average Precision) for *N*=10 and 48 combinations of Rocchio parameters *m* and *α*, *β*, *γ*

We have tested and verified our overall retrieval approach and Conceptual Interlingua on the TREC-5 and TREC-6 Chinese test collections through a series of retrieval experiments. Tables 1 and 2 below illustrate the incremental improvements in performance to be gained from the combination of approaches we have used. Simple term indexing results in an average precision of 0.3572 in a monolingual test over the TREC-5 collection. Term-based English-Chinese cross-language retrieval, using machine translation of terms (Alis Technology's Gist-in-Time system), gives average precision of 0.2408 in a similar test. Augmenting this run by concept indexing through the Conceptual Interlingua yields an improvement of 3.6% and 4.6% respectively. If this run is in turn augmented by pseudo-relevance feedback, there is a further 3.1% monolingual and 8.9% cross-language improvement in average precision. If bi-grams are then extracted from the documents and query (translation) and used in matching following the above linear combination, a final 9% and 9.3% improvement is observed.

TREC-5 Test Collection

	Monolingual	% Gain	Cross-Language	% Gain
Term Indexing (MT)	.3572		.2408	
+ Conceptual Interlingua	.3701	3.6%	.2518	4.6%
+ Relevance Feedback	.3817	3.1%	.2742	8.9%
+ Bi-gram Indexing	.4161	9%	.2998	9.3%
		16%		**24.5%**

Table 1: Incremental improvements in Average Precision from combination of retrieval techniques

These improvements in performance are replicated on the TREC-6 test collection in the same way in Table 2. The aggregate result is performance 16% and 10% over the term-based baseline for monolingual retrieval in TREC-5 and TREC-6, with 24.5% and 25.3% improvements in cross-language retrieval in the same way. This, we felt, provided a firm foundation from which to launch our TREC-9 submissions.

TREC-6 Test Collection

	Monolingual	% Gain	Cross-Language	% Gain
Term Indexing (MT)	.5010		.3091	
+ Conceptual Interlingua	.5151	2.8%	.3170	2.6%
+ Relevance Feedback	.5208	1.1%	.3472	9.5%
+ Bi-gram Indexing	.5509	5.8%	.3875	11.6%
		10%		**25.3%**

Table 2: Incremental improvements in Average Precision from combination of retrieval techniques

5. TREC-9 Results and Analysis

We submitted a monolingual run (TWmono3CItdn) and a cross-language run (TWe2c3CItdn) for evaluation by NIST. Both runs correspond to the best parameter settings for the training collection as explained in the previous section. Our final results for TREC 9 monolingual performance show an average precision of 0.3041. This monolingual run is above the median in 18 of the 25 topics. The average difference above the median is 0.052 (20.5% above the median).

Our cross-language run achieved 0.1312 average precision, which is below the median (0.1460). The cross-language results are above the median in 13 of the 25 topics and a difference with the median of -0.015 (10.9% below the median). This cross-language run achieved 42% of our monolingual run performance, which is considerably lower than the 70% we obtained in the training set. While we have not yet conducted an in-depth analysis of what caused this low cross-language performance relative to our monolingual baseline, we suspect it to be primarily related to gaps in the translation resources used. Even from a superficial analysis of the results, it is clear that there were gaps in translation, both in the machine translation and the Conceptual Interlingua. Examples are '*Daya Wan electric plant*', '*computer hackers*', '*Tiananmen Square*', etc. We have also noticed that the machine translation system consistently translated '*China*' as '瓷器' (in the sense of "Mom's best…") instead of '中国' (the nation). This was compensated for however by the fact that the correct translation of '*China*' had been captured in the Conceptual Interlingua. Since most of the queries were about China however, this is likely to have impacted the final performance of some queries.

6. Conclusion

Our TREC-9 experiments reported here are part of an ongoing set of experiments that evaluate the performance of the CINDOR system over a wide range of languages. Our work on automatic generation of Conceptual Interlingua resources here has, as desired, generated a general approach and a corresponding set of tools that can now be applied toward the rapid addition of other languages. Our results indicate that the automatically generated Conceptual Interlingua can contribute to improved retrieval performance for cross-language information retrieval over a simple term-based baseline.

The research version of CINDOR used here has certainly benefited from integration with the retrieval capabilities of SMART. This has had the further advantage of allowing us to experiment with retrieval models using a combination of indexing vocabularies and a combination of different sources of evidence for cross-language retrieval.

Despite the apparent success of our TREC-9 participation, especially in our monolingual Chinese runs, we believe that there remain avenues along which we can further enhance the performance of the CINDOR system. We continue to pursue research directed at improving our cross-language retrieval precision in all languages by processing and matching of named entities and multi-word terms across languages and in the area of word sense disambiguation in the framework of WordNet for our Conceptual Interlingua. We hope to realize the fruits of these efforts in future evaluations.

7. Bibliography

[LDC 2000]

Linguistic Data Consortium, "List of Chinese Resources over the Internet". http://morph.ldc.upenn.edu/Projects/Chinese

[Miller 1990]

Miller G., "WordNet: An On-line Lexical Database", *International Journal of Lexicography*, Vol. 2, No. 4, Special Issue, 1990.

[Ruiz *et al* 2000]

Ruiz, M. E., Diekema, A., and Sheridan, P. "CINDOR Conceptual Interlingua Document Retrieval:TREC-8 Evaluation". *Proceedings of the eighth Text Retrieval Conference (TREC-8)*, NIST special publication, 2000.

[Salton 1971]

Salton, G. *The SMART retrieval system: Experiments in automatic document processing*. Englewood Cliffs, NJ. Prentice-Hall. 1971.

[Singhal *et al* 1996]

Singhal A, Buckley C, and Mitra M, "Pivoted Document Length Normalization", In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, August 1996, ACM Press, pages 21-29.