Clairvoyance CLEF-2003 Experiments

Yan Qu, Greg Grefenstette, David A. Evans
Clairvoyance Corporation
5001 Baum Boulevard, Suite 700
Pittsburgh, PA 15213-1854
{yqu, grefen, dae}@clairvoyancecorp.com

Abstract

In CLEF 2003, Clairvoyance participated in the bilingual retrieval track with the German and Italian language pair. As we did not have any German-to-Italian translation resources, we used the Babel Fish translation service provided by Altavista.com for translating German topics into Italian, with English as a pivot language. Then the translated Italian topics were used for retrieving Italian documents from the Italian document collection. The translated Italian topics and the document collections were indexed using three different kinds of units: (1) linguistically meaningful units, (2) character 6-grams, and (3) a combination of 1 and 2. We submitted three automatic runs with the three indexing units.

1. Introduction

Clairvoyance participated in the CLEF 2003 bilingual retrieval track using the German and Italian language pair. As we did not have German-to-Italian translation resources, we used the free Babel Fish translation service provided by Altavista.com for translating German topics into Italian, with English as a pivot language. The resulting translated Italian topics were used for retrieving Italian documents from the Italian document collection. The translated Italian topics and the document collections were indexed using three different kinds of features: (1) linguistically meaningful units (e.g., words and NPs), (2) character 6-grams, and (3) a combination of 1 and 2. We submitted three automatic runs, each based on one of the three indexing units. In the following sections, we describe the details of our submission and present the performance results.

2. CLARIT Cross-Language Information Retrieval

In CLEF 2003, we adopted query translation as the means for bridging the language gap between the query language and the document language for cross-language information retrieval. For German-to-Italian information retrieval, first, a German query string was translated into Italian via machine translation; then the translated Italian topics were used for retrieving Italian documents from the Italian document collection. For query and document processing, we used the CLARIT system [1], in particular, those components encompassing a newly developed Italian NLP module (for extracting Italian phrases), indexing (term weighting and phrasal decomposition), retrieval, and "thesaurus extraction" (for extracting terms to support pseudo relevance feedback).

2.1 Query Translation via a Pivot Language

The Babel Fish translation service (altavista.com) provides translation between selected language pairs including German-to-English and English-to-Italian. It does not provide translation service between German and Italian directly. So we used English as a pivot language, first translating the German topics to English and then translating the English topics into Italian. As an illustration of typical results for this process, Figure 1 provides the translations from Babel Fish for Topic 141.

Even though there was increased degradation in query quality after translation, we felt that, except for translation of proper names, the quality of the translation from German to English and from English to Italian by Babel Fish was adequate for the purpose of cross-language information retrieval. We quantitatively evaluate this impression in Section 3.

2.2 Italian Topic Processing

Once the topics were translated into Italian, we extracted two types of terms from the topics: (1) linguistically meaningful units or character n-grams.

To extract linguistically meaningful units, we used CLARIT Italian NLP. This NLP module makes use of a lexicon and finite-state grammar for extracting phrases such as NPs. The lexicon is based on the Multext Italian

(1) Original German topic from CLEF-2003:

Briefbombe für Kiesbauer .
Finde Informationen über die Explosion einer Briefbombe im Studio der Moderatorin Arabella Kiesbauer beim Fernsehsender PRO7.

(2) English translation of (1) by Babel Fish:

Letter bomb for gravel farmer. Find information about the explosion of a letter bomb in the studio of the host Arabella gravel farmer with the television station PRO7.

(3) Ideal English topic from CLEF-2003:

Letter Bomb for Kiesbauer
Find information on the explosion of a letter bomb in the studio of the TV channel PRO7 presenter Arabella Kiesbauer.

(4) Italian translation of (2) by Babel Fish:

Bomba della lettera per il coltivatore della ghiaia. Trovi le informazioni sull'esplosione di una bomba della lettera nell'studio del coltivatore della ghiaia di Arabella ospite con la stazione PRO7 della televisione.

(5) Ideal Italian topic from CLEF-2003:

Lettera Bomba per Kiesbauer
Recupera le informazioni relative all'esplosione di una lettera bomba nello studio della presentatrice della rete televisiva PRO7.

