First participation of University and Hospitals of Geneva to Domain-Specific Track in CLEF 2008

Julien Gobeill, Patrick Ruch
University and Hospitals of Geneva, Switzerland
julien.gobeill@sim.hcuge.ch

Abstract

We participate in 2008 to our first Domain-Specific Track, with the aim to establish a baseline for our Information Retrieval engine in an unknown domain for us. We are specialized in Natural Language Processing in the biomedical domain, and we participate to the medical Image track and to TREC Genomics for four years with textual strategies, as queries expansions with controlled vocabularies, pattern recognition and vectorial space models. The technical component of our cross-language search engine is a generic toolkit, EasyIR, with which we can perform Text Categorization and Information Retrieval. The strategy applied for the 2008 Domain-Specific track is as simple as possible, as we want only to establish a baseline for EasyIR in a new track. For the English monolingual task, we choose to work with the title, the descriptive text and some types of classification terms to index documents. For the German queries to English collection bilingual task, we choose to perform a simple retrieval on the German collection in one hand, and to collect the descriptors of the retrieved documents in order to make cross-lingual query expansion in the other hand. Unfortunately, our results cannot be seen as fair, as we achieve MAP of 0.171 for the monolingual task and MAP of 0.132 for the bilingual task. Nevertheless, comparing to several baseline runs of other participants for DS CLEF 2007, our baseline run achieves equal performances. Possibilities to improve for the next DS CLEF are best tuning of our system with the benchmark, and an efficient use of the controlled vocabularies.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [Database Management]: Languages - Query Languages

General terms

Measurement, Performance, Experimentation

Keywords

Image Retrieval, Text categorization, multimodal retrieval

1 Introduction

The Cross Language Evaluation Forum (CLEF) is a challenge which occurs each year since 2000. The goal of this challenge is to evaluate the participants on a common multilingual task, to establish a state of the art of the techniques used in a domain, and to build a benchmark for future evaluations. The Domain-Specific (DS) Track has started since 2000 with the goal to retrieve relevant documents in a scientific documents structured collection. The DS Task is for few years focused on bibliographic databases in the social sciences domain. The goal of this task is to retrieve relevant documents, in respect to a query, in a multilingual collection, using titles, abstracts and human-assigned descriptors (1).

Our team is specialized in Natural Language Processing in the biomedical domain, as we regularly participate to the TREC Genomics Track (2; 3) and to the ImageCLEF medical retrieval Track (4; 5). In these challenges, we usually use simple textual strategies with thesaural resources in order to compose our runs. The

technical component of our cross-language search engine is a generic toolkit, EasyIR, which can perform Text Categorization at high precision for high rank (6) – above 90% for Medical Subject Headings terms – and Information Retrieval. Our first participation to the 2008 DS Track is motivated by the aim to establish a baseline for our Information Retrieval engine in an unknown domain for us, where some controlled vocabularies can be used for query expansion and more efficient retrieval.

We participate to the English monolingual task, and to the German queries to English collection bilingual task. As the aim of our first participation is only to obtain a baseline evaluation of our engine in this track, we only submit one run per task, with the simplest possible strategy.

2 Data and Strategies

The 2008 collection is the same as in 2007. The concerned collection for the tasks we participate – English monolingual and German queries to English collection bilingual tasks – comprises documents from two different sources. On one hand, the German Indexing and Retrieval Testdatabase in its forth version (GIRT-4 German) contains 151,319 German documents dealing with social science and covering the years 1999-2000; a pseudo-parallel English version of this collection, GIRT-4 English, contains the same documents translated in English. On the other hand, the social science database Sociological Abstracts from Cambridge Scientific Abstracts (CSA-SA) contains 20,000 documents, covering the years 1994-1996.

