Thomson Legal and Regulatory at NTCIR-5: Japanese and Korean Experiments

Isabelle Moulinier and Ken Williams
Thomson Legal and Regulatory
Research and Development Group
610 Opperman Drive, Eagan, MN 55123, USA
Isabelle.Moulinier@thomson.com

Abstract

Thomson Legal and Regulatory participated in the CLIR task of the NTCIR-5 workshop. We submitted formal runs for monolingual retrieval in Japanese and Korean, as well as for bilingual English-to-Japanese retrieval. We employed enhanced tokenization for our Japanese and Korean runs and applied a novel selective pseudo-relevance feedback scheme for Japanese. Our bilingual search participation was a straightforward application of an off-the-shelf Machine Translation system to transform an English query into a Japanese query.

Unfortunately we cannot draw many conclusions from our participation, as our experiments were hampered by technical difficulties, particularly with our tokenization and stemming components.

Keywords: *pseudo-relevance feedback, online resources.*

1 Introduction

Thomson Legal and Regulatory participated in the Cross-Lingual Information Retrieval task of the NTCIR-4 workshop. For this year's campaign, we participated in three subtasks: monolingual Japanese retrieval, monolingual Korean retrieval, and bilingual retrieval from English to Japanese. Characteristics of the tasks and collections are described in [6].

In the past, we have found that linguistic processing of document texts as well as queries can have a significant impact on system effectiveness. Since we lack sufficient personal familiarity with Korean and Japanese to implement or evaluate tokenization and stemming ourselves, we elected to use an off-the-shelf commercial linguistic analyzer for these two tasks. As detailed below, however, this third party software was not accurate enough, which negatively impacted our results.

For our monolingual Japanese retrieval system, we re-evaluated a novel selective pseudo-relevance feedback scheme that we developed for the 2005 Cross Language Evaluation Forum (CLEF) on European languages [7]. As with our CLEF results, we found that our scheme helped performance on several queries, but to our surprise also hindered performance on other queries, rendering its overall benefit statistically insignificant.

In Section 2, we briefly present our base retrieval system as well as our pseudo-relevance feedback approach. Section 3 summarizes our experiments with Japanese and Korean, including our results with pseudo-relevance feedback. Section 4 reports some preliminary bilingual experiments.

2 System overview

2.1 The Win system

The WIN system is a full-text natural language search engine, and corresponds to TLR/West Group's implementation of the inference network retrieval model. While based on the same retrieval model as the INQUERY system [2], WIN has evolved separately and focused on the retrieval of legal material in large collections in a commercial environment that supports both Boolean and natural language searches [8].

2.1.1 Indexing

We used words as our basic indexing units. Words are identified using a third party tokenizer. For our experiments, we used the tokenizer included in the LinguistX toolkit commercialized by Inxight [5]. Where appropriate, words are also stemmed using the same toolkit. In addition, the stemmer identifies compound terms as well as their components. We rely on this feature to index Korean compound terms.

2.1.2 Document scoring

WIN supports various strategies for computing term beliefs and scoring documents. We used a standard *tf-idf* scheme for computing term beliefs in all our runs. The belief of a single concept is given by $tf_{norm} * idf_{norm}$, where

$$tf_{norm} = \frac{\log(tf + 0.5)}{\log(tf_{max} + 1.0)}$$
$$idf_{norm} = \frac{log(C + 0.5) - log(df)}{log(C + 1.0)}$$

and tf is the number of occurrences of the term within the document, tf_{max} is the maximum number of occurrences of any term within the document (a weak approximation for document length), df is the number of documents containing the term and C the total number of documents in the collection.

The document is scored by combining term beliefs using a different rule for each query operator [2]. The final document score is an average of the document score as a whole and the score of the best portion. The best portion is dynamically computed based on query term occurrences.

2.1.3 Query formulation

Query formulation transforms natural language text into a belief network with simple concepts and operators. The belief network is then used for scoring documents. Simple concepts typically correspond to terms in the query, while operators add structure and may represent phrases and compounds.

Currently, our Japanese and Korean processing identifies stopwords and considers all other tokens in the query as concept. In addition, Korean processing uses the compound operator (implemented here as a proximity search) after the stemmer has identified compounds and their components.

2.2 Pseudo-relevance feedback

We have incorporated a pseudo-relevance feedback module into our search system. We follow the approach outlined by Haines and Croft [3]. In addition, we adapted our CLEF experiments on selective pseudo-relevance feedback for Japanese search.

2.2.1 Basic pseudo-relevance feedback approach

We select terms for query expansion using a Rocchio-like formula and add the selected terms to the query. The added terms are weighted either using a fixed weight or a frequency-based weight.

