Integration of Arabic to a Cross-Lingual Retrieval Tool: Challenges and Perspectives

René Schneider, Thomas Mandl, and Christa Womser-Hacker

University of Hildesheim
Marienburger Platz 22
D-31141 Hildesheim
{rschneid,mandl,womser}@uni-hildesheim.de

Abstract

The ambition of this paper is to resume briefly the challenges that Arabic offers in Cross-Lingual Information Retrieval, to show the potential of MIMOR[1], a retrieval system, that has proved to be succesful for cross-lingual retrieval tasks, and to propose string matching techniques for feature unification instead of stemming techniques.

Keywords

Information Retrieval, Multi-Linguality, Fusion, Merging, String Matching techniques

1 Introduction

Recent years have shown an increasing interest for Arabic in several fields of natural language processing. One of these fields is information retrieval, where a query that is formulated by a user is matched with objects of any media in a data base. In one specific case of information retrieval, the user has an interest in information (e.g. text documents) that are formulated in a language that he does not know at all. This task can be solved by so called cross-lingual retrieval engines where queries and/or results are translated crosswise to overcome this gap. This problem becomes even more tedious whenever the writing system, the language family, or language typology between target language and source language differ from each other.

Arabic, as a root-inflecting language of the hamito-semitic language family with its own writing system offers all these challenges whenever it is matched against an indo-european language like French or German. Since the translation of the whole collection of text documents cannot be translated automatically simply due to their vast amount, machine translation has to be restricted to queries and/or the results selected by the user after being retrieved as similar or relevant for his information needs. This paper shows the potential of MIMOR, a retrieval tool that has proved to be succesful in cross-lingual information retrieval

[1] Multiple Indexing and Method-Object Relations

and outlines a project plan for the integration of Arabic into the system. Special attention will be given to string matching techniques and multiple sequence alignemt to overcome the fact that n-grams and stemming techniques are less fruitful for cross-lingual retrieval with Arabic.

2 Cross-Lingual Information Retrieval for Arabic

Former works in the field of cross-lingual information retrieval for Arabic (Hasnah/Evens. 2001, Labed Jilani/Hauoala 2001, Xu/Fraser/Weischedel 2002) have pointed out the challenges that Arabic offers: a) orthographic variations, b) the complexity of the morphology, c) the prevelance of broken plurals, d) the ambiguity of the words, e) the omission of short vowels, and f) the wide use of synonyms. Therefore, several techniques (e.g. keyword match, n-grams, stemming and spelling normalization) that are widely used in information retrieval have low impact in cross-lingual retrieval for Arabic.

Besides that the major focus on cross-lingual information retrieval with Arabic has been on the language pairs English-Arabic and French-Arabic (see Oard 2002, www.hahooa.com). Since the full integration of linguistic resources is time and effort consuming, we have therefore decided a) to make use of online-available linguistic ressources as offered e.g. by tarjim.ajeeb.com using English as a pivot-language[2], b) to firstly concentrate on language independent empirical strategies c) to enrich this empirical approach with string matching techniques that are able to catch the specific structure of root-inflecting languages.

3 The MIMOR Model

MIMOR is modelled as an open information retrieval system which is able to combine individual approaches of information retrieval within one meta system and which could be expanded at different positions over time. On the one hand the performance of special retrieval devices can be explored and on the other hand additions can be made at any time. This joint efforts seem to be very fruitful for both perspectives.

MIMOR integrates users' relevance assessments to learn which combinations of object representations and IR functionality lead to good performance of the overall system. An internal evaluation procedure which is realized via a blackboard model permanently registers which resource produces good results and which one does not. Good techniques gain high weights, bad ones are extinguished over time.

3.1 Basic assumptions

The MIMOR model takes advantage of the main outcomes of TREC[3]. One of the most important result from this study is that many information retrieval systems perform similarly well in terms of recall and precision but do not lead to the same sets of documents. This means that the systems find the same percentage of relevant documents, but the overlap between their results often is low. Because of these findings, fusion seems to be a promising strategy. This is confirmed by many application experiments of fusion techniques in information retrieval (e.g. McCabe, Chowdhury, Grossmann & Frieder 1999). It turned out that fusion of various information retrieval techniques can improve the overall quality of a system. Fusion methods delegate a task to several systems and integrate their results into one

[2] A pivot language is a language used for communication between people with different native languages like a *lingua franca*. Pivot languages can form an instance in the translation process in multilingual environments.