A typical composition of a document contains different useful features for indexing, as title, author names, type of document, and publication date. An abstract is present for 96% of the GIRT-4 German documents and 94% of the CSA documents – but only for 17% of the GIRT-4 English translated documents. Additional thesaurus descriptors and classification codes belonging to controlled vocabularies are manually added to each document. For the GIRT-4 collections, descriptors are issued from the GESIS IZ Thesaurus; for the CSA-SA collection, they are issued from the CSA Thesaurus of Sociological Indexing Terms. See figures 1-3 for an example of a document for each collection.

```
 <DOC>
    <DOCID>CSASA-1-EN-9600289</DOCID>
    <TITLE-EN>Structural Tightness and Social Conformity: Varying the Source of
External Influence</TITLE-EN>
    <AUTHOR>Roberts, Lance W.</AUTHOR>
    <AUTHOR>Boldt, Edward D.</AUTHOR>
    <AUTHOR>Guest, Anne</AUTHOR>
    <AUTHOR-AFFILIATION>Dept Sociology U Manitoba, Winnipeg R3T 2N2</AUTHOR-
AFFILIATION>
    <DOCTYPE>Abstract of Journal Article</DOCTYPE>
    <PUBLICATION-YEAR>1990</PUBLICATION-YEAR>
    <COUNTRY-CODE>US</COUNTRY-CODE>
    <CONTROLLED-TERM-EN>Hutterites</CONTROLLED-TERM-EN>
    <CONTROLLED-TERM-EN>Conformity</CONTROLLED-TERM-EN>
    <CONTROLLED-TERM-EN-MINOR>Manitoba</CONTROLLED-TERM-EN-MINOR>
    <CONTROLLED-TERM-EN-MINOR>College Students</CONTROLLED-TERM-EN-MINOR>
    <CLASSIFICATION-TEXT-EN>social psychology; personality and social roles
(individual traits, social identity, adjustment, conformism, and
deviance)</CLASSIFICATION-TEXT-EN>
    <FREE-TERM-EN>social conformity, structural tightness thesis; test data;
Hutterites/undergraduates, Manitoba;</FREE-TERM-EN>
    <TEXT-ENG>Structural tightness is defined as the capacity to impose collective
role expectations on community members. An attempt is made to reconceptualize this
term so that the findings in a cross-cultural conformity study may be brought into
a different light. Theoretical considerations are made in order to break down an
ecocultural model provided by others working in the field. It is this
conceptualization that puts forth the original definition of structural tightness
that is debated. To test these notions, test data were obtained from ethnic
Hutterites and 51 undergraduates in Manitoba. Findings suggest that the
theoretical rationale put forth is plausible and support the proposed
reconceptualization.</TEXT-ENG>
  </DOC>
```

Figure 1: example of a document from the CSA-SA collection.

```
<DOC>
<DOCNO>GIRT-DE19909343</DOCNO>
<DOCID>GIRT-DE19909343</DOCID>
<TITLE-DE>Die sozioökonomische Transformation einer Region : Das Bergische Land
von 1930 bis 1960</TITLE-DE>
<AUTHOR>Henne, Franz J.</AUTHOR>
<AUTHOR>Geyer, Michael</AUTHOR>
<PUBLICATION-YEAR>1990</PUBLICATION-YEAR>
<LANGUAGE-CODE>DE</LANGUAGE-CODE>
<CONTROLLED-TERM-DE>Rheinland</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>historische Entwicklung</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>regionale Entwicklung</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>sozioökonomische Faktoren</CONTROLLED-TERM-DE>
<METHOD-TERM-DE>historisch</METHOD-TERM-DE>
<METHOD-TERM-DE>Aktenanalyse</METHOD-TERM-DE>
<CLASSIFICATION-TEXT-DE>Sozialgeschichte</CLASSIFICATION-TEXT-DE>
<ABSTRACT-DE>Die Arbeit hat das Ziel, anhand einer regionalen Studie die
Entstehung des "modernen" fordistischen Wirtschaftssystems und des sozialen
Systems im Zeitraum zwischen 1930 und 1960 zu beleuchten; dabei geht es auch um
das Studium des "Sozial-imaginären", der Veränderung von Bewußtsein und Selbst-
Verständnis von Arbeitern durch das Erlebnis und die Erfahrung der Depression, des
Nationalsozialismus und der Nachkriegszeit, welches sich in den 1950er Jahren
gemeinsam mit der wirtschaftlichen Veränderung zu einem neuen "System"
zusammenfügt.</ABSTRACT-DE>
</DOC>
```