2.2.2 Selective pseudo-relevance feedback

Pseudo-relevance feedback (PRF) is known to be useful on average but can be detrimental to the performance of individual queries. At CLEF, we took a first step towards predicting whether or not PRF would aid individual queries [7]. We follow the same approach for our NTCIR runs.

PRF parameters were selected based on training data from previous participation. Using these values, we constructed a simple prediction rule that identifies those queries where PRF was very detrimental. Our decision rule is composed of two components: $bestcore$, the score of the top ranked document and $maxscore$, the maximum score any document can achieve for a given query, computed by setting the tf_{norm} factor in belief scores to 1. Our prediction rule is of the form:

if $maxscore \geq Min\_MS\_Value$
and ($maxscore < MS\_Threshold$ **or** $bestscore \geq Min\_TD\_Value$)
then
Apply PRF

Using training data, we searched for the best parameters in this three-dimensional space (*Min_MS_Value*, *MS_Threshold*, and *Min_TD_Value*).

Our intuitive reasoning for such a rule is that a query with a high maximum score includes infrequent terms and is precise. We can then apply PRF to find additional documents. On the other hand, a query with a low maximum score includes frequent terms and we can not guarantee that the returned documents are relevant. Similarly, when the score of the top ranked document is low, we do not assume that the document is relevant.

3 Monolingual experiments

Our experiments for NTCIR-5 were hindered by technical difficulties with our new tokenization and stemming components. We used a third party linguistic toolkit, commercialized by Inxight [5], to identify Japanese word boundaries and Korean compound parts. We used the latest release (at the time) as dictionary coverage for both languages had improved according to our provider. Unfortunatly, we lated discovered that some of our Japanese tokens were incorrect. In particular, the offsets returned by the tokenizer sometimes resulted in words with no characters, and some offsets extended much beyond the end of the tokenized text. This problem prevented us from assessing whether the larger lexicon coverage actually translated to better retrieval performance.

Our Japanese runs focused on pseudo-relevance feedback, and in particular on evaluating the selective approach we derived for CLEF. Table 1 show

that pseudo-relevance feedback significantly improved performance but that applying PRF selectively did not result in additional improvement. The runs correspond to the following settings:

- tlrrd-J-J-D-01: selective PRF, selecting 5 terms from the top 5 documents. Queries use the Description field only.
- tlrrd-J-J-D-02: PRF, selecting 10 terms from the top 10 documents. Queries use the Description field only.
- tlrrd-J-J-D-05: no PRF. Queries use the Description field only.
- tlrrd-J-J-T-03: PRF, selecting 10 terms from the top 5 documents. Queries use the Title field only.
- tlrrd-J-J-T-04: no PRF. Queries use the Title field only.

The difference between runs tlrrd-J-J-D-01 and tlrrd-J-J-D-02 does not reflect the influence of selectively applying pseudo relevance feedback as parameters are set differently. We report comparable runs in Table 2. We note that selective PRF neither improve nor degrade performance in a significant manner. A per-query analysis shows that the selection rule blocks the application of pseudo-relevance feedback for 7 queries. This prediction is correct 4 times out of 7. However, the rule failed to prevent detrimental application of PRF for over 10 queries. This is not surprising as we crafted the rule for precision rather than recall [7].

Our official Korean runs are reported in Table 3. These runs do not use relevance feedback. Compounds are handled as natural phrases, that is to say a proximity operator of 3. During our analysis, we found that compound identification was a problem. In particular, our lemmatization tool generated a number of alternatives for compounds. A larger number of alternatives has an impact on document scoring. We also realized that the lack of edited stopword list is affecting our retrieval performance.

In future work, we would like to compare different versions of our linguistic toolkit, in order to evaluate whether enhanced lexicon coverage translated into better retrieval performance. In addition, we would like to evaluate alternative tokenization and stemming methods.

4 Bilingual experiments using online resources

Our involvement with bilingual retrieval was minimal due to lack of resources. We submitted runs for the English-to-Japanese task.

Our approach consisted of building a translation layer on top of our monolingual search engine without changing the underlying search engine. We used the online resource Babelfish [1] to implement the English-to-Japanese translation layer and programmatically integrated these tools into our workflow.

Using past NTCIR queries, we noticed that Babelfish failed to translate some words, mostly proper names or technical terms. In such cases, we expanded queries by adding the Katakana transliteration of the non-translated terms. We used transliteration because the Japanese language typically transliterate foreign words based on phonetics into Katakana words. We used expansion because 'latin' words are often found in Japanese news documents. We use the ICU library to support transliteration [4].

4.1 Results

We report our submitted and base runs in Table 4. Overall, our results are less than satisfactory. Bilingual runs achieve between 33% and 43% of the monolingual runs. However, the use of transliteration was beneficial: it enhanced retrieval performance although the difference is not always significant.

