Université de Neuchâtel
Faculté des Sciences
Institut d'Informatique

Influence of Language Morphological Complexity on Information Retrieval

par

Ljiljana Dolamic

Thèse

présentée à la Faculté des Sciences
pour l'obtention du grade de Docteur ès Sciences

Acceptée sur proposition du jury

Prof. Jacques Savoy, directeur de thèse
Université de Neuchâtel, Suisse

Prof. Martin Braschler, rapporteur
Zürcher Hochschule für Angewandte Wissenschaften, Winterthur, Suisse

Prof. Fabio Crestani, rapporteur
Universitá della Swizzera Italiana, Lugano, Suisse

Prof. Peter Kropf, rapporteur
Université de Neuchâtel, Suisse

Soutenue le 22 janvier 2010

Université de Neuchâtel

FACULTE DES SCIENCES
Secrétariat-Décanat de la faculté
- Rue Emile-Argand 11
- CP 158
- CH-2009 Neuchâtel

IMPRIMATUR POUR LA THESE

Influence of Language Morphological Complexity on Information Retrieval

Ljiljana DOLAMIC

UNIVERSITE DE NEUCHATEL

FACULTE DES SCIENCES

La Faculté des Sciences de l'Université de Neuchâtel,
sur le rapport des membres du jury

MM. J. Savoy (directeur de thèse), P. Kropf,
M. Braschler (Zürich University of Applied Sciences)
et Fabio Crestani (Università della Svizzera italiana)

autorise l'impression de la présente thèse.

Neuchâtel, le 23 avril 2010

Le doyen :
F. Kessler

UNIVERSITE DE NEUCHATEL
FACULTE DES SCIENCES
Secrétariat - décanat de la faculté
Rue Emile-Argand 11 - CP 158
CH-2009 Neuchâtel

Abstract

Keywords: Stemmer, Natural Language Processing with Indo-European Languages, Search Engines with Asian Languages, Cross-Language Information Retrieval (CLIR), Automatic Translation

In this dissertation two aspects of information retrieval are elaborated. The first involves the creation and evaluation of various linguistic tools for languages less studied than English, and in our case we have chosen to work with the two Slavic languages Czech and Russian, and three languages widely spoken on the Indian subcontinent, Hindi, Marathi and Bengali. To do so we compare various indexing strategies and IR models most likely to obtain the best possible performance. The second part involves an evaluation of the effectiveness of queries written in different languages when searching collections written in either English or French. To cross the language barriers we apply publicly available machine translation services, analyze the results and then explain the poor performances obtained by the translated queries.

Acknowledgements

I would like to express my gratitude in first place to professor Jacques Savoy for his support, guidance and encouragement during my PhD thesis.

I would also like to thank Prof. Martin Braschler (Zurich University of Applied Sciences), Prof. Fabio Crestani (Univesitá della Swizzera Italiana) and Prof. Peter Kropf (University of Neuchâtel) for the time they consecrated to evaluate this PhD thesis.

My gratitude goes also to my former colleague Dr. Samir Abdou for a warm welcome and help in the beginning of my dissertation work and for all the work he has done for our group.

I would also like to thank my colleagues and friends Claire Fautsch and Heiko Sturzrehm, thank you for all nice discussions we have had and time we have spent together.

I am grateful to my husband Igor and our daughter Vasilija for their infinite love and support.

This research was supported in part by the Swiss National Science Foundation under Grant #200021-113273.

Contents

To Vasilija

Chapter 1

Introduction

Information retrieval (IR) involves the representation, storage, organization and accessing of information items (Salton, 1971), where the main goal is to find information within various document collections that are possibly relevant to a user's information needs. The information sought and the collections searched may be written in the same or in two or more different languages, and thus we make a distinction between monolingual (MLIR) and cross-language information retrieval (CLIR).

1.1 Motivation

Due to the rapid growth of the Internet within the domain of information retrieval a number of challenges have resulted. In the beginning of its development the content of the Internet has been mainly written in English. In recent years this balance has shifted a great deal toward other languages. Over last eight years (2000-2008)[1] the number Internet users in the English language has grown by 227 %, while for certain other languages this growth was much more pronounced. Use of the Russian language on the Internet during this same period for example has increased by 1225.8 %. Moreover, the number of internet pages written in these languages and the number of readers have both grown accordingly. In this context, the growing importance of languages other than English has prompted the development of tools and techniques needed to enable automated data processing in various languages.

Accessing information in languages other than English has become an important field of interest in IR. It began with the introduction of the Spanish language to the *ad-hoc* track in TREC 1995 (Harman, 2005). Since 1999 the NTCIR[2] evaluation campaigns have

[1] http://www.internetworldstats.com/
[2] http://research.nii.ac.jp/ntcir/

been held in Japan concentrated mainly on Far East languages such as Japanese, Korean and Chinese. Then the CLEF[3] evaluation campaign covered a variety of less popular European languages (from an IR point of view), and the testing of IR systems expand through introducing Russian in CLEF 2003, Portuguese in CLEF 2004, Bulgarian and Hungarian in CLEF 2005, and Czech in CLEF 2007 and lastly the Persian language in the CLEF 2008 campaign. The first FIRE[4] evaluation campaign led to further expansion of available languages by providing test collections in various widely spoken languages in the Indian sub-continent, namely Hindi, Marathi and Bengali.

The Web's multilingual environment has provided other challenges to the IR community, and one very important issue for them has been the ability to access information regardless of the language in which it is available. In fact, given that some users speak or have knowledge of several languages, and that in bilingual countries such as Canada or Finland, or in multilingual ones such as Switzerland, India and the European Union, there has been an increasingly important need for simple and effective information access regardless of a user's language. For example a lawyer working on EU law and who wishes to respond to a request written in a given language must be able to easily identify relevant legal texts written in the English, German or Czech languages. Furthermore, with expanding globalization, managers of multinational companies (IBM, Nestlé) or international organizations (WTO, UN) are sometimes faced with similar situations. Moreover, many other users may experience difficulties expressing their information needs in a foreign language even though they may have some understanding of documents written in that language (Oard & Resnik, 1999). Unfortunately, monolingual searching does not provide adequate solutions to these various situations, and thus an effective cross-language IR system would prove very helpful in overcoming many language barriers (Grefenstette, 1998). Even in situations where search results cannot be readily translated into a user's native language the user might be provided with good candidates for manual translation.

1.2 Objectives

The various trends described above increase the importance of the effective multilingual processing. In order to develop an effective multilingual IR system, a monolingual search engine capable of handling any language is required (Savoy, 2004). Thus one of the main goals of this dissertation is to provide an analysis of monolingual IR within the context of morphologically complex but less studied languages, particularly from the IR point of view. To adapt an existing system such that it provide optimal performance for a new

[3] http://clef.iei.pi.cnr.it/
[4] http://www.isical.ac.in/~fire/2008/index.html

language, a certain number of linguistic tools will be needed over and above a proper IR model, and this is especially true for those languages that are more morphologically complex (Pirkola, 2001).

In IR the application of a stemming procedure in the document indexing phase is assumed to be a good practice in order to improve the retrieval effectiveness. For some languages applying a stemmer may result in modest (if any) yet not statistically significant improvement (Dolamic & Savoy, 2009; Harman, 1991) while for others this procedure may prove to be very important in order to obtain acceptable retrieval performances (Dolamic & Savoy, 2007; Tomlinson, 2004). By creating stemming procedures for various languages and then evaluating their impact on monolingual retrieval effectiveness, our goal was to determine whether greater benefits might be achieved from word transformations, particularly for those languages having greater rather than simple morphological variations (Sproat, 1992), such as English. Our aim was also to study behavior of standard IR procedures, such as the IR models or indexing strategies on test collections written in various languages. Finally, by participating in different evaluation campaigns, our aim was to evaluate our solution by comparing our results to those presented by other participants.

The second main objective of this research is to determine the effectiveness of bilingual searches, and more precisely whether (and when) publicly available machine translation services could be used as an effective means of negotiating language barriers. Does translation quality and thus retrieval effectiveness depend on the relations and characteristics of the source and target languages? Or only on translation services? Or does the retrieval effectiveness of machine translated queries depend on the underlying IR model?

1.3 Concepts and Definitions

1.3.1 Information Retrieval (IR)

The main goal of information retrieval (IR) is to find which "documents" within a given collection would correspond to user information needs as expressed in a query. The information searched might exist in a variety of formats (text, table, picture, video, sound, music ...) and in different languages. The term "document" must be understood as a generic representing a news article, a scientific paper, a paragraph in an encyclopaedia, a bibliographic record, an image, a video sequence .... The work presented in this thesis is limited to text retrieval. For this reason the term "document" is considered as a synonym for a text document.

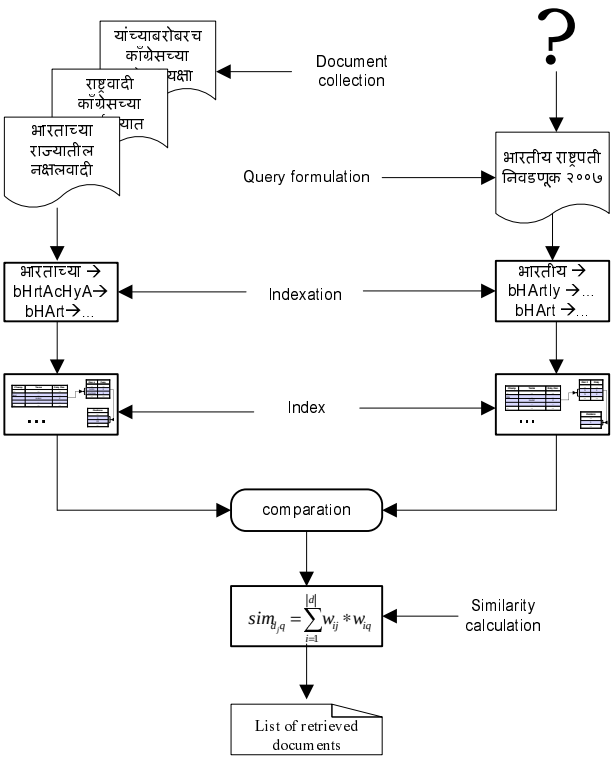

FIGURE 1.1: IR System

To perform a search in a collection for documents corresponding to a user's information need and then compare them to the query, one might use a simple sequential scan of a text file. This procedure is, however inefficient and time consuming, so to enable efficient searching the documents are indexed. Indexing represents the first phase of the IR process. At the end of this phase an index is created in the form of inverted file, containing each indexing term or feature associated to a list of documents in which it is found. An index may also contain other information like term weight, its position in the document, etc. In the second phase the user's query is submitted to the indexing process, usually the same as that used for document indexing (see Figure 1 for an illustration of the classical IR system).

