Description of NTU QA and CLIR Systems in TREC-9

Chuan-Jie Lin, Wen-Cheng Lin and Hsin-Hsi Chen

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, TAIWAN, R.O.C.
E-mail: {cjlin, denislin}@nlg2.csie.ntu.edu.tw, hh_chen@csie.ntu.edu.tw
Fax: +886-2-23628167

1. Introduction

National Taiwan University (NTU) Natural Language Processing Laboratory (NLPL) took part in QA and CL tracks in TREC9.

For the QA Track, we proposed three models, which integrate the information of Named Entity, inflections, synonyms, and co-reference. We plan to evaluate how each factor effects the performance of a QA system.

For the Cross-Language Track, we employed two approaches in Chinese-English information retrieval (Bian and Chen, 1998; Chen, Bian, and Lin, 1999) to English-Chinese information retrieval. We plan to explore their usability.

2. QA Track

In TREC-8, we've experimented on expanding questions by adding inflections of verbs and nouns, as well as the their synonyms (Lin and Chen, 1999). However, the performance was not so good as our expectation. This year we propose three models to see whether expansion is helpful or not. Model 1 is a base model. Only inflections are added. Model 2 adds synonyms from WordNet (Miller, 1990). And Model 3 tries to resolve co-reference in a simple way. Each of them will be described in detail in later sections.

Besides, we select answers according to the named entities that the question might be relevant. Our QA system will guess the interested entity type by looking at the questions. Position of the interested answer terms is also important. If the length of answering sentences is longer than restricted length, the final answer text has to include the actual answer. We also propose a method to implement this idea. The proposed algorithm will be described later.

2.1. Model Description

2.1.1. Interested Entity Type

After taking a question as input, our system first guesses which entity type the question is interested in. The method is simply rule-based. If the question starts with "who", "when", and "where", it may ask for a person name, a time/date expression, and a location name, respectively. If it starts with "what" or "which", or it is a "Name a ..." -type question, then the system goes on to look at the first noun behind it. We collected some keywords to indicate the interested entity types, such as "country" for location name, "person" for personal name, and so on.

2.1.2. Named Entity Extraction

Named entity extraction plays an important role in our experiments. It is introduced while deciding question focusing, doing question expansion, and measuring similarity between document passage and question sentence.

For named entity extraction, we employ several named entities dictionaries, such as gazetteer, a collection of family name, *etc*. Different from simply dictionary look-up, these dictionaries also include other useful information. For a personal name, we can know that it is a family name, a male first name, or a female first name. For a country name, we can get its adjective form as well as how to call its people. For other location names, it provides the names of provinces or countries it belongs to as well. Organization names are accompanied by their abbreviations. We have not employed the information of types of personal names and the superior administrative division yet.

Time/date expression is simply keywords (Sunday, January, etc.) The resolution of expressions like "yesterday", "last week" are still undergoing. Other named entities like quantity and numbers are not handled yet.

2.1.3. Base Model - Question Expansion by Named Entity and Inflection Forms

In Base Model, we first decide if there is a named entity in the question sentence. If so, we record its equivalence (e.g. abbreviation of an organization name). Notice that a named entity can be more than one word. For the rest words in the question sentence, we remove stop words and attach the root form and all the inflection forms of each of them. These newly invited terms are for the use of similarity comparison later.

The next step is to segment documents into passages as comparison units. The document set we use this year is the set of the 50 most relevant documents to the questions. The relevant document set is offered by NIST. In the Base Model, a passage is simply a sentence.

For each passage, we also identify named entities in it, but their equivalences are not attached. The inflections are not added either. This is because we have already introduced them in the question side.

Then we measure its similarity to the expanded question sentence. For each word (or phrase) occurs in the passage and also in the expanded question, it contributes a score to the similarity. By the recent experiment, if it is a named entity, it contributes 2 points; otherwise 1 point. If it occurs in the original question, the contributed score is doubled.

Besides, if a word (or a phrase) does not occur in the question but is of the interested type of the question, the FOCUS tag is set and the position of this word is recorded.

While giving answers, those words (or phrases) that are assigned the FOCUS tag are reported first. The passage of higher score is considered to be more possible to carry the answer and is ranked higher.

To meet the length restriction, we have to truncate the passages longer than 250 bytes. We decide the focusing center of each answering passage first. Truncate characters 125 bytes ahead of the center and also the exceed part if the remaining passage is still longer than 250 bytes. For those assigned a FOCUS tag, the center is the average position of all the found named entities of interest. For those did not, the center is the average position of words that also occur in the question sentence.