[3] Text Retrieval Conferences

final result presented to the user. In information retrieval, fusion is mostly implemented as a combination of several algorithms where different probabilities for the relevance of a document query relation are integrated into one final similarity measure.

3.2 Formalization and Technical Background

MIMOR's fusion approach takes advantage of heterogeneity or diversity. It samples users' relevance feedback to predict optimal method-object relations where methods are indexing algorithms or retrieval machines. These are assigned to the characteristics of users and documents with the goal of improving the overall retrieval quality. From a computational viewpoint, MIMOR is designed as a linear combination of the results of different retrieval systems. The contribution of each system or algorithm to the fusion result is governed by a weight for that system.

$$RSV_{MIMOR}(doc_i) = \frac{\sum_{system=1}^{N} (\omega_{system}\, RSV_{system}(doc_i))}{N}$$

A central aspect in MIMOR is learning. The weight of the linear combination of each information retrieval system is adapted according to the success of the system measured by the relevance feedback of the users. A system which gave a high retrieval status value (RSV) and consequently a high rank to a document which then received positive relevance feedback should be able to contribute with a higher weight to the final result. The following formula enables such a learning process, which is also illustrated in figure 1:

$$\omega_{system} = \varepsilon\, RF_{user}(doc_i)\; RSV_{system}(doc_i)$$

$$\varepsilon \quad learning\ rate$$

However, the optimal combination may depend on the context and especially on the users' individual perspectives as well as the characteristics of the documents. Therefore, MIMOR needs to consider context.

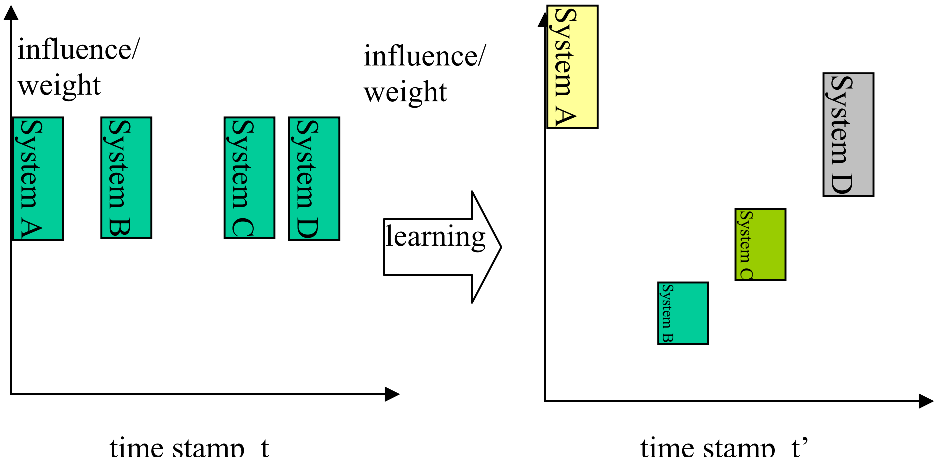

Figure 1: Learning the optimal linear combination.

3.3 Modelling Context via Clusters

The performances of IR systems differ from domain to domain. TREC found out that particular characteristics of the documents relevant for the indexing procedure may be responsible for this effect. In one experiment, for example, optimal similarity functions especially for short queries could be developed (Kwok & Chan 1998). MIMOR is based upon the idea that formal properties of queries and documents can be exploited in order to improve the overall fusion system. Within fusion, the weight of the weight of a system should be high for the type of documents it was optimised for only. Some characteristics of text documents seem to be good candidates for such distinctions. Length, difficulty, syntactic complexity and even layout can be assessed automatically. In MIMOR, these properties are modelled as clusters whereas all documents having a property in common belong to the same cluster. For each cluster an individual MIMOR model is developed with own weights for all participating systems. The clustering process is not restricted to algorithms based on unsupervised learning. Pre-defined classes and even human assignment are compatible with MIMOR.