The system used for experiments in this thesis included an automatic indexing process able to transform a full text document into a reduced sized index, and incorporated the following steps:

1. Transliteration - transformation of the text written in different scripts or encoding standards into the Latin (in our context).

2. Tokenization - transforms document texts into a set of tokens or terms through splitting the input text at white spaces or punctuation marks.

3. Case normalization - transforms each token into lower case.

4. Handling special words - handles special words such as acronyms and emails according to the analysing mechanism chosen.

5. Stopword removal - eliminates non content-bearing terms (e.g., "and", "the", ...), thus reduces the inverted file size. For certain languages and IR models this may have a significant influence on retrieval performance (Dolamic & Savoy, 2010).

6. Accent removal - replacing accented letter by its unaccented version improves the retrieval effectiveness for certain languages (Hollink et al., 2004).

7. Word normalization - terms are substituted with a stem or lemma, by means of various word normalization methods: morphological analysis, stemming, truncation or n-gram method (McNamee & Mayfield, 2004).

The above-mentioned steps are selected according to the language in which the document is written and the indexing strategy applied.

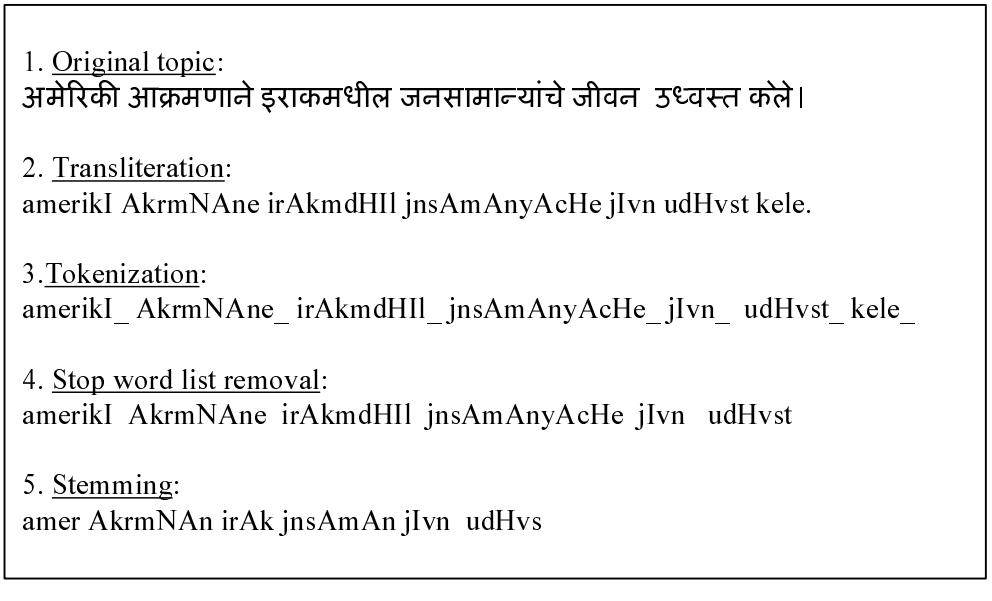

FIGURE 1.2: Indexing Process

Figure 1.2 shows an example of this algorithm's application for a topic written in the Marathi language. As illustrated, some of the above algorithm's steps were omitted. For example, in the transliteration process for this language, upper and lower case of the same Latin letter are used to represent different characters of the Devanagari script (in which the Marathi language is written). As a consequence the case normalisation step was not performed.

In the next phase, the IR system compares the query and the documents based on the terms found in the query, calculates the similarity between the two. Next, based how the query is represented, the matching process detects and classifies the documents according to their relevance, providing the user with a list of documents sorted by their decreasing order of relevance.

1.3.2 IR Models

Retrieval model must specify precisely how documents (or information items) and users information needs (requests) are represented and how these representations are used to extract and sort retrieved items. In this second part, mathematical formulae are applied to measure and calculate the similarity between query and documents surrogates. Retrieval models can be classified in several groups, while the models used in the experiments undertaken for our publications belong to two major groups: vector-space (VSM) and probabilistic. This section provides a basic description of the models used to carry out evaluations in this thesis.

1.3.2.1 Vector-Space Models

In the vector-space model (Salton, 1971), query q and the document d are represented as vectors in a multidimensional space, with every term belonging either to a document or query being a dimension in this space. In this thesis we do not make a distinction between the document and its representation as a vector of weighted terms d_j. Figure 1.3 shows an example of this notion, with query q containing three indexing terms (t_1, t_2, t_3), document d_1 containing only terms t_1 and t_3 and document d_2 containing all three indexing terms.

With each term t_i and document d_j we usually associate a weight w_{ij} reflecting the importance of the term t_i in describing the semantic content of d_j. Term weight w_{ij} can be calculated in various ways (Salton & McGill, 1984). The *tf idf* weighting scheme with weights calculated according to formula in Equation 1.1 is commonly used.

$$w_{ij} = tf_{ij} \cdot idf_i \tag{1.1}$$

In Equation 1.1 tf represents the number of occurrences of term t_i in document d_j while idf_i is the *inverse document frequency* of term t_i, generally calculated as follows:

$$idf_i = \log \frac{n}{df_i} \tag{1.2}$$

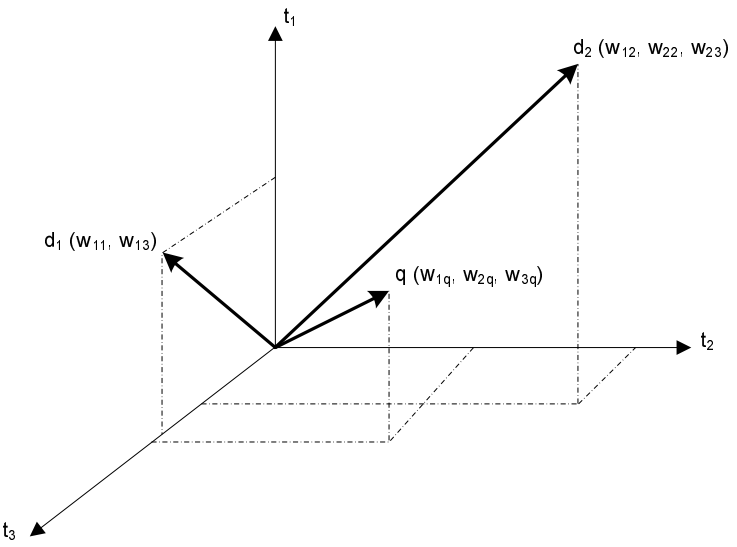

FIGURE 1.3: Vector space model

where n represents the total number of documents in the collection while df_i indicates the document frequency of the term t_i (number of documents in which the term t_i appears). More recent weighting measures also take document length into account.

Documents retrieved by the system are classified based on their degree of similarity with the query, calculated based on the term's t_i weight w_{ij} in document d_j and its weight w_{iq} in the query q. The similarity between the query and the document can be calculated as a dot product for example (Equation 1.3), or by computing cosine angle between the query and the document vector (Equation 1.4) where t indicates the number of indexing terms (or features) included in the representation space and $|q|$ the number of terms in the query. Cosine measure has the advantage of returning value between 0 (no term in common) and 1. Thus, instead of trying to predict weather or not documents are relevant to the query, it ranks them according to their degree of similarity to the query. This measure is used as a baseline measurement in the enclosed publications.

$$sim(d_j, q) = \sum_{i=1}^{|q|} w_{ij} \cdot w_{iq} \tag{1.3}$$

$$sim(d_j, q) = \frac{\sum_{i=1}^{|q|} w_{ij} \cdot w_{iq}}{\sqrt{\sum_{n=1}^{t} w_{nj}^2} \times \sqrt{\sum_{m=1}^{|q|} w_{mq}^2}} \tag{1.4}$$

When taking document length into account we could make use of more complex IR models. Term's presence in a shorter document might provide stronger evidence than in the longer document. Many vector space models have been proposed, with their

main goal being to adjust the document length normalization. Two of these models, namely "*Lnu-ltc*" (Buckley et al., 1996) or "*dtu-dtn*" (Singhal et al., 1999) are used in publications described in the Section 2.3 and Section 2.2 respectively. For the model "*Lnu-ltc*" Equation 1.5 calculates the weight assigned to the document term (*Lnu*) while Equation 1.6 determines the weight assigned to the query term (*ltc*).

$$w_{ij} = \frac{\frac{[\ln(tf_{ij})+1]}{[\ln(mean\ tf)+1]}}{(1 - slope) \cdot pivot + slope \cdot nt_i} \tag{1.5}$$

$$w_{qj} = \frac{(\ln(tf_{qj}) + 1) \cdot idf_j}{\sqrt{\sum_{k=1}^{t} ((\ln(tf_{qk}) + 1) \cdot idf_k)^2}} \tag{1.6}$$

Equation 1.7 and 1.8 give the document and query term weights respectively for the model "*dtu-dtn*".

$$w_{ij} = \frac{[\ln(\ln(tf_{ij}) + 1)] \cdot idf_j}{(1 - slope) \cdot pivot + slope \cdot nt_i} \tag{1.7}$$

$$w_{qj} = [\ln(\ln(tf_{qj}) + 1)] \cdot idf_j \tag{1.8}$$

In the above formulae nt_i represents the number of distinct indexing term in document d_i and the *pivot* and *slope* are used to adjust term weight normalization value, according to document length. This formulation prevents the retrieval system from favoring short documents over those articles longer than the mean, corresponding to the pivot value.