2.1.4. Model 2 – More Expansion by Synonyms

Besides the basic structure of Base Model, we also expand questions by the synonyms of ordinary nouns or verbs, i.e., those which are not named entities. Synonyms are obtained by looking up the WordNet (Miller, 1990). We do so because we want to save those answers written in different terms.

2.1.5. Model 3 – Passage with Co-Reference Resolved

This model is also based on the Base Model. But we want to resolve co-reference problem first before measuring similarity with the question sentence. We proposed a simple strategy to do so: take the first sentence as a passage. If the next sentence contains pronouns (except “it”), it is merged into the previous passage. Or if the next one contains a phrase of the pattern “the A” and the word “A” occurs in the previous passage, it is merged into the previous one, too. It can help resolve anaphora problem as well as the co-referential noun phrases.

2.2. Evaluation

Table 1 lists the results of our three models. We submitted three runs, each run for each model, i.e., qantu01 for Base Model, and so on. Each answer text can be judged as Wrong, Correct, and Unsupported. "Unsupported" means that the document associated to the answer text does not really support the answer. The Strict Evaluation only counts Correct ones, and the Lenient Evaluation takes both Correct and Unsupported ones as correctly answered.

Table 1. Results of Three Models in the QA Track at TREC-9

Run ID	Strict		Lenient		Strict (Debugged)	
	MMR	Failed	MMR	Failed	MMR	Failed
qantu01	0.315	377 (55.3%)	0.348	354 (51.9%)	0.367	333 (48.8%)
qantu02	0.315	376 (55.1%)	0.341	354 (51.9%)	0.372	324 (47.4%)
qantu03	0.278	394 (57.8%)	0.309	370 (54.3%)	0.319	354 (51.8%)

By Table 1, less than half of the questions failed to be answered. It is better than last year that we only answered 1/3 of the questions correctly. There are 24 more questions in average answered by unsupported documents.

Comparing the performance of different models, Base Model and Model 2 are almost the same, but Model 3 is worse than the other two. Model 2 answered nine more questions than Base Model did, but Base Model offered unsupported answers at higher ranks than Model 2 did in the Lenient Evaluation. Model 3 is worse in either evaluation.

It seems that adding synonyms does not help a lot. It even lows down the speed. The most difficulties we met in QA are often paraphrases, not only synonyms. Therefore, it might be more efficient to tackle the paraphrases problem.

The reason that Model 3 worked badly may be the over-simplified co-occurrence resolution. For those questions failed to be answered here but successful in the other two runs, it was often the case that the passages containing the answer texts have been expanded into large ones. The occurrence of co-reference candidates is too frequent to simply concatenate sentences.

But co-reference resolution is helpful for question answering. During the investigation, we found that a portion of questions can be answered by keyword matching with co-reference resolved. To integrate the co-reference resolution part into the system, or find an alternative way to tackle it will be another important future work.

3. Cross-Language Track

Query translation is usually employed to unify the languages in queries and documents in Cross Language Information Retrieval (CLIR). Bian and Chen (1998) proposed a hybrid approach that integrated both lexical and corpus knowledge to translate queries. A bilingual dictionary provides the translation equivalents of each query term, and the word co-occurrence information trained from a target language text collection is used to disambiguate the translation. Target polysemy is another problem in CLIR. Bian, Chen and Lin (1999) augmented a pseudo context to a query term to restrict its use in the target language. The contextual information is trained from a source language text collection.

The above approaches performed well in Chinese-English information retrieval (CEIR). In TREC-9, we made experiments in English-Chinese information retrieval (ECIR) to see that if these two approaches work.

3.1. Query Translation

In Cross-Language track, we adopted query translation to unify the language of queries and documents. Then we retrieved Chinese documents using a monolingual information retrieval system. We proposed two models to translate queries. The first one is CO model, which uses co-occurrence information trained from a text collection in source language to select the best translation equivalents of source language query terms. The second one is A1W model, which resolves the target polysemy problems by augmenting some restriction words. A TREC topic is composed of several fields. In our experiments, only the title and description fields were used to generate queries.