Figure 1: Topic 141 and its translations from Babel Fish

lexicon[1], which was expanded by adding punctuations and special characters. In addition, entries with accented vowels were duplicated by substituting the accented vowels with their corresponding unaccented vowels followed by an apostrophe (“’”). The final lexicon contained about 135,000 entries. An Italian stop word list[2], which contained 433 entries, was used to filter out stop words. The grammar specified the rules for constructing phrases, especially NPs, and morphological normalization rules for normalizing morphological variants to their root forms, e.g., “previsto” to “prever”. In CLEF 2003 experiments, we extracted Adjectives, Verbs, and NPs as indexing terms.

Another way to construct terms is to use overlapping character n-grams. We have observed that our lexicon-based term extraction did not have complete coverage for morphological normalization. The n-gram approach we adopted was aimed at mitigating such an effect. For the submissions, we have used overlapping 6-grams, as it was previously reported to be effective [2]. Spaces and punctuations were included in the character 6-grams.

2.3 CLARIT Indexing and Retrieval

CLARIT indexing involves statistical analysis of a text corpus and construction of an inverted index, with each index entry specifying the index word and a list of texts. CLARIT allows the index to be built upon full documents or variable-length subdocuments. We used subdocuments as the basis for indexing and document scoring in our experiments. The size of a subdocument was in the range of 8 sentences to 12 sentences.

CLARIT retrieval is based on the vector space retrieval model. Various similarity measures are supported in the model. For CLEF 2003, we used the dot product function for computing similarities between a query and a document:

$$sim(P, D) = \sum_{t \in P \cap D} W_P(t) \cdot W_D(t).$$

where $W_P(t)$ is the weight associated with the query term t and $W_D(t)$ is the weight associated with the term t in the document D. The two weights were computed as follows:

[1] http://www.lpl.univ-aix.fr/projects/multext/LEX/LEX.SmpIt.html

[2] Obtained from http://www.unine.ch/Info/clef/

$$W_D(t) = TF_D(t) \cdot IDF(t).$$

$$W_P(t) = C(t) \cdot TF_P(t) \cdot IDF(t)$$

where IDF and TF are standard inverse document frequency and term frequency statistics, respectively. *IDF(t)* was computed with the target corpus for retrieval. The coefficient *C(t)* is an “importance coefficient”, which can be modified either manually by the user or automatically by the system (e.g., updated during feedback).

2.4 Post-Translation Query Expansion

Query expansion through (pseudo) relevance feedback has proved to be effective for improving IR performance [3]. We used pseudo relevance feedback for augmenting the queries. After retrieving some documents for a given topic from the target corpus, we took a set of top ranked documents, regarding them as relevant documents to the query, and extracted terms from the these documents. The terms were ranked based on the following formula:

$$Prob2(t) = log(R_t + 1) \;\; x \left(log(\frac{N - R + 2}{N_t - R_t + 1} - 1) - log(\frac{R + 1}{R_t} - 1) \right)$$

where N is the number of documents in the target corpus, N_t is the number of documents in the corpus that contain term t, R is the number of documents for feedback that are (presumed to be) relevant to the topic, and R_t is the number of documents that are (presumed to be) relevant to the topic and contain term t.

3 Experiments

We submitted three automatic runs to CLEF 2003. All the queries used the title and description fields (Ttitle+Description) of the topics provided by CLEF 2003. The results presented below are based on relevance judgments of 42 topics, which have relevant documents in the Italian corpus. The three runs were:

- ccwrd: with linguistically meaningful units as indexing terms
- ccngm: with character 6-grams as indexing terms
- ccmix: a combination of linguistic units and character 6-grams as indexing units

With the ccmix run, the combinations were constructed through a simple concatenation of the terms nominated by ccwrd and ccngm. We ran Italian monolingual experiments to obtain the baseline with ideal translations after obtaining the relevance judgments from CLEF 2003.