1.3.2.2 Probabilistic Models

The principal of the probabilistic models states that documents should be ranked according to their estimated probability of relevance to a given query (Probability Ranking Principle, (Robertson, 1977)). More precisely the goal is to determine as accurate as possible the probability of retrieved document belonging to the set of relevant documents for the given query (marked R) or the set of non-relevant documents (marked $\bar{R}$). Probabilistic models must thus estimate, as accurately as possible, the probability $P(R|d_j)$ of document d_j belonging to the set of relevant documents. In this notation, the underlying query q is implicit. The probability $P(\bar{R}|d_j)$ that the same document belongs to the set of the non-relevant documents is usually estimated as $1 - P(R|d_j)$. It could thus be said that a document is considered relevant if the probability of it being relevant is higher than the probability of it being irrelevant. The documents can thus be ranked according to their relevance odds, calculated as a ratio of the two probabilities as follows:

$$O(R|d_j) = \frac{P(R|d_j)}{P(\bar{R}|d_j)} \tag{1.9}$$

Based on the Bayesian formulation we can express:

$$P(R|d_j) = \frac{P(d_j|R) \cdot P(R)}{P(d_j|R) \cdot P(R) + P(d_j|\bar{R}) \cdot P(\bar{R})} \tag{1.10}$$

and therefore the odds ratio (of the current request q) is:

$$O(R|d_j) = \frac{P(R|d_j)}{P(\bar{R}|d_j)} = \frac{P(d_j|R) \cdot P(R)}{P(d_j|\bar{R}) \cdot P(\bar{R})} \tag{1.11}$$

In Equation 1.11 $P(d_j|R)$ is the probability of retrieving a document from the set of relevant documents for the given query. Since the values $P(R)$ and $P(\bar{R})$ are assumed to be the same for all the documents in the collection, the similarity formula can be transformed as follows:

$$O(R|d_j) \approx \frac{P(d_j|R)}{P(d_j|\bar{R})} \tag{1.12}$$

In the absence of any additional hypotheses able to simplify this process, it is still difficult to estimate $P(d_j|R)$ directly. What is needed in fact are not precise probabilities but rather a ranking according to them, thus allowing constants (such as $P(R)$) to be ignored. Second, we assume that a document's relevance is independent of that of another document. Third we assume that a document's relevance depends on its content, or more precisely on the terms occurring within it. In this case we also assume that their presence (or absence) is mutually independent. In Equation 1.12 the important estimation in order to calculate the retrieval status value (RSV) is $P(d_j|R)$ and we can estimate it as follows:

$$RSV(d_j, q) = RSV(d_j) = P(d_j|R) \approx \prod_{t_i \in d_j} P(t_i|R) \cdot \prod_{t_i \notin d_j} (1 - P(t_i|R)) \tag{1.13}$$

where $P(t_i|R)$ is the probability that term t_i occurs, given that the document belongs to the relevant set. To estimate these underlying probabilities as accurately as possible, various models have been suggested, and the three most important ones are described as follows.

Okapi The formula of the Okapi or BM25 model (Robertson et al., 2000) implemented in our system and applied in our publications is given in Equation 1.14. This model is based on the 2-Poisson model (Harter, 1975), and accounts for both term frequency and document length to determine the probability that the given document would be relevant to the query.

$$RSV(d_j, q) = \sum_{i=1}^{|q|} qtf_i \cdot idf_i \cdot \frac{tf_{ij} \cdot (k_1 + 1)}{tf_{ij} + k_1 \cdot \left[1 - b + b \cdot \frac{|d|}{avdl}\right]} \tag{1.14}$$

$$idf_i = \log \frac{N - df_i + 0.5}{df_i + 0.5} \tag{1.15}$$

where qtf_i represents the frequency of term t_i in the query q, n the number of documents in the collection, df_i the number of documents in which the term t_i appears and $|d|$ the length of the document d_j. Average document length is represented by $avdl$ while b and k_1 are constants, usually set to $b = 0.75$ and $k_1 = 1.2$.

Divergence from Randomness (DFR) The *Divergence from Randomness* paradigm, proposed by Amati & van Rijsbergen (2002) estimates the importance of a term in the retrieved document using the following two information measures:

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = -\log_2 \left[Prob^1_{ij}(tf)\right] \cdot (1 - Prob^2_{ij}) \tag{1.16}$$

$Prob^1_{ij}$ refers to the probability of having tf occurrences of the given term t_i in document d_j purely by chance, based on the chosen probabilistic model. The more the term's frequency in the document diverges from that expected by the chosen model, the lower its probability. As such the underlying term provides more information on the document's content, thus making it a good candidate for a content bearing term. On the other hand for non-content bearing terms this probability is high, given that the distribution of such terms usually follows the model chosen.

The Inf^2_{ij} (first normalization of the informative content) measures the risk of accepting the given term as a good document descriptor. In Equation 1.16 the probability $Prob^2_{ij}$ is calculated for the set of documents containing a given term t_i and represents a probability of encountering a new occurrence of the term in the given document, based on the statistics of this document set. For the models derived from this paradigm and used in our work, the exact formulae may be found in the attached publications.

$$tfn_{ij} = tf_{ij} \cdot log_2 \left[1 + \frac{c \cdot \bar{l}}{l_i}\right] \tag{1.17}$$

Before being used for calculating probabilities $Prob^1_{ij}$ and $Prob^2_{ij}$ term frequency tf_{ij} is resized according to a second normalization (see Equation 1.17) so as to account for documents length. In Equation 1.17 $\bar{l}$ is the average document length in the collection while l_i is the length of the document d_i.

Language Model (LM) Unlike other probabilistic models suggested in IR, in its attempts to estimate a document's relevance for a given query this model adapts a different point of view. By building the probabilistic language model from each document d, the documents are ranked according to the probability of generating the query using

the language model M_{d_i} built from the given document d_i. A similarity between the document d_i and query q is defined as the probability that the query q would be generated by the document's d_i language model M_{d_i}.

$$RSV(d_i, q) = P(q|M_{d_i}) \tag{1.18}$$

If a query is to be considered as a set of terms (e.g., unigram model), then RSV may be calculated according to Equation 1.19.

$$RSV(d_i, q) = P(t_1, t_2 \cdots t_n|M_{d_i}) = \prod_{j=0}^{n} P(t_j|d_i) \tag{1.19}$$

$$P(t_j|d_i) = \frac{tf(t_j|d_i)}{\sum_k tf(t_k|d_i)} \tag{1.20}$$

The simplest way to estimate the probability $P(t_j|d_i)$ is by applying the maximum likelihood estimate expressed by Equation 1.20. The problem with using this kind of estimation however is that the calculated RSV will be equal to zero for all documents missing at least one query indexing term. To avoid this problem, various smoothing techniques have been proposed (Hiemstra, 2000; Zhai & Lafferty, 2004) such that they assign a non null probability to those query terms not appearing in the document.

1.3.3 Performance Measurements

Precision and recall are two measurements used to evaluate an IR system's ability to return good responses while also discarding non-relevant ones. The precision (Equation 1.21) represents the proportion of retrieved documents that are relevant while the recall (Equation 1.22) represents the proportion of relevant documents that are retrieved. Thus to evaluate the list of results returned by the system the precision at different recall levels can be calculated, providing the precision recall curve as shown in Figure 1.4.

$$Precision = \frac{|Relevant \bigcap Retrieved|}{|Retrieved|} \tag{1.21}$$

$$Recall = \frac{|Relevant \bigcap Retrieved|}{|Relevant|} \tag{1.22}$$

To calculate the precision and recall it is necessary to know all relevant documents for a given query. This information is usually not available in the real situation. In this purpose test collections were created, providing a controlled environment containing a

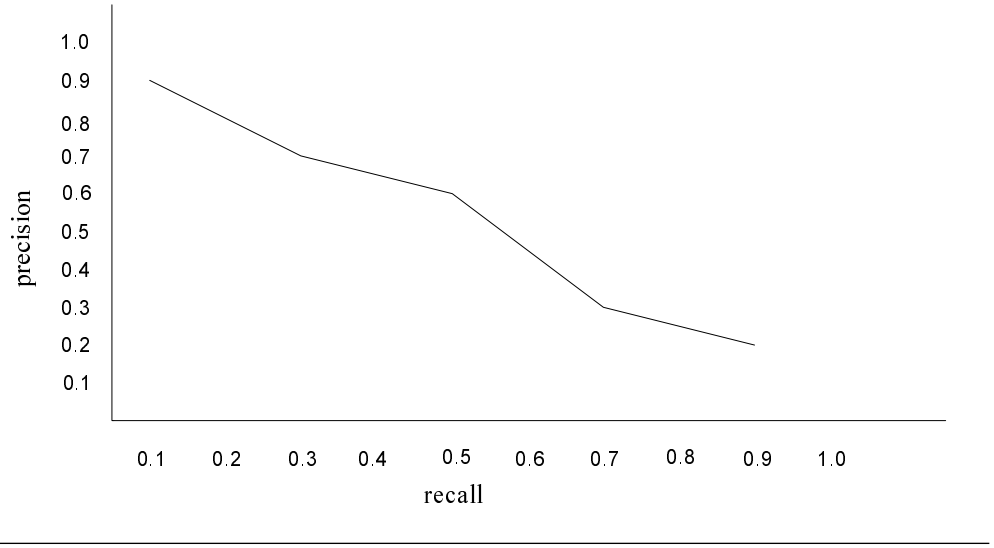

FIGURE 1.4: Precision recall curve

set of documents, a set of topics and also relevance assessments for all the topics in the set.

$$AP(q) = \frac{1}{m} \sum_{j=0}^{m} Precision(R_j) \tag{1.23}$$

In order to describe the system performance using just one value different solutions have been proposed. To combine the precision, recall and the rank of the retrieved and relevant document, we use Average Precision (AP) as given by Equation 1.23 (Buckley & Voorhees, 2005). For a given query the average precision value is determined after each relevant document is retrieved. The MAP performance measure is the mean of the AP achieved by each query in the test collection.

On the other hand not all users are interested in retrieving all relevant items but rather would like to find a single relevant item in response to their request. In this context mean reciprocal rank (MRR) may be applied. For any given query r represents the rank of the first relevant document retrieved and the query's performance is computed as $1/r$, or the reciprocal rank (RR). This value varies between 1 (the first retrieved item is relevant) and 0 (no correct response), serves as a measure of any given search engine's ability to extract one correct answer and list it among the top-ranked items. We thus believe that the MRR value should closely reflects the expectation of those Internet surfers who are looking for a single good response to their requests.