In CO model, the English queries were translated into Chinese as follows: First, all sentences in queries were parsed, so that phrases could be identified. We used Apple Pie to parse sentences. Second, we collected the translation equivalents of each phrase or word by looking up an English-Chinese bilingual dictionary. The stop words were ignored. If translation equivalents of a word could not be found, we tried again after stemming. The stemmer what we used is Porter's stemmer. Third, by using the co-occurrence information, we selected the best translation equivalents. We adopted mutual information (MI) (Church, *et al*., 1989) to measure its strength. The mutual information was trained from a text collection in target language, i.e. Academia Sinica Balance Corpus (ASBC) (Huang, *et al*., 1995). For each Chinese word, we collected its mutual information with other words within a window of size 3. For a query term, we compare the MI values of all the translation equivalent pairs (x, y), where x is the translation equivalent of this term, and y is the translation equivalent of another query term within a sentence. The word pair (x_i, y_j) with the highest MI value is extracted, and the translation equivalent x_i is regarded as the best translation equivalent of this query term. Selection is carried out based on the order of the query terms.

Take phrase "China's silk industry" as an example. The translation equivalents of each word are shown in Table 2. Table 3 lists the mutual information scores of some word pairs of translation equivalents. Consider the term "China" first. The translation equivalent pair with the highest MI score is <中國 , 絲綢>. Therefore, '中國' is selected as the translation of the word "China". Then we select the best translation equivalent of the words "Silk" and "Industry". In a similar way, '絲綢' and '工業' are selected, respectively.

Table 2. Translation Equivalents of Each Term in "China's silk industry"

Term	Translation Equivalents
China	中, 中國, 中華, 神州, 神洲, 華, 華夏
Silk	布帛菽粟, 帛絲, 絲綢, 綾子, 綾羅, 綢, 綢子, 綢緞, 緞子, 蠶絲, 紬
Industry	工商界, 工業, 行業, 產業, 業界

Table 3. The Mutual Information Scores for Some Word Pairs

	China			Silk		Industry			
	中	中國	中華	絲綢	蠶絲	工商界	工業	行業	產業
中							-0.276507		0.106157
中國				**5.069064**			1.035561	0.844666	-0.373534
中華									
絲綢		**5.069064**							
蠶絲									
工商界									
工業	-0.276507	**1.035561**							
行業		0.844666							
產業	0.106157	-0.373534							

In order to resolve target polysemy problem, we augmented some words to restrict the use of a translated query term in A1W model. In this model, the English queries were translated by CO model, and the translation equivalents of augmented words were added to target language queries. The augmented restriction words of a source language query term are those words that frequently co-occur with it within a window. We selected the terms that have only one translation to augment the original query term. The co-occurrence information was trained from TREC6 text collection (Harman, 1997), and the mutual information was used to measure the strength.

Assume that the terms of an English query E are E_1, E_2, E_3, …, and E_n, and their translations are C_1, C_2, C_3, …, and C_n. For each term E_i, we augment a sequence of words EW_{i1}, EW_{i2}, …, and EW_{imi} to it. EW_{ij} has only one translation and the MI score of the pair (E_i, EW_{ij}) exceeds a threshold. Assume that the corresponding Chinese translations of EW_{i1}, EW_{i2}, …, and EW_{imi} are CW_{i1}, CW_{i2}, …, and CW_{imi}, respectively. The list (C, CW_{i1}, CW_{i2}, ..., CW_{imi}) is an *augmented translation result* for E.

We assigned different weights to the translations of original terms and augmented restriction terms. They were determined by the following formula:

$$\text{weight}(C_i) = \frac{1}{n+1} \quad (1)$$

$$\text{weight}(CW_{ij}) = \frac{1}{(n+1) * \sum_{k=1}^{n} m_k} \quad (2)$$

Where n is number of words in a query Q; C_i is the translation of query term E_i; CW_{ij} is the translation of augmented restriction term EW_{ij} and m_k is the number of words in a restriction for C_k. In this formula, we assume that the sum of the weights of all the terms is 1. The words C_1, ..., C_n take n/(n+1) total weight. In this way, the weight of each C_i is 1/(n+1). The remaining 1/(n+1) is distributed to those CW_{ij}'s equally.

3.2. IR System

Our Information Retrieval system is based on vector space model. The index terms are Chinese character bigrams, and the term weighting function is tf*idf. When a query is submitted to this IR system, it computes the similarities of this query and all documents, then returns top rank documents. We adopt cosine vector similarity formula to measure the similarity of a query and a document. Higher score means that the query and the document are more similar.

3.3. Results

We submitted two runs: ECIRNTUCO and ECIRNTUA1W. The former used CO model to translate queries and the latter used A1W model. Table 4 shows the results of these two runs. Figure 1 shows the average precision of each topic.