Figure 2: example of a document from the GIRT-4 German collection.

```
<DOC>
<DOCNO>GIRT-EN19901932</DOCNO>
<DOCID>GIRT-EN19901932</DOCID>
<TITLE-EN>The Socio-Economic Transformation of a Region : the Bergische Land from
1930 to 1960</TITLE-EN>
<AUTHOR>Henne, Franz J.</AUTHOR>
<AUTHOR>Geyer, Michael</AUTHOR>
<PUBLICATION-YEAR>1990</PUBLICATION-YEAR>
<LANGUAGE-CODE>EN</LANGUAGE-CODE>
<CONTROLLED-TERM-EN>Rhenish Prussia</CONTROLLED-TERM-EN>
<CONTROLLED-TERM-EN>historical development</CONTROLLED-TERM-EN>
<CONTROLLED-TERM-EN>regional development</CONTROLLED-TERM-EN>
<CONTROLLED-TERM-EN>socioeconomic factors</CONTROLLED-TERM-EN>
<METHOD-TERM-EN>historical</METHOD-TERM-EN>
<METHOD-TERM-EN>document analysis</METHOD-TERM-EN>
<CLASSIFICATION-TEXT-EN>Social History</CLASSIFICATION-TEXT-EN>
</DOC>
```

Figure 3: the corresponding document to figure 2 in the GIRT-4 English collection.

Strategy for the English monolingual task. We choose to perform a simple Information Retrieval process for this task. For the GIRT-4 English collection, the title, abstract, controlled terms and classification texts are concatenated in a bag of words in order to index each document. For the CSA-SA collection, the title, text and classification texts are used in a same way. The keystone of our strategy in ImageCLEF and TREC Genomics is the automatic assignments of descriptors to documents and queries, in order to synthesize the concepts of a document in a kind of intermediate language (7). As human-generated keywords are already associated with each documents in the DS Track collection, and as we have no expertise of these bibliographic controlled vocabularies – and as the submitted runs are supposed to establish a baseline and to be as simple as possible – we choose to not work deeply with controlled vocabularies terms for document indexing. Moreover, when studying the Working Notes of the previous DS Track, we choose to not to use the controlled vocabularies in order to make query expansions, as several participating teams report than this technique leads no significant improvements (1; 8; 9). Therefore, this run is as basic as we could.

Strategy for the German queries to English collection bilingual task. For this task, our strategy is lightly more sophisticated. As GIRT-4 offers translated version across German and English, we firstly choose to perform a simple Information Retrieval process in the GIRT-4 German collection, in respect to the German

queries, in order to obtain a first ranking. We don't perform any translation of the queries, whereas it seems to be an effective strategy in the previous DS Track (1). Then, we use this ranking in order to make query expansion: for each query, we select the 10 most relevant retrieved documents in GIRT-4 German, and then we parse their corresponding documents and descriptors in the GIRT-4 English in order to extract the 5 most frequent English descriptors. These English descriptors are added to the queries in order to perform a second retrieval in the CSA-SA corpus in order to obtain a second ranking. The two ranking are then normalized and merged into a final ranking, with weights of 75% for the first ranking and 25% for the second one. We don't use at any time the provided vocabulary mappings.

3 Methods

Two main modules constitute the skeleton of EasyIR, our Information Retrieval engine: the regular expression component, and the vector space component. Each of the basic classifiers implements known approaches to document retrieval. The first tool is based on a regular expression pattern matcher (10). The second classifier is based on a vector space engine. This second tool is expected to provide high recall in contrast to the regular expression-based tool, which should privilege precision. The former component uses tokens as indexing units and can be merged with a thesaurus, while the latter uses stems (Porter). See (11) for more precisions about our engine.

The mean average precision (map): is the main measure for evaluating ad hoc retrieval tasks (for both monolingual and bilingual runs). Following (12), we also use this measure to tune the Information Retrieval system. We use the parameters obtained by a previous tuning on a small set of OHSUMED abstracts: 1200 randomly selected abstracts were used to select the weighting parameters of the vector space classifier and the best combination of these parameters with the regular expression-based classifier.