1.4 Thesis Overview

This chapter provides a brief introduction and Chapter 2 gives the short description of the various publications upon which this thesis is based on. Finally the Chapter 3 contains the conclusion of our research. Examples of the documents extracted from the different test collections used in our research are shown in Appendix B, while Appendix C contains algorithms for the light stemming procedures used.

Chapter 2

Presentation of the Publications

The content of this dissertation is based on the following five publications.

- Ljiljana Dolamic, Jacques Savoy
  "Indexing and Stemming Approaches for the Czech Language"
  In *Information Processing and Management*, 45(6), pages 714-720, 2009.

- Ljiljana Dolamic, Jacques Savoy
  "Indexing and Searching Strategies for the Russian Language"
  In *Journal of the American Society for Information Science and Technology*, 60(12), pages 2540–2547, 2009.

- Ljiljana Dolamic, Jacques Savoy
  "Comparative Study of Indexing and Search Strategies for the Hindi, Marathi and Bengali Language"
  Special Issue of ACM Transaction on Asian Language Information Processing on IR for Indian languages, to appear

- Jacques Savoy, Ljiljana Dolamic
  "How Effective is Google's Translation Service in Search?"
  In *Communications of the ACM*, 52(10), pages 139-143, 2009.

- Ljiljana Dolamic, Jacques Savoy **"Retrieval Effectiveness of Machine Translated Queries"**
  In *Journal of the American Society for Information Science*, to appear

These publications address various questions concerning motivations and objectives of this dissertation and attempts to provide answers to some of them. The following sections give a short overview of the content and contributions for each of these publications. The complete articles, containing results, discussions as well as full list of references can be found in the Appendix A.

2.1 Indexing and Stemming Approaches for the Czech Language

Czech morphology includes seven cases, three genders and two numbers for both nouns and adjectives, and thus is more complex then that of the English or German languages. Thus we may assume that stemming might have a positive effect on the IR in this language. In this article we propose two stemming approaches for the Czech language. A "light" stemming procedure removes only grammatical cases from nouns and adjectives (see the attached publication for the algorithm used). We also designed and implemented a more aggressive stemmer that, in addition to the inflectional suffixes removed by the light one, also removes a certain number of frequent derivational suffixes.

The Czech language is written in the Latin alphabet, plus eight diacritic letters used to represent specific pronunciation instances (see Document B.1 in Appendix B for an example). Even though for certain European languages it has been shown that removing diacritics improves retrieval effectiveness in mean (Hollink et al., 2004), our previous experience (Dolamic & Savoy, 2007) tends to show that for the Czech language this is not the case. For this reason, diacritics have not been removed in the experiments presented in this article.

Performed experiments allow us to conclude that aggressive stemming approach provides the best retrieval performance for this language. The average difference between the aggressive stemming approach and the approach without stemming of 46% is larger and more important than for other languages (Tomlinson, 2004). Moreover, the performances differences are always statistically significant. On the other hand, when compared to the light stemming approach or n-gram indexing, differences were rather small (i.e. 2.7% and 3.5%, respectively) and never statistically significant. From these results we thus

learned that stemming, whether light or aggressive was an effective tool when performing the search in the Czech language, and that the language independent n-gram indexing, with n fixed at 4 can be used as an effective alternative.

2.2 Indexing and Search Strategies for the Russian Language

Russian is a Slavic language written in the Cyrillic script (see Document B.2 in Appendix B for an example). As most languages belonging to this language family it has a very rich inflectional morphology (Malhebre, 1995). This is the case with verbal as well as noun or adjectival morphology (i.e., 6 cases, 3 genders, 2 numbers for nouns and adjectives), and thus in this work we attempted to determine to what extent Russian's complex morphology might influence the IR process.

To do so we proposed and evaluated two different stemmers. First, a light suffix striping algorithm consisting of 57 rules that remove only inflectional suffixes from nouns and adjectives and performs small number of normalization steps (see Algorithm C.1 in Appendix C). This algorithm will of course remove suffixes from other POS sharing those sets of suffixes used by nouns and adjectives, although in the Russian language this is rather rare. This, however, was not our main preoccupation, since our goal was to conflate different forms of nouns and adjectives into the same stem, believing that these POS convey the most meaning in documents as well as in topics. Secondly, we proposed the more aggressive suffix stripping algorithm, a stemmer applying the same set of rules as the light one adding certain number of rules to remove the most frequent derivational suffixes. In this article we also address the comparative retrieval effectiveness of the language independent n-gram scheme as well as the effectiveness of the Snowball[1] stemmer for Russian.

The experiments described in this article led us to conclude that for Russian the stemming procedure improves retrieval performance to a much larger extent than for other European languages. This may be partially explained by the fact that while in the English language a noun may be singular or plural (e.g., horse vs. horses) and different cases are expressed by means of prepositions, in Russian a noun (and even a name) may have up to twelve distinct forms, according to its number and case.

Even thought the light stemming approach proved in average to be the best performing indexing strategy for Russian, closer inspection shows that the number of topics resulting in performance increases is somewhat greater for the aggressive than for light

[1] http://snowball.tartarus.org/

stemming approach (67 vs. 61) when compared to the approach omitting the stemming procedure. For both approaches there were performance decreases for the same number of topics (18). Also, better retrieval performances resulted for those indexing strategies using our light or aggressive stemmer, when compared to either the Snowball or 4-gram indexing schemes, although these differences were never statistically significant.

2.3 Comparative Study of Indexing and Search Strategies for the Hindi, Marathi and Bengali Language

The experiments described in this publication are based on the collections provided by the first FIRE evaluation campaign for the Hindi, Marathi and Bengali languages. Hindi, Marathi and Bengali are all members of the Indo-Aryan language family and although Sanskrit is their common root language, they have certain rather distinct morphological features. While Hindi and Marathi languages are written in the Devanagari script (see Document B.3 in Hindi and B.4 in Marathi in Appendix B), Bengali has its own script (see Document B.5 in Appendix B for an example written in Bengali). In this article light and more aggressive stemming algorithms were proposed for each of these languages. The light stemmer for the Hindi language consists of only 19 rules removing case endings from nouns and adjectives as well as gender markers (see Section C.2 in Appendix C for the light stemming algorithm). For the Marathi language the light stemmer consists of 51 rules removing the case-number-gender suffixes from nouns and adjectives (see Section C.3 in Appendix C for the light stemming algorithm). As for the Bengali language the light stemmer includes 70 rules (see Section C.4 in Appendix C) dealing with case ending, article suffixes and measure words usually added to the noun stems. These differences in the number of rules for each stemmer should be expected, given the variety of inflectional morphology in the languages studied (two cases in Hindi, eight in Marathi, seven in Bengali, etc.).

We also proposed aggressive stemming algorithm for these Indic languages. Unlike those we proposed for other languages (e.g. Czech, Russian, Bulgarian, Persian, ...) the aggressive stemmers proposed for these three languages did not simply incorporate the light stemming algorithm, adding to it certain number of rules usually to remove most frequent derivational suffixes. To handle irregular inflectional morphology of these languages, the algorithm rather detects certain suffixes within words, and then removes it with part of the word following the suffix altogether.

Retrieval performances obtained using proposed stemming strategies were compared to those resulting from word-based indexing strategy, but without involving the stemming

phase. Apart from the word-based indexing strategies described for each of languages in questions, we tested the retrieval performances after applying two language-independent strategies, namely 4-gram (where the length of n-gram has been established experimentally) and trunc-4 method by simply truncating words (e.g., "computer" → "comp"). The truncation length of four letters was also established experimentally, and for all three languages this length proved to perform the best.

The experimental results presented in this publication led us to the conclusion that for these Indian languages the aggressive stemming treatment is needed in order to obtain better retrieval effectiveness. In fact aggressive stemming approach provided better retrieval performance than the light one for all three languages. The trun-4 approach (McNamee et al., 2009) proved to be the best for both Hindi and Bengali languages.

2.4 How Effective is Google's Translation Service in Search?

In this article we tried answer the question "How effective is IR in certain languages when using automatically translated queries compared to the monolingual search in the same language?". To do so we used the collection of French language news documents made available through CLEF 2001-2006 evaluation campaigns. The collection also contains 310 topics written in this language out of which 299 have relevant items in the collection. This same set of topics, when automatically translated from the German and English language by Google translations services[2] was used in order to perform a bilingual search, using title only query formulation. The results obtained by the queries originally written in English have also been compared to the retrieval performance of the topics originally written in the German language. We also used the possibility to translate directly from German to French and used these translated queries to compare the performance achieved when using English as pivot language.

The evaluation results presented in this publication tend to show that using Google translation services to automatically translate English topics into French leads to a retrieval effectiveness of around 88% of that obtained from a monolingual search. This strategy provided relevant items among first five retrieved documents for 213 topics out of 299, compared to 241 for the monolingual search. For the topics originally written in the German language the retrieval performances was somewhat lower (-30.2% when compared to the monolingual results), mostly due to the poor handling of German compound constructions.

[2] http://translate.google.com/

Finally, for the topics originally written in German translated by Google to French using the English as a pivot language, retrieval performance decreases by 20.5% compared to the monolingual search. This evaluation leads us to conclude that automatic translation is more effective for some language pairs than for others, and when compared to direct translations, the use of a pivot language does not always result in performance decreases.

To compare the effectiveness of different translation services, we had the same set of 299 topics originally written in German and English translated by other services, namely Babelfish[3] and Prompt[4]. In this case we used both the short (title only) and the medium (title, description) query formulation. The results show that the topics translated by different translation services tend to perform at the same level. More precisely, 206 (68%) topics translated by Babelfish and 212 (70.9%) translated by Prompt retrieved at least one relevant item among top five, compared to 213 for the topics translated by Google (title only query formulation). These findings were merely a first attempt, and we thus concluded that to understand the some of the difficulties occurring when translating a query from one language to another a more profound analysis was needed. This will be the subject of the next paper.

2.5 Retrieval Effectiveness of Machine Translated Queries

In the series of tests described in this publication we analyzed the effectiveness of bilingual search into English language and "informational retrieval cost" incurred by bilingual search. In other words user searches within a collection of documents written in English by writing queries in their own language. Questions on retrieval effectiveness were answered through analyzing the effects of query source language, morphological differences with the English language and the quality of translation services involved.