All the experiments were run with post-translation pseudo relevance feedback. The feedback-related parameters were based on training over CLEF 2002 topics. The settings for German-to-Italian retrieval were: extracting T=80 terms from the top N=25 retrieved documents with the Prob2 method. For the n-gram based indexing and the mixed model, an additional term cutoff percentage set to P=0.01. For the word-based indexing, the percentage cutoff is set to P=0.25. For Italian monolingual retrieval with words as indexing terms: T=50, N=50, P=0.1. For Italian monolingual retrieval with n-grams and the mixed model as indexing terms: T=80, N=25, P=0.05.

Table 1 presents the results for our submitted runs, and Table 2 presents results for our training runs with CLEF 2002 topics. The monolingual baselines for the ccwrd*, ccngm*, and ccmix* runs are the optimal monolingual runs based on word based indexing, 6-gram based indexing, and mixed indexing, respectively. While the 6-gram based indexing produced higher average precision compared with the word based indexing for the CLEF 2002 topics, it significantly underperformed word based indexing for CLEF 2003 topics. Further examination is required to account for the difference in behavior of the two indexing methods for the two top sets.

Run ID	Indexing Units	Recall	AP	% mono AP
ccwrd	Adj+VP+NPs	541/809	0.2303	67.2% (of 0.3428)
ccngm	Character 6-grams	456/809	0.1624	54.3% (of 0.2993)
ccmix	Adj+VP+NPs, character 6-grams	505/809	0.2098	61.7% (of 0.3402)

Table 1: German-to-Italian retrieval performance with CLEF 2003 topics. All three runs are our submitted runs.

Run ID	Indexing Units	Recall	AP	% mono AP
ccwrd2002	Adj+VP+NPs	643/1072	0.1823	61.6% (of 0.2959)
ccngm2002	Character 6-grams	648/1072	0.2133	68.1% (of 0.3132)
ccmix2002	Adj+VP+NPs, character 6-grams	675/1072	0.2147	60.1% (of 0.3574)

Table 2: German-to-Italian retrieval performance with CLEF 2002 topics.

Topics	2002 Avg. Prec	2003 Avg. Prec
(1) Translated English (from German) to Italian Performance change compared with (2) Performance change compared with (3)	0.1549 (-24.2%) (-36.7%)	0.1748 (-16.1%) (-45.3%)
(2) Ideal English to Italian Performance change compared with (3)	0.2048 (-16.4%)	0.2083 (-34.8%)
(3) Ideal Italian	0.2449	0.3197

Table 3: Performance comparison between different versions of topics

Table 3 presents a comparison between different versions of the Italian topics for CLEF 2002 and CLEF 2003 topics. The average precision statistics were computed with word based indexing and with no feedback. Even through comparing different topic statements is not justified methodologically [4], the comparison gives us a rough estimate of the quality of the translation module. Translation from English to Italian decreased performance in the range of 16.4% to 34.8%, while adding another layer of translation from German to English decreased performance further by 16.1% to 24.2%. This shows that translation service such as Babel Fish still needs to be improved for better CLIR performance.

4 Conclusions

Due to the lack of resources, our participation in CLEF 2003 was limited. We succeeded in submitting three runs for German-to-Italian retrieval, examining word based indexing and n-gram based indexing. Our results with CLEF 2002 and CLEF 2003 did not provide firm evidence of which indexing method is better. Future analysis is required in this direction.

References

[1] Evans, D.A., and R.G. Lefferts. CLARIT–TREC Experiments. *Information Processing and Management*, Vol.31, No.3, pp.385–395, 1995.

[2] McNamee, P., and J. Mayfield. Scalable Multilingual Information Access. In C. Peters, editor, *Working Notes for the CLEF 2002 Workshop*, pp.133–140, 2002.

[3] Ballesteros, L., and W. B. Croft. Statistical Methods for Cross-Language Information Retrieval. In G. Grefenstette, editor, *Cross-Language Information Retrieval*, Chapter 3. Kluwer Academic Publishers, Boston, pp.23–40, 1998.

[4] Voorhees, E. The Philosophy of Information Retrieval Evaluation. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems: Proceedings of the CLEF 2001 Workshop, Lecture Notes in Computer Science 2406*, Springer, pp.355–370, 2001.