4 Cross-Lingual Information Retrieval with MIMOR

In CLEF[4] 2003, a fully automatic MIMOR system was applied to cross-language information retrieval with the four languages English, French, German and Spanish using English as source language because most of the web based translation services offer translations to and/or from English. The employed tools included Lucene 1.31[5], MySQL 4.0.122[6] and JAVA TM-based snowball3[7] analyzers.

In a first step after formal pre-processing, customized snowball stemmers were used to stem the data and to eliminate stop words. Then the collection was indexed by Lucene and MySQL and the topics were translated into French, German and Spanish via machine translation tools (FreeTranslation, Reverso and Linguatec5[8]). Examining the various translation tools, it became apparent that the quality of the machine translations was not satisfying, but that, at the same time, the individual translation systems did not show the same weaknesses nor made the same mistakes. Due to this fact, fusion of the various systems was applied as well in the context of multilinguality. According to the respective language the topics were also stemmed and then merged together forming an entire query. The experiments on the training data strongly favoured Lucene as retrieval engine in comparison to MySQL which was reflected in the defaults of weight setting.

To further improve retrieval quality, blind relevance feedback (BRF) was implemented. Expansion terms were selected by applying the Robertson selection value or the Kullback-Leibler (KL) divergence measure (Carpineto et al. 2001). Thus, the submitted runs used BRF KL from the top five documents adding 20 terms.

To gain more insights, a number of additional experiments were conducted beyond CLEF 2003. On the one hand the isolated information retrieval systems were examined and it could be proved that by applying more intensive optimization techniques the solo performance of each individual system could be improved (Hackl et al. 2004). On the other hand it turned out that BRF worked generally well but using the original (perfect) translations of the CLEF queries instead of the automatically translated ones for all four languages, BRF as well as the fusion runs had a negative influence on the overall performance.

[4] Cross-Lingual Evaluation Forum

[5] http://jakarta.apache.org/lucene/docs/index.html

[6] http://www.mysql.com/

[7] http://jakarta.apache.org/lucene/docs/lucene-sandbox/snowball/

[8] FreeTranslation: http://www.freetranslation.com/, Reverso: http://www.reverso.net/, Linguatec Personal Translator: http://www.linguatec.net/online/ptwebtext/index.shtml

5 Multiple Sequence Alignment for Feature Unification

Former studies (e.g. Schneider/Renz 2000) have shown that stemming for information retrieval and information extraction tasks might be replaced by a robust lemmatization procedure that consists of a combination of term frequency and string matching: This approach is based on the empirical investigation of several corpora, that showed that there is a strong connection between the frequency of a word and the probability for being a lemma or root. By this, frequency lists provide information concerning the wellformedness or grammaticality of a word form. This approach is close to previous work (Dichy and Farghaly 2003) insofar as it

To compute the similarity relationship between the different elements of the frequency list, we made use of the Levenshtein distance (Levenshtein 1975) though it represents a useful string matching technique making use of three basic operations, namely the insertion, deletion and substitution of symbols, whereas substitution can be seen as the consecutive application of deletion and insertion. However, since the method is robust both with negative and positive aspects, its disadvantages may be compensated through multiple string alignment techniques.

So far, similarity or distance between the two strings is defined by the minimal number of operations that is necessary for transforming one symbol sequence into another. It may be expressed in a metric value by dividing the total number of transformations through the length of the longest string.

The general cycle of the learning algorithm, is described as follows: the first loop starts with the initial element of the ranked list, i.e. the element with the highest significance or degree--of--interest and compares this element with every succeeding element of the list. Each element, bearing a lower similarity to the top element as indicated by the threshold value is put into one class with the top element as class representative. Both list elements are simultaneously taken out of the list. In the second and all consecutive loops, the algorithm proceeds in the same way, comparing the respective top element with the remaining list elements until the list is shrunken to a group of elements that shows no significant similarity to all the other elements.

The algorithm is deterministic insofar as it does not allow any ambiguity concerning the membership of the elements to a certain class, i.e. any variant, although it might have the same or even a lower similarity to one or several other elements is assigned to the respective initial list element. Thus, emphasis is given to the rank which is also deterministic for the differentiation between the class representative and the different class members. Besides the algorithm does not produce any information loss as a result of stemming, but rather groups similar words around a possibly fully inflected word all words have greatest similarity to. Similarity is calculated by computing the cells and finding the minimal path of a two-dimensional matrix, and is therefore usable for root-inflecting languages too.