The English document collection was made available through CLEF 2001-2006 evaluation campaigns, together with the set of topics written in this language (used in a baseline monolingual search), and a corresponding set of topics in the German, French, Spanish and Chinese languages.

We performed the baseline monolingual search using 284 topics (title only queries). The retrieval performances obtained by the monolingual search were then compared to the retrieval performances resulting from the search performed by the corresponding topics automatically translated from German, French, Spanish and Chinese languages by both Google translation services and Yahoo![5].

[3] http://babel.altavista.com/
[4] http://translation2.paralink.com/
[5] http://babelfish.yahoo.com

Given its morphological distance from the English language, Chinese language proved to be the most demanding for both translation services. The decrease in MAP averaged over the four models used, for topics originally written in this language when compared to monolingual searches was 18.2% for the Google translation and 45.1% for the topics translated by Yahoo!. Similar tests were also performed on the Japanese language (not reported in this publication) with a somewhat restrained set of 185 topics. The results obtained showed tendencies similar to those obtained for the Chinese language.

For the two translation services involved, the "easiest" language to be translated proved to be Spanish for Google and French for Yahoo!, thus resulting in much lower retrieval effectiveness cost. Although the differences compared to the monolingual search were statistically significant even for these languages.

In this paper one of our main conclusions is that the principal reason for such decreases in retrieval performances involves the incorrect translations of names, especially those names having specific meanings in the source language (e.g., "El Niño" in Spanish), the polysemy attached to the word in the source language, different morphological and grammatical categories and the difficulties in translating compound constructions (especially in the case where German was a source language).

Chapter 3

Conclusion

3.1 Contributions

3.1.1 Monolingual IR

The main objective of this part of my thesis was studying, those languages that are morphologically complex but less studied, particularly from the IR point of view. This leads to the following question: "How does one determine that one language is morphologically more complex than another?" Morphology is a linguistics field involving the study of word structures and their formation, and more particularly their inflectional[1] and derivational[2] characteristics. The morphological complexity of a language can be determined by examining these characteristics:

- the number of part of speech (POS) undergoing inflectional changes
- the number of cases multiplied by number of grammatical genders and grammatical numbers
- the extent to which the affixes are distinct
- the possible variations of a given stem according to the attached suffix
- the degree of stability relative to the affix order, etc.

In this work both Slavic languages (Czech and Russian) and the Indian languages (Hindi, Marathi, and Bengali) were studied. Although written in different scripts, these languages, share some basic characteristics. They all have obvious word boundaries, thus

[1] Inflection is defined as use of morphological methods to form inflectional word forms who indicate grammatical relations between words.

[2] Derivational morphology is concerned with derivation of new word forms using derivational affixes. Compounding can be seen as another method of forming new words.

implying that the words are proper indexing units. Inflection in these languages is achieved exclusively through suffixing. Derivations are usually obtained through applying suffixes, even thought there are certain number of prefixes their usage is less frequent, as for all languages belonging to the Indo-European language family.

The Czech language is written in the Latin script and is characterized by seven cases (nominative, genitive, dative, accusative, vocative, instrumental and locative), three genders (masculine, feminine and neuter) and two numbers (singular and plural). Declinable POS include nouns, together with personal names, adjectives, pronouns and numbers. In Russian, the same set of POS is declined through six cases, three genders and two numbers. Although Hindi and Marathi are written in the same Devanagari script they have very distinct characteristics. In Hindi the only fully declinable POS are nouns, comprising 2 cases only. Direct, corresponding to nominative case in the Czech and Russian, and the oblique case. The word relations in this language are expressed using a noun in its oblique form with addition of postpositions. The Hindi language has two numbers and two grammatical genders. On the other hand, in the Marathi nouns, adjectives and proper nouns are declined through eight cases, three genders and two numbers. Finally, Bengali is written in the Bengali or Bangla script. This language has no grammatical gender. Only nouns are declined through seven cases and two numbers. In this language determiners are added to the nouns before case suffixes to express definite article.

Taking into consideration previously described possibilities of measuring the language morphological complexity (e.g., POS undergoing inflectional changes and $\#\ cases \cdot \#\ genders \cdot \#\ numbers$) and morphological characteristics of the languages studied, we can say that the Russian, Czech and Marathi languages can be considered as being more complex than both the Bengali or Hindi languages.

In order to deal with inflectional suffixes, light stemming procedures have been created for previously described languages. Stemming procedures are usually created through focusing mostly on the nouns and adjectives, for in our opinion these are the POS that convey a topic's meaning. For this reason we do not account for a language's verbal morphology. The results obtained by our light stemming procedure reveal an average increase in MAP of 13% for the Bengali language, 13.9% for Marathi, 19.3% for Hindi up to 42% for Czech and 90% for Russian, compared to the results obtained with the word indexing strategy when applying no word normalization whatsoever. The extreme values obtained for Russian can however be partially attributed to the specificity of this collection (average document length of only 19 terms per document). From a statistical point of view these differences were always significant, and clearly greater and more important than those obtained for English (Fautsch & Savoy, 2009; Tomlinson, 2004).

The derivation of the new word forms in these languages is performed mostly by suffixing, and aggressive stemming procedures were created to deal with derived word forms for the languages involved. These procedures remove a certain number of the most frequent suffixes. Although a certain number of prefixes are present in these languages, their usage is not that common. Moreover, usually their usage changes the word meaning, and for these reasons aggressive stemming procedures do not remove any prefixes. When we compare the performances of the aggressive stemming to no stemming approach, we can see that the statistically significant improvement in MAP for these languages (46% for the Czech, 87.5% for the Russian, 27.6% for the Hindi, 41.6% for the Marathi and 16.9% for the Bengali language) are larger and more important than for the English language (Fautsch & Savoy, 2009).

Upon comparing the performances of light and aggressive stemmers however, only the relative 24.3% improvement in MAP for the Marathi language would be considered as statistically significant. The influence that the aggressive stemming has on the IR performance for Marathi can be partially explained by the specificity of this language's morphology and the very common use of derivations. Moreover, the number and order of the derivational suffixes is not stable (Navalkar, 2001). For the other languages studied, incorporating aggressive stemming into the indexing process results in an increase in the MAP for Czech (2.7%), Hindi (6.9%) and Bengali (3.5%) while for Russian there is a slight decrease in MAP (-1.5%), even though these differences are never statistically significant.

The obtained results and previously given description of the studied languages morphology bring us to the conclusion that the positive effect of the word normalization, like stemming is much more important for the languages considered to be morphologically more complex (Russian, Czech and Marathi) than for the languages having fewer morphological variations (Bengali and Hindi).

During indexing process, words having no precise meaning (the, a, and, etc.) and thus representing noise are removed for two main reasons. First, in order for matching between query and documents to be performed based on pertinent matches. Second, to reduce the size of the index (Manning et al., 2008). To achieve these effects for languages studied we proposed stop lists. While the removal of stop words has little impact on retrieval performances for Czech, Russian, Marathi and Bengali ($1 - 3\%$), for the Hindi language incorporating this step in the indexing process has a remarkable effect, providing a statistically significant improvement in the retrieval effectiveness. This average improvement of more than 20% in MAP results from some of this language's particular features. First, in the Hindi language grammatical relations are expressed by means of postpositions, and the number of such words in one sentence (query for

example) can be large. When present they provoke a mismatch between query and documents. Second, in this language some words we consider to be without a clear meaning (e.g. "and") may be expressed using different terms. In this case the given term may not be as frequent as expected, and this feature interferes with a common assumption that the non pertinent terms appear in large number of documents.

Given our goal of ensuring useful conclusions for the various test collections and the different languages in which they were written, we considered evaluating retrieval performances from a broader perspective and under various conditions. For this reason we evaluated different indexing strategies with respect to various IR models, ranging from the classical *tf idf* to more complex probabilistic models. We can also conclude that the models derived from *Divergence from Randomness* (DFR) paradigm usually provide better retrieval performances than the vector space or language model, for all languages studied. On the other hand from the statistical point of view performance differences between the DFR and Okapi models were usually not significant. The implemented Language Model resulted in somewhat lower performance than the Okapi and DFR models, but yet for the most part it outperforms the standard vector-space model.

Finally, by participating in different CLEF (Peters et al., 2008, 2009) and FIRE [3] evaluation campaigns we were able to compare the solutions elaborated with other approaches, by which we have fulfilled the final objective of this dissertation.

3.1.2 Cross-Language IR

To perform an effective bilingual search the language barrier needs to be crossed. For the scope of this dissertation we chose to analyze and evaluate the retrieval performances achieved by the queries automatically translated using publicly available translation services. In our opinion it is also important to understand when and why a translation fails to provide the search terms needed. Thus, we have evaluated both performances of the queries originally written in the language close to the target language (in our case French in Section 2.4 and English in Section 2.5) and the languages both visually and morphologically distinct to the target language (Section 2.5).

Our findings show that writing a query in one language and than requesting an automatic translation services to translate it before lunching the search, does indeed decrease retrieval effectiveness when compared to a monolingual search. We found that the level of this decrease depends largely on the quality of the machine translation tools and source-target language pairs involved. In the case of the Google translation service for example when English is the source and French the target language, the MAP levels

[3] http://www.isical.ac.in/~fire/2008/working_notes.html

decrease by 12% on average when compared to a monolingual search, yet when German is the as the source language this difference is much more pronounced, with decrease being up to 30%.

These tendencies were also confirmed with English as the target language, with retrieval performances decreasing by approximately 7% for Spanish as a source language (this language is considered as being "close" to the English language) to more than 18% in the case of Chinese as a source, this language being morphologically and visually distant from English. For Yahoo! the differences between manually translated queries and automatically translated queries are always larger than those for Google, varying from -17.5% for French as a source language to -45.1% for Chinese as a source.

Through making a query-by-query analysis we were able to shed some light on the reasons for the poor retrieval performance of machine translated queries. These reasons may be separated into a few distinct groups. First, the presence of proper names in a query formulation has two effects: 1) a name spelled differently in the language involved is left unchanged by translation services (e.g. Solénitsyne, Solzhenitsin, Solschenizysn) and 2) a name has a specific meaning in the source language and thus is translated by the machine translation service (e.g., "El Niño‘" in Spanish). We also found that the polysemy[4] and synonyms (e.g., car vs. automobile, film vs. movie) can influence retrieval effectiveness. The important influence on the retrieval performance had choice of different grammatical constructions (e.g., gold vs. golden, European vs. Europe) as well as compound constructions. The influence of the compound constructions was important especially in the case of German as a source language.