During our analysis, we have found several issues relating to our query translation and search. First, we noticed that machine translation has difficulty translating the short TITLE queries as these queries were not proper sentences but rather lists of terms and phrases. In particular, we found that the translated queries had little in common with the original Japanese queries. Also, we noticed that our query formulation did not correctly remove translated noise phrases (e.g. "find documents") as the translation did not map to our monolingual resources.

Future work may investigate alternative translation techniques, in particular dictionary and corpus-based as they seem more appropriate to the translation of short queries. This may also include some limited work on translation disambiguation to select the appropriate translations for a given query context.

5 Conclusions

While we can imagine a future in which constructing an Information Retrieval system for a novel language or pair of languages would be a simple matter of piecing together the appropriate components for linguistic analysis and data retrieval, we are clearly not at that point yet. Our attempts at using our well-tested and proven search engine with industry-leading translation and linguistic analysis components did not result in competetive overall retrieval systems. We speculate on the reasons behind this: general translation systems are tailored for natural language translation, not

	Relax				Rigid			
Run	MAP	Above med	Equal med	Below med	MAP	Above med	Equal med	Below med
tlrrd-J-J-D-01	0.3596⋆	17	0	30	0.2681⋆	17	3	27
tlrrd-J-J-D-02	0.3827⋆	24	0	23	0.2891⋆	20	0	27
tlrrd-J-J-D-05	0.3113	13	0	34	0.2375	10	5	32
tlrrd-J-J-T-03	0.3672	22	0	25	0.2694	22	0	25
tlrrd-J-J-T-04	0.3321	10	0	37	0.2366	11	0	36

Table 1. Japanese runs: average precision and comparison to the median. Comparison to median is performed with median information specific to T and D runs, as appropriate. We denote statistically significant differences between the base run and the corresponding PRF run by a ⋆. We used the paired t-test with a p-value of 0.01.

	Relax	Rigid
Run	MAP	MAP
tlrrd-J-J-D-01 without selection	0.3638	0.2734
tlrrd-J-J-D-01	0.3596	0.2681
tlrrd-J-J-D-02	0.3827	0.2891
tlrrd-J-J-D-02 with selection	0.3840	0.2888

Table 2. Japanese runs: average precision with and without selective pseudo-relevance feedback.

	Relax				Rigid			
Run	MAP	Above med	Equal med	Below med	MAP	Above med	Equal med	Below med
tlrrd-K-K-D-01	0.3313	6	1	43	0.2985	6	1	44
tlrrd-K-K-T-02	0.3051	9	0	41	0.2774	13	0	37

Table 3. Korean runs: average precision and comparison to the median. Comparison to median is performed with median information specific to T and D runs, as appropriate.

	Relax				Rigid			
Run	MAP	Above Med	Equal Med	Below Med	MAP	Above Med	Equal Med	Below Med
tlrrd-E-J-T-01	0.1336	6	0	41	0.1018	10	0	37
tlrrd-E-J-T-01, no transliteration	0.0928				0.0709			
tlrrd-E-J-D-02	0.1023	4	3	40	0.0919	5	5	37
tlrrd-E-J-D-02, no transliteration	0.0975				0.0738			

Table 4. Bilingual runs: average precision and comparison to the median. Comparison to median is performed with median information specific to T and D runs, as appropriate.

search-query translation; commercial linguistic analysis tools are not yet sufficiently well-tested in the field; search engines developed for one set of languages may require significant re-tuning to perform effectively on another set of languages. Overall, it seems clear that developing retrieval systems in new domains still requires significant investment of resources.

Acknowledgements

The authors would like to thank the NTCIR-5 organizers for their effort.

References

[1] http://babelfish.altavista.com.

[2] W. B. Croft, J. Callan, and J. Broglio. The inquery retrieval system. In *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, Spain, 1992.

[3] D. Haines and W. Croft. Relevance feedback and inference networks. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[4] http://icu.sourceforge.net.

[5] http://www.inxight.com/products/oem/linguistx.

[6] K. Kishida, K.-H. Chen, S. Lee, K. Kuriyama, N. Kando, H.-H. Chen, and S. H. Myaeng. Overview of clir task at the fifth ntcir workshop. In *Proceeding of the Fifth NTCIR Worshop*, 2005.

[7] I. Moulinier and K. Williams. Thomson Legal and Regulatory experiments at CLEF-2005. In *Workshop notes of the 2005 Cross-Language Information Retrieval and Evaluation Forum*, 2005.

[8] H. Turtle. Natural language vs. boolean query evaluation: a comparison of retrieval performance. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.