The benefit of the lexical clusters generated through the calculation of binary and multiple string distances is twofold: a) the number of features may be reduced considerably by replacing variants through their lemma within the document vectors and b) by expanding short term queries with their generated variants.

It is our ambition to combine pairwise comparision techniques and the derived clusters with multiple sequence alignment techniques (Sankoff/Kruskal 1983) to achieve higher homogenity within the clusters.

6 Conclusions

In this paper we reported ongoing work on the integration of Arabic to an existing retrieval engine (MIMOR) that has recently started and put a special focus on the use and benefits of string matching techniques and multiple sequence alignment to enable feature unification and feature reduction in cross-lingual information retrieval. Although many questions remain open, we think that these two methodologies will be a solid fundament for the integration of a new language into an existing retrieval system.

References

Baeza-Yates, R., / Ribeiro-Neto, B. (eds.) (1999), *Modern Information Retrieval.* Harlow et al.. Addison-Wesley.

Carpineto, C.; de Mori, R.; Romano, G.; Bigi, B. (2001), "An Information-Theoretic Approach to Automatic Query Expansion" in: *ACM Transactions on Information Systems*, 19(1) 1-27.

Dichy, J. / A. Farghaly (2003) "Roots & Patterns vs. Stems: on what grounds should a multilingual database centred on Arabic be built?", in *Proceedings of the MT Summit IX Workshop on Machine Translation for Semitic Languages: Issues and Approaches*, September 23, 2003, New Orleans, Louisiana, U.S.A.

Hackl, R./ Kölle, R./ Mandl, T./ Plödt, A./ Scheufen, J.H. / Womser-Hacker, C. (2004), „Multilingual Retrieval Experiments with MIMOR at the University of Hildesheim" In: Peters, C., Braschler, M., Gonzalo, J. & Kluck, M. (eds.): *Evaluation of Cross-Language Information Retrieval Systems. Proceedings of the CLEF 2003 Workshop*. Berlin et al.: Springer [LNCS], to appear.

Hasnah, A. / Evens, M. (2001), "Arabic/English Cross Language Information Retrieval Using a Bilingual Dictionary", in: *Proceedings of the ACL/EACL 2001 Workshop on Arabic Language Processing: Status and Prospects*, July 6, 2001, Toulouse, France.

Labed Jilani, L. / Haouala, H. (2001), "A Tool for Arabic Documents Indexing and Retrieval from a Web Virtual Library", in: *Proceedings of the ACL/EACL 2001 Workshop on Arabic Language Processing: Status and Prospects*, July 6, 2001, Toulouse, France.

Levenshtein, V. (1975), "On the Minimal Redundancy of Binary Error-Correcting Codes" *Information and Control* 28 (4): 268-291.

McCabe, M.C./ Chowdhury, A./ Grossmann, D. / Frieder, O. (1999), "Unified Framework for Fusion of Information Retrieval Approaches", in: *Proc. Eighth ACM Conference on Information and Knowledge Management (CIKM)*. Kansas City, Missouri, USA. pp. 330-334.

Oard, D. / Gey, Fredric (2002): The TREC-2002 Arabic/English CLIR Track. http://citeseer.nj.nec.com/578170.html, retrieved 01-15-2004.

Sankoff, D. / Kruskal, J.(1983) , *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley.

Schneider R. / Renz I. (2000), "The Relevance of Frequency Lists for Error Correction and Robust Lemmatization", in: *Actes JADT 2000 (= 5es Journées Internationales d'Analyse Statistique des Données Textuelles)*, Lausanne (Suisse), pp. 43-50.

Womser-Hacker, C. (1997), *Das MIMOR-Modell. Mehrfachindexierung zur dynamischen Methoden-Objekt-Relationierung im Information Retrieval.* Universität Regensburg, Informationswissenschaft, Habilitationsschrift.

Xu, Jinxi / Fraser, Alexander / Weischedel Ralph (2002)., "Arabic Information Retrieval: Empirical studies in strategies for Arabic Retrieval", in: *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval.*