Finally, we can conclude that even though machine translated queries do in fact result in a decrease in retrieval performances, they can be used as a means of negotiating the language barriers. However, the retrieval performance of queries handled in this way depends largely on source-target language pair.

3.2 Future Work

In this dissertation we show that the choice of the appropriate linguistic tools depends greatly on the underlining language. For this reason and for the reason of the rapid growth of the number of Internet pages available in those languages which are less studied, we believe that extending this research to other languages is important.

[4] The ambiguity of an individual word or phrase that can be used (in different contexts) to express two or more meanings.

At the same time this research was mostly based on the $ad-hoc$ collections. If we take the English language as an example, incorporating stemming into the indexing procedure improves the retrieval performance for $ad-hoc$ retrieval (Fautsch & Savoy, 2009), while on the other hand using the same stemming strategies hurts the performances on the blog collection (Fautsch & Savoy, 2008). These results raise the following question: "Would the same tendencies be true for other languages and other collection types?."

Appendix A

Selected Publications

Indexing and Stemming Approaches for the Czech Language

Ljiljana Dolamic, Jacques Savoy

Computer Science Department
University of Neuchatel
2009 Neuchâtel, Switzerland
{Ljiljana.Dolamic, Jacques.Savoy}@unine.ch

Abstract. This paper describes and evaluates various stemming and indexing strategies for the Czech language. Based on Czech test-collection, we have designed and evaluated two stemming approaches, a light and a more aggressive one. We have compared them with a no stemming scheme as well as a language-independent approach (*n*-gram). To evaluate the suggested solutions we used various IR models, including Okapi, *Divergence from Randomness* (DFR), a statistical language model (LM) as well as the classical *tf idf* vector-space approach. We found that the *Divergence from Randomness* paradigm tend to propose better retrieval effectiveness than the Okapi, LM or *tf idf* models, the performance differences were however statistically significant only with the last two IR approaches. Ignoring the stemming reduces generally the MAP by more than 40%, and these differences are always significant. Finally, if our more aggressive stemmer tends to show the best performance, the differences in performance with a light stemmer are not statistically significant. .

Keywords. Czech Language; Stemming, Evaluation, Slavic languages.

1 Introduction

Slavic languages dominate in Eastern and Central Europe (e.g., Serbo-Croatian, Russian, Polish, Bulgarian or Czech), and their distinct linguistics features (e.g., the use the various grammatical cases marked by suffixes) must be taken into account in an efficient IR system (Manning *et al*., 2008). However, the IR community has only a very small number of test-collections available for this family of languages. As an exception, we could mention the Bulgarian language for which the last two CLEF evaluation campaigns have produced a test-collection (Peters *et al*., 2008). Unlike the morphology of other Slavic languages however, the grammatical cases are usually not explicitly indicated by a given suffix in the Bulgarian morphology (with the exception of the infrequent vocative case). Thus, experiments drawn for this language cannot be applied directly to other Slavic languages.

The CLEF 2007 campaign (Dolamic & Savoy, 2008) produces also a shorter test-collection for the Czech language, and the main objective of this paper is to describe the main morphological difficulties when working with this language. We also proposed and evaluated a suitable stemmer for this Slavic language. In IR it is assumed that applying a stemmer will conflate several word variants into the same stem, and thus improve the pertinent matching between query and document surrogates. For example, when a query contains the word "horse," it seems reasonable to also retrieve documents containing the related word "horses." Moreover, stemming procedures will also reduce the size of inverted files.

When designing a stemmer, we may create a "light" suffix-stripping procedure by removing only the morphological inflections by conflating the singular and plural word forms (e.g., "door" and "doors") or feminine and masculine variants (e.g., "actress" and "actor") to the same stem. More sophisticated approaches will remove derivational suffixes (e.g., "enhance" and "enhancement") use to generate a new part-of-speech word from a given stem. Even though a different stemming procedures have been suggested for various European languages (e.g., Snowball project, CLEF, TREC and NTCIR campaigns (Peters *et al*., 2008; Harman, 2005), no stemming algorithm with its evaluation is available for the Czech language.

The rest of this paper is organized as follows. Section 2 describes different stemming approaches while Section 3 depicts the main characteristics of our test-collection. Section 4 briefly describes the IR applied during our experiments. Section 5 evaluates the performance of various IR models, in addition to two stemming approaches for the Czech language. The main findings of this paper are presented in the conclusion.

2 Related Work

In the IR domain we usually assume that stemming is an effective means of enhancing retrieval efficiency by conflating several different word variants into a common form. Most stemming approaches achieve this through applying morphological rules for the language involved (e.g., see (Lovins, 1968) and (Porter, 1980) for the English language). In such cases suffix removal is also controlled through the adjunct of quantitative restrictions (e.g., '-ing' would be removed if the resulting stem had more than three letters as in "running", but not in "king") or qualitative restrictions (e.g., '-ize' would be removed if the resulting stem did not end with 'e' as in "seize"). Certain *ad hoc* spelling correction rules are applied to improve conflation accuracy (e.g., "running" gives "run" and not "runn"), due to certain irregular grammar rules, usually applied to facilitate easier pronunciation. However, applying an algorithmic stemmer does not guarantee that we always obtain either the correct stem or an existing word in the corresponding language.

Compared to other languages having more complex morphologies (Sproat, 1992), English is considered quite simple and the use of a dictionary to correct stemming procedures could be more helpful for those other languages such as French (Savoy, 1993). When a language has an even more complex morphology, deeper analysis could be required (e.g., for Finnish (Korenius *et al*., 2004), where lexical stemmers

are clearly more elaborate and not always freely available (e.g., Xelda system at Xerox). They are more labor intensive and their implementation is complex. Moreover their use depends on a large lexicon and a complete grammar for the language involved. These application also requires more processing time and could thus be problematic, especially when document collections are very large and dynamic (e.g., within a commercial search engine on the Web). Additionally, lexical stemmers must be capable of handling unknown words such as geographical names, products, proper names or acronyms (out-of-vocabulary problems). Lexical stemmers thus cannot be viewed as error-free approaches. Finally, it must be recognized that when inspecting language usage and real corpora, the observed morphological variations are less extreme than those that might be imagined when inspecting the grammar. Kettunen & Airo (2006) indicate for example that in theory Finnish nouns have around 2,000 different forms, yet in actual collections the occurrence of most of these forms is rare. As a matter of fact in Finnish, 84% to 88% of the occurrences of inflected nouns are generated by only six out of a possible 14 cases.

While stemming schemes are normally designed to work with general texts, some may also be especially designed for a specific domain (e.g., in medicine) or a given document collection, such as that developed by Xu & Croft (1998), which used a corpus-based approach. This more closely reflects language usage (including word frequencies and other co-occurrence statistics), instead of a set of morphological rules in which the frequency of each rule (and therefore its underlying importance) is not precisely known.

Few stemming procedures[1] have been suggested for European languages other than English. The proposed stemmers usually pertain to the most popular languages (Peters *et al*., 2008; Tomlinson, 2004) and some of them, like the Finnish language, seem to require a deeper morphological analysis (Korenius *et al*., 2004) to achieve good retrieval performance.

Algorithmic stemmer ignores word meanings and tends to make errors, usually due to over-stemming (e.g., "organization" is reduced to "organ") or to under-stemming (e.g., "European" and "Europe" do not conflate to the same root). Most of the studies so far have been involved in evaluating IR performance for the English language, while studies on the stemmer performance for less popular languages are less frequent. For example, Tomlinson (2004) evaluated the differences between Porter's stemmer strategy (Porter, 1980) and lexical stemmers (based on a dictionary of the corresponding language) for various European languages. For the Finnish and German languages, lexical stemmer tends to produce statistically better results, while for seven other languages performance differences were insignificant.

Based on these facts, the rest of this paper will address the following questions: 1) Does stemming affect IR performance for the Czech language (and to which extent)? 2) For this language, is a light stemming approach more effective than more complex suffix-stripping algorithms?

[1] Freely available at the Web site http://snowball.tartarus.org/ or http://www.unine.ch/info/clef/

3 Czech Morphology and Stemming Strategies

When creating stemming procedures for the Czech language we adopted the same strategy as for the other European languages for which we have created stemmers during the past years. We believe that effective stemming should focus mainly on nouns and adjectives (sustaining most of the meaning of a document), thus ignoring numerous verb forms (tending to generate more stemming errors when taken into account).

The Czech language belongs to the Slavic languages and is written, as for example the Polish language with our Latin alphabet with the addition of eight diacritics used to specify a particular pronunciation (e.g., 'č', 'ň', 'ř', 'd', 't'). As with the Latin or the German languages, the Czech and usually other Slavic languages use various grammatical cases marked by suffixes (e.g., the noun "city" in Russian could be written as "город" (nominative), "города" (genitive) or "городе" (locative)). These linguistic elements indicate that Czech inflections are more complex than the English ones which are mainly limited to the final '-s'[2].

All nouns in the Czech language belong to the three distinct genders (masculine, feminine, or neutral). Moreover, all nouns are declined both in number (singular, plural)[3]; and using seven grammatical cases (nominative, genitive, dative, accusative, vocative, locative, and instrumental), with very few exceptions (a handful of indeclinable borrowed words). Each combination gender-case has its own set of characteristic paradigms, including hard-stem types, soft-stem types, and special types. For example, masculine noun "muž" (husband) appears as such in the nominative case singular, but varies in other cases "muže" (genitive, accusative), "mužovi"(dative, locative), "muže" (accusative), "muži" (vocative) or "mužem" (instrumental) with plural forms of this noun being "mužové," mužů," "mužům," "muže," "mužích," and "muži". From this example, we can see that the suffix denoting a case could be ambiguous in the sense that the same suffix may appear in other cases ("muže" could be the accusative or genitive singular form). Moreover, the stem (e.g., "muž" in our case) does not change after adding the appropriate suffix (unlike other languages like Finnish (Korenius *et al*., 2004)). Although this phenomenon can also occur in the Czech language, it is less frequent that in other languages. Finally, it is important to know that suffixes denoting cases occur also with proper names (e.g., with Paris, "Paříž" (nominative), "Paříže"(genitive), "Paříži"(dative), or with Ann, "Anna" (nominative), "Anny" (genitive), and "Anně" (dative)). It is also important to notice that the stemming unlike lemmatization doesn't not always produce result with a correct lexical meaning (e.g., neuter noun "moře" (sea) and its different forms "moři" (dative), "mořem"(instrumental) conflate into "moř", the corresponding stem that does not appear as it in the dictionary).