Table 4. Results of Run ECIRNTUCO and ECIRNTUA1W

Run	Average precision	R-Precision	Rel_ret
ECIRNTUCO	0.0244	0.0218	206
ECIRNTUA1W	0.0197	0.0173	91

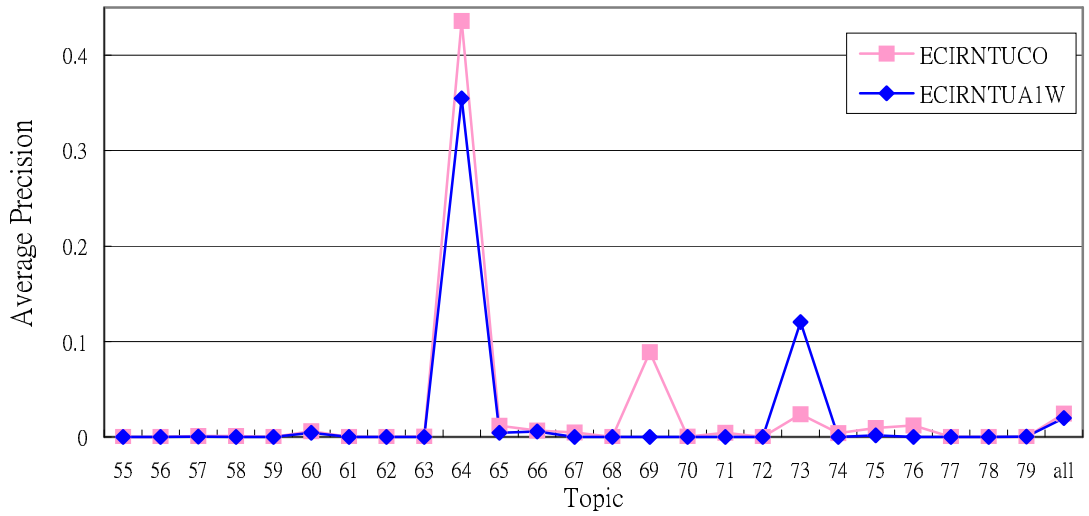

Figure 1. The Average Precision of Each Topic in Run ECIRNTUCO and ECIRNTUA1W

We found that the IR system had some bugs, so the results were not correct. After the IR system was corrected, we did three new runs: CO, A1W and MONO. Runs CO and A1W used the same translated topics as ECIRNTUCO and ECIRNTUA1W. For comparison, we did a monolingual information retrieval run, i.e. MONO. The results of these three runs are shown in Table 5. Figure 2 shows the average precision of each topic in new runs and the median average precision of all TREC9 English-Chinese information retrieval runs.

Table 5. Results of Run CO, A1W and MONO

Run	Average precision	R-Precision	Rel_ret
CO	0.0784 (32.78%)	0.0950 (34.75%)	328
A1W	0.0787 (32.9%)	0.0950 (34.75%)	329
MONO	0.2392	0.2734	563