4 Results and Discussion

We then describe each task separately.

4.1 English monolingual task

For this task, our run achieves a R-precision of 22.69%, and a map of 17.14%. These performances make us the lasts of the two rankings and are relatively far from the best ones (respectively around 40% for R-precision and 38% for map). This could be considered relatively weak, but once again, the aim of our participation is only to establish a baseline with simple methods in this DS track. Nevertheless, a closer look to the previous DS Track Working Notes shows that several teams participating to DS Track this year submitted last year equivalent runs (13; 14), even if the two Tracks cannot be directly compared as queries have changed. We assume that the performance of our run is fair relatively to our expertness and our background in this domain, and that we will be able to submit more efficient runs in the future DS Tracks.

4.2 German queries to English collection bilingual task

The result of this task is quite similar. Our run achieves a R-precision of 18.80% – which is not the worst R-precision of the ranking – and a map of 17.14%. As for the English monolingual task, we find several runs with equivalent performances in the previous DS Track. As we didn't tune our system, and we didn't use strong use of the controlled vocabularies and their mapping, we assume once again that we have a lot of room for improvement for the future evaluations.

5 Conclusion and Future Work

For the future DS Track, we need to invest more time in an efficient tuning of our engine with the previous benchmark. A more in-depth state of the art of the successful techniques used this year, followed by a more

efficient use of the controlled vocabularies in order to make query expansion, and automatic translations of queries, should be planned too.

References

1. *The Domain-Specific Track at CLEF 2007.* **V Petras, S Baerisch, M Stempfhuber.** CLEF 2007 Proceedings.
2. *TREC 2007 Genomics Track Overview.* **W Hersh, A Cohen, L Ruslen, P Roberts.** TREC 2007 Proceedings.
3. *Vocabulary-driven Passage Retrieval for Question-Answering in Genomics.* **J Gobeill, F Ehrler, I Tbahriti, P Ruch.** TREC 2007 Proceedings.
4. *Overview of the ImageCLEF 2007 Medical Retrieval and Annotation Tasks.* **H Muller, T Deselaers, E Kim, J Kalpathy–Cramer,T M Deserno, W Hersh.** ImageCLEF 2007 Proceedings.
5. *University and Hospitals of Geneva at ImageCLEF 2007.* **X Zhou, J Gobeill, P Ruch and H Muller.** CLEF 2007 Working notes.
6. *Automatic Assignment of Biomedical Categories: Toward a Generic Approach.* **Ruch, P.** 22(6), 2006, Bioinformatics, pp. 658-64.
7. *Query and Document Translation by Automatic Text Categorization: A Simple Approach to Establish a String Textual Baseline for ImageCLEFmed 2006.* **J Gobeill, H Muller and P Ruch.** 2006. ImageCLEF.
8. *Experiments in Classification Clustering and Thesaurus Expansion for Domain Specific Cross-Language Retrieval.* **Larson, R R.** CLEF 2007 Proceedings.
9. *Domain-Specific IR for German, English and Russian Languages.* **C Fautsch, L Dolamic, S Abdou, J Savoy.** TREC 2007 Proceedings.
10. *A tool to search through entire file systems.* **Wu, U Mamber and S.** Proceedings of the USENIX Winter 1994 Technical Conference, San Francisco, pp. 23-32.
11. *Learning-Free Text Categorization.* **P Ruch, R Baud, and A Geissbuhler.** 2003, LNAI 2780, pp. 199-208.
12. *Combining classifiers in text categorization.* **Croft, L Larkey and W.** 1996, SIGIR, ACM Press, New York, pp. 289-297.
13. *Domain-Specific Cross Language Retrieval: Comparing and Merging Structured and Unstructued Indices.* **Eibl, J Kursten and M.** TREC 2007 Proceedings.
14. *XRCE's Participation to CLEF 2007 Domain-specific Track.* **Renders, S Clinchant and JM.** TREC 2007 Proceedings.