As with many languages, the suffixes assigned to adjectives agree with the attached noun in case, gender and number. These language characteristics result in large num-

[2] As for other natural languages, the English knows exceptions such as "mouse" and "mice" or the "'s" in "Paul's book" to denote the genitive case in some circumstances

[3] As for other natural languages, some words occur only in singular or plural form (e.g., "nůžky," scissors).

ber of suffixes being added to adjectives compared to other languages like German (having a rather limited set of suffixes (e.g., '-en', '-es')). Our stemmer denoted as "light" contains 52 rules for removing these grammatical case endings from nouns and adjectives (inflectional suffixes only). A complete description of this stemmer is given in the Appendix. In the case of part of speech other than nouns and adjectives sharing the same set of suffixes (this being rather rare in the Czech language), they will also be removed. In such cases, the suggested strategy will certainly produce an incorrect stem. However defining the POS of each surface word is the first step of a lexical stemmer. Their use depends also on a large lexicon and a complete grammar for the language involved. These application also requires more processing time and could thus be problematic, especially when document collections are very large and dynamic (e.g., within a commercial search engine on the Web). Additionally, lexical stemmers must be capable of handling unknown words such as geographical names, products, proper names or acronyms (out-of-vocabulary problems). On the other hand, light algorithmic stemmers have shown to be effective for different European languages (Savoy, 2006).

Derivational Czech morphology is accomplished by means of prefixation and suffixation of a stem, a usual construction with the Indo-European languages. Usually, the part-of-speech of the stem changes after adding a suffix (e.g., '-ial' in "commerce" and "commercial"). In our work we addressed only suffixes because adding a prefix usually changes more the original meaning of a stem (e.g., "prehistory" vs. "historic" from the stem "history"). In the Czech language, derivational suffixes are added before case endings. We designed and implemented a more aggressive stemmer denoted "aggressive" in this paper which, besides removing inflectional suffixes, removes certain frequent derivational suffixes as for example (e.g., "klavír" (piano) → "klavírista" (pianist)). Both suggested stemmers address other morphological characteristics of the Czech language as fleeting 'e' (e.g. "zámek" (lock, nominative sing.) → "zámku" (genitive, dative, vocative, and locative sing.)) or consonant alternations (e.g. "ruka" (hand, nominative sing.) → "ruce" (dative and locative sing.)). Such irregularities, also present in the English language, are usually integrated to smooth the pronunciation.

Finally, to define pertinent matches between search keywords and documents, we removed very frequently occurring terms having no important significance (e.g., the, in, but, some). For the Czech language, the suggested stopword list contains 467 forms (determinants, prepositions, conjunctions, pronouns, and some very frequent verb forms). In the process generating this stopword list we have followed the guidelines suggested by (Fox, 1990). Both stemmers and the suggested stopword list for the Czech language are freely available at `www.unine.ch/info/clef/`.

4 Test-Collections

The evaluations reported in this paper were based on the Czech collection built during the CLEF 2007 evaluation campaign. This corpus consists of newspaper articles extracted from the *Mladá fronta Dnes* (year 2002) and *Lidové Noviny* (year 2002) news-

papers. A typical document begins with a short title (tag <TITLE>), usually followed by the first paragraph under the <HEADINGS> tag, and finally the body (<TEXT> tag). As shown in Table 1, the mean number of indexing terms per article is around 212.6 while the whole corpus contains 81,735 articles.

Size	# docs	# docs mean terms	# queries	# rel. docs /query
178 MB	81,735	212.6	50	15.24

Table 1. Some statistics from the Czech test-collection

The topics available covered various subjects (e.g., "NATO Summit Security," "Human cloning," "VIP Divorces") including both regional ("Kostelic Olympic Medals") and more international coverage ("Causes of Air Pollution"). Topics #411 ("Best Picture Oscar") or #413 ("Reducing Diabetes Risk") owns the smallest number of pertinent articles (2) while Topic #415 ("Drug Abuse") has the greatest number of correct answers (47).

Based on the TREC model, each topic was structured into three logical sections comprising a brief title (examples given upper), a one-sentence description, and a narrative part specifying the relevance assessment criteria. In our experiments, we used only the title part of the topic formulation in order to reflect more closely queries sent to commercial search engines. Using only the title section, queries had a mean size of 2.98 search terms.

Finally, since the title part of the request "Cosmetic procedures" was corrupted in the original topic formulation (replaced by the narrative part of the previous topic) we changed this topic title part into "kosmetický procedury" (the Czech translation of the corresponding English version).

5 IR Models

To evaluate our proposed two stemming approaches with respect to various IR models, first we used the classical *tf idf* model wherein the weight attached to each indexing term was the product of its term occurrence frequency (tf_{ij} for indexing term t_j in document d_i) and the logarithm of its inverse document frequency (idf_j). To measure similarities between documents and the request, we computed the inner product after normalizing (cosine) the indexing weights (Manning *et al*., 2008).

To complement this vector-space model, we have implemented probabilistic models, such as the Okapi (or BM25) approach (Robertson *et al*., 2000), and one model derived from *Divergence from Randomness* (DFR) paradigm (Amati & van Rijsbergen, 2002) wherein two information measures formulated below are combined:

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = -\log_2[Prob^1_{ij}] \cdot (1 - Prob^2_{ij}) \quad (1)$$

in which $Prob^1_{ij}$ is the pure chance probability of finding tf_{ij} occurrences of the term t_j in a document. On the other hand, $Prob^2_{ij}$ is the probability of encountering a new occurrence of term t_j in the document, given tf_{ij} occurrences of this term had already been found.

To model these two probabilities, we used the I(n_e)C2 model based on the following estimates:

$$\text{Prob}^1_{ij} = \left(\frac{n_e + 0.5}{n+1}\right)^{tfn_{ij}}$$

$$\text{and Prob}^2_{ij} = 1 - \left(\frac{tc_j + 1}{df_j \cdot (tfn_{ij} + 1)}\right) \quad (2)$$

$$\text{with } tfn_{ij} = tf_{ij} \cdot \ln\left(1 + \frac{c \cdot \text{mean dl}}{l_i}\right) \quad \text{and } n_e = n \cdot \left(1 - \left(\frac{n-1}{n}\right)^{tc_j}\right)$$

where tc_j is the number of occurrences of term t_j in the collection, df_j indicates the number of documents in with the term t_j occurs, n the number of documents in the corpus, l_i the length of document d_i, *mean dl* (= 212), the average document length, and c a constant (fixed empirically at 1.5).

Finally, we also used an approach based on a language model (LM) (Hiemstra, 2000), known as a non-parametric probabilistic model. Various implementations and smoothing methods might also be considered within this language model paradigm. In this paper we adopted a model proposed by Hiemstra (2000; 2002) as described in Equation 3 using the Jelinek-Mercer smoothing (Zhai & Lafferty, 2004), a combination of an estimate based on document ($P[t_j \mid d_i]$) and one based on the whole corpus ($P[t_j \mid C]$).

$$\text{Prob}[q_i \mid q] = \text{Prob}[d_i] \cdot \prod_{tj \in Q} [\lambda_j \cdot \text{Prob}[t_j \mid d_i] + (1-\lambda_j) \cdot \text{Prob}[t_j \mid C]]$$

$$\text{with Prob}[t_j \mid d_i] = \left(\frac{tf_{ij}}{l_i}\right)$$

$$\text{and Prob}[t_j \mid C] = \left(\frac{df_j}{lc}\right) \quad \text{with } lc = \sum_{k=1}^{t} df_k \quad (3)$$

where λ_j is a smoothing factor (fixed at 0.35 for all indexing terms t_j), df_j indicates the number of documents indexed with the term t_j, and lc is a constant related to the size of the underlying corpus C.

6 Evaluation

In order to measure retrieval performance, we have adopted the mean average precision (MAP) computed by TREC_EVAL (Buckley & Voorhees, 2005) based on maximum of 1,000 retrieved items. To statistically determine whether or not a given search strategy is statistically better than another, we have applied the bootstrap methodology (Savoy, 1997), with the null hypothesis H_0 stating that both retrieval schemes produce similar performance. In the experiments presented in this paper statistically significant differences were detected by a two-sided test (significance level α=5%). Such a null hypothesis would be accepted if two retrieval schemes returned statistically similar means, otherwise it would be rejected.

6.1 IR Models Evaluation

Given the methodology previously described, Table 2 depicts the MAP using three stemming approaches with four IR models. In the last column we have also included a language-independent indexing approach based on 4-gram (McNamee & Mayfield, (2004). Under this indexing scheme, words are decomposed by overlapping sequences of 4 letters (this value of 4 was selected because it produced the best IR performance). For example, the sequence "prime minister" generates the following 4-grams {"prim," "rime," "mini," "inis," … and "ster"}.

	none	light	aggressive	4-gram
tf idf	0.1357*	0.2040*	0.2095*	0.1918*
Okapi	0.2040*	0.2990	0.3065	0.2957*
DFR-I(n_e)C2	**0.2208**	**0.3042**	**0.3135**	**0.3125**
LM	0.2054*	0.2813*	0.2882*	0.2785*

Table 2. Mean average precision (MAP) of various IR models and different stemmers

Finally, we have compared the retrieval effectiveness of the IR model with and without the stopword list. The performance differences were small (in mean, around 1%) and did not give any evidence of significant impact of stopword list removal on MAP, for this language at least. Of course, the inverted file was reduced as well as the query processing time.

6.2 Stemming Evaluation

Facing a language with more complex inflectional morphology than English, we may infer that applying stemming will improve the MAP. However to which extent (if it really exists) is not, a priori, known. This section will address these questions using different IR models.