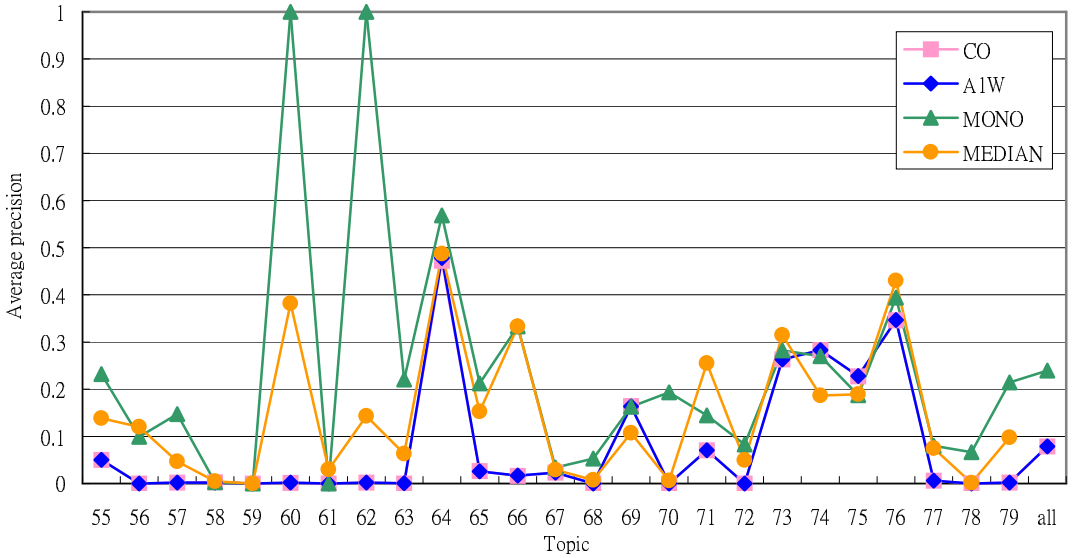

Figure 2. The Average Precision of Each Topic of Run CO, A1W, MONO and MEDIAN

Since our IR system is a simple system that is based on vector space model, its performance is not very good. The average precision of CO model is 0.0784 that is 32.78% of monolingual information retrieval. In our previous work, the performance of CO model in Chinese-English information retrieval is 56.96% of monolingual information retrieval. Although we cannot compare the performance directly since the document collection and IR system in these two experiments are not the same, it seems that the performance of CO model in ECIR is worse than that in CEIR. There are some factors that influence the performance. First, the degree of ambiguity of English is high. In this experiment, 264 query terms after removing stopwords have translation equivalents in our English-Chinese bilingual dictionary. Among these, only 42 terms have unique translation. On average, an English query term has 8.3 translation equivalents. Second, a stemmer was used instead of a morphological analyzer. Translation equivalents of many stemmed words cannot be found. Third, some phrases cannot be identified. We tried to identify phrases by parsing queries, but not all phrases can be identified. For example, "human rights violations" was identified as a minimal NP, and the NP "human rights" is missed. Therefore we failed to retrieve the translation of "human rights", that is "人權". The words "human" and "rights" were translated into "人類" and "權 益" respectively.

The performance of A1W model is almost the same as CO model. The average precision of A1W model is 0.0787 that is 32.9% of monolingual information retrieval. When we augmented restriction terms to an original query term, we also added noises. On the other hand, some good terms were not added. Recall that we only augmented co-occurrence terms that have only one translation. Many good terms are ambiguous so that they cannot be added. For example, since 'programmer' has four translation equivalents, it could not be selected as a restriction for 'computer'.

4. Conclusion

For the QA Track, we proposed three models this year. These models can help us to see the usefulness of each proposed factor. Base Model uses the information of named entity and its equivalence, as well as the information of inflection forms of general nouns and verbs. Synonyms of nouns and verbs are proved to be of little use. Simple co-reference resolution causes a drawback because of the wrongly merged passages.

At the Cross-Language Track, we proposed two models to translate queries: CO and A1W model. Two runs were submitted to cross-language track, and the performances are not so good as our expectation. The major reason is that our IR system has bugs. After correcting the bugs, we redid the experiment. The average precisions of CO and A1W model are 0.0784 and 0.0787 respectively. The improvement of A1W model was limited. How to select augmented restriction terms is a problem. We will make more experiments to see what strategy is more appropriate.

Reference

Bian, G.W. and Chen, H.H. (1998) "Integrating Query Translation and Document Translation in a Cross-Language Information Retrieval System." *Machine Translation and Information Soup*, Lecture Notes in Computer Science, No. 1529, Spring-Verlag, 250-265.

Chen, H.H., Bian, G.W. and Lin, W.C. (1999) "Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval." *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics,* 215-222.

Church, K. *et al.* (1989) "Parsing, Word Associations and Typical Predicate-Argument Relations." *Proceedings of International Workshop on Parsing Technologies*, 389-398.

Harman, D.K. (1997) *TREC-6 Proceedings*, Gaithersburg, Maryland.

Huang, C.R., *et al.* (1995) "Introduction to Academia Sinica Balanced Corpus." *Proceedings of ROCLING VIII*, Taiwan, 81-99.

Lin, C.J. and Chen, H.H. (1999) "Description of Preliminary Results to TREC-8 QA Task." *TREC-8 Proceedings*, Gaithersburg, Maryland, http://trec.nist.gov/pubs/trec8/papers/NTU_TREC8_QA.pdf

Miller, G. (1990) "Five Papers on WordNet." *Special Issue of International Journal of Lexicography* 3.

MUC (1998) *Proceedings of 7th Message Understanding Conference*, http://www.muc.saic.com/proceedings/proceedings_index.html.

Salton, G. and Buckley, C. (1988) "Term Weighting Approaches in Automatic Text Retrieval." *Information Processing and Management*, Vol. 5, No. 24, 513-523.