If we use retrieval performance without stemming, marked "none" in Table 2 as a baseline, we can see that both stemming strategies, "light" and "aggressive", performed better than the baseline. Applying our statistical testing, we found that all performance differences were always statistically significant when compared to an approach ignoring the stemming stage. If we average the performance over four models given, we find an increase of 42% with the "light" stemmer and 46% with the more "aggressive" one. These relative improvements are clearly large and more important than with other languages (Tomlinson, 2004) (+4% with the English language, +4.1% Dutch, +7% Spanish, +9% French, +15% Italian, +19% German, +29% Swedish, +40% Finnish).

When comparing different stemming strategies we can see that the "aggressive" stemmer performs slightly better, 2.7% in average over four models. The retrieval performance differences were in this case never statistically significant.

Denoted as "4-gram" in Table 2 are shown retrieval performances of the given IR models when language independent 4-gram indexing strategy (without applying a

stemming procedure). The performance difference between 4-gram indexing strategy and word-based indexing is rather small (e.g., in average -1% over "light" and -3.5% over "aggressive") and is never statistically significant.

When analyzing query-by-query the effect of applying a stemmer, and limiting our investigations of the best performing model (DFR-I(n_e)C2), we found that after applying our light or more aggressive stemmer, the performance was increased for 41 queries while, for the remaining 9 queries, the average precision (AP) decreases. In this case, Topic #418 ("Bülent Ecevit's Statements") has the greatest improvement, starting with an AP of 0.25 without stemming to 0.6797 (+172%) with our light stemmer and 0.7526 (+201%) with the more aggressive approach. Explanation for this improvement could be found in the fact that personal names in Czech, as in other Slavic languages are changed through cases. Genitive form of the name found in this query ("Prohlašení Bülenta Ecavita") as well as other forms found in relevant documents, after stemming conflate to its nominative form enabling a pertinent matching. Also, Topic #441 ("Space tourists") cannot retrieve any relevant articles without stemming (AP 0.0), retrieves the first relevant document in second place with both stemmers (e.g., AP 0.3568 with light stemmer). None of the terms forming the query ("Vesmírní turisté"), exists in relevant documents in the same word form (they occur as "vesmírný", "vesmírnou", "turista"). Of course, applying a stemmer may sometimes hurt the AP as shown by Topic #407 ("Australian Prime Minister") having an AP of 0.9325 without stemming to 0.5616 (-39.8%) with our light stemmer and 0.5925 (-36.5%) with the more aggressive approach. In this case nouns "premiér" (prime minister) and "premiéra" (first night, premiere) conflate to the same stem resulting in retrieving large number of non-relevant articles.

Finally it is interesting to know that some topics could be classify as hard because for all indexing strategies and IR models they achieve a MAP smaller than 0.1. In our experiments, we have found seven such topics (#403, #411, #422, #425, #428, #436, #439). Those topics mostly contain either too general terms (e.g., Topic #436 "VIP divorces") or certain spelling errors (e.g., in Topic #411 "Best Picture Oscar", Academy Award's name was spelled with a K ("Oskar") in the topic and with a C ("Oscar") in relevant documents).

7 Conclusion

In this paper, we present the main aspects of the Czech morphology and we suggested two stemmers for this Slavic language, one removing only inflectional suffixes (denoted "light") and a second algorithm that removes also some frequent derivational suffixes (denoted "aggressive"). Both approaches contain some rules to correct orthographic irregularities. A stopword list containing 467 forms was also suggested. These linguistic tools are freely available on the Internet.

Using the most effective current IR models, we have evaluated our stemming approaches and found that the best performing IR model is derived from *Divergence from Randomness* (DFR) paradigm. This approach performs statistically better than a

language model or the classical *tf idf* while the difference with the Okapi model was not statistically significant.

Our various experiments clearly show that a stemming procedure improves retrieval effectiveness when applied to the Czech language (mean improvement of around +45%, larger than those found for other European languages). From a statistical point of view, the differences are always significant when comparing to an approach ignoring stemming.

From comparing different stemming strategies, it seems that the more aggressive stemming approach produces better MAP than does a light stemmer, but the difference between these two stemming schemes is never statistically significant.

Acknowledgments
This research was supported in part by the Swiss NSF under Grant #200021-113273.

References

Amati, G., & van Rijsbergen, C.J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM-Transactions on Information Systems*, 20(4), 357-389.

Buckley, C., & Voorhees, E.M. (2005). Retrieval system evaluation. In E.M. Voorhees, D.K. Harman (Eds): *TREC. Experiment and Evaluation in Information Retrieval* (pp. 53-75). The MIT Press, Cambridge (MA).

Dolamic, L., & Savoy, J. (2008). Stemming approaches for east European languages. In C. Peters, V., Jijkoun, T., Mandl, H., Müller, D.W., Oard, A., Peñas, D. Santos, D. (Eds.): *Advances in Multilingual and Multimodal Information Retrieval* (pp. 37-44). LNCS #5152, Springer-Verlag, Berlin.

Fox, C. (1990). A stop list for general text. ACM-SIGIR Forum, 24, 19-35.

Harman, D.K. (2005). Beyond English. In *TREC Experiment and Evaluation in Information Retrieval*, E.M. Voorhees, D.K. Harman (Eds), The MIT Press, Cambridge (MA), 153-182, 2005.

Hiemstra, D. (2000). Using language models for information retrieval. CTIT Ph.D. Thesis.

Hiemstra, D. (2002). Term-specific smoothing for the language modeling approach to information retrieval. The importance of a query term. In *Proceedings of the ACM-SIGIR*, Tempere, 35-41.

Kettunen, K. & Airo, E. (2006). Is a morphologically complex language really that complex in full-text retrieval? In *Advances in Natural Language Processing* (pp. 411-422). LNCS #4139, Berlin: Springer.

Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2004). Stemming and lemmatization in the clustering of Finnish text documents. In *Proceedings of the ACM-CIKM*. The ACM Press, Washington (DC), 625-633.

Lovins, J.B. (1968). Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 11(1), 22-31

Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge (UK).

McNamee, P., & Mayfield, J. (2004). Character *n*-gram tokenization for European language text retrieval. *IR Journal*, 7(1-2), 73-97.

Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A. & Santos, D. (Eds.). (2008). *Advances in Multilingual and Multimodal Information Retrieval*. LNCS #5152, Springer-Verlag, Berlin.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130-137.

Robertson, S.E., Walker, S., & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. *Information Processing &Management*, 36(1), 95-108.

Savoy, J. (1993). Stemming of French words based on grammatical category. *Journal of the American Society for Information Science*, 44(1), 1-9.

Savoy, J. (1997). Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management*, 33(4), 495-512.

Savoy, J. (2006). Light stemming approaches for the French, Portuguese, German and Hungarian languages. In *Proceedings ACM-SAC*, The ACM Press, Dijon, 1031-1035.

Sproat, R. (1992). Morphology and computation. Cambridge: The MIT Press.

Tomlinson, S. (2004). Lexical and algorithmic stemming compared for 9 European languages with Hummingbird SearchServer™ at CLEF 2003. In *Comparative Evaluation of Multilingual Information Access Systems* (pp. 286-300). LNCS #3237, Springer-Verlag, Berlin.

Xu, J., & Croft, B. (1998). Corpus-based stemming using cooccurrence of word variants. *ACM-Transactions on Information Systems*, 16(1), 61-81.

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2), 179-214.

Appendix: Description of our Czech Light Stemmer

```
CzechStemmer(word) {
  RemoveCase(word);
  RemovePossessives(word);
  Normalize(word);
  return;
}

RemoveCase(word) {
  if (word ends with “-atech”) then remove “-atech” return;
  if (word ends with “-ětem”) then remove “-ětem” return;
  if (word ends with “-etem”) then remove “-etem” return;
  if (word ends with “-atům”) then remove “-atům” return;
  if (word ends with “-ech”) then remove “-ech” return;
  if (word ends with “-ich”) then remove “-ich” return;
  if (word ends with “-ích”) then remove “-ích” return;
  if (word ends with “-ého”) then remove “-ého” return;
  if (word ends with “-ěmi”) then remove “-ěmi” return;
  if (word ends with “-emi”) then remove “-emi” return;
  if (word ends with “-ému”) then remove “-ému” return;
  if (word ends with “-ěte”) then remove “-ěte” return;
  if (word ends with “-ete”) then remove “-ete” return;
  if (word ends with “-ěti”) then remove “-ěti” return;
  if (word ends with “-eti”) then remove “-eti” return;
  if (word ends with “-ího”) then remove “-ího” return;
  if (word ends with “-iho”) then remove “-iho” return ;
  if (word ends with “-ími”) then remove “-ími” return;
  if (word ends with “-ímu”) then remove “-ímu” return;
```

```
  if (word ends with “-imu”) then remove “-imu” return;
  if (word ends with “-ách”) then remove “-ách” return;
  if (word ends with “-ata”) then remove “-ata” return;
  if (word ends with “-aty”) then remove “-aty” return;
  if (word ends with “-ých”) then remove “-ých” return;
  if (word ends with “-ama”) then remove “-ama” return;
  if (word ends with “-ami”) then remove “-ami” return;
  if (word ends with “-ové”) then remove “-ové” return;
  if (word ends with “-ovi”) then remove “-ovi” return;
  if (word ends with “-ými”) then remove “-ými” return;
  if (word ends with “-em”) then remove “-em” return;
  if (word ends with “-es”) then remove “-es” return;
  if (word ends with “-ém”) then remove “-ém” return;
  if (word ends with “-ím”) then remove “-ím” return;
  if (word ends with “-ům”) then remove “-ům” return;
  if (word ends with “-at”) then remove “-at” return;
  if (word ends with “-ám”) then remove “-ám” return;
  if (word ends with “-os”) then remove “-os” return;
  if (word ends with “-us”) then remove “-us” return;
  if (word ends with “-ým”) then remove “-ým” return;
  if (word ends with “-mi”) then remove “-mi” return;
  if (word ends with “-ou”) then remove “-ou” return;
  if (word ends with “-[aeiouyáéíýě]”) then remove “-[aeiouyáéíýě]” return;
  return;
}

RemovePossessives(word) {
  if (word ends with “-ov”) then remove “-ov” return;
  if (word ends with “-in”) then remove “-in” return;
  if (word ends with “-ův”) then remove “-ův” return;
  return;
}

Normalize(word) {
  if (word ends with “čt”) then replace by  “ck” return;
  if (word ends with “št”) then replace by  “sk” return;
  if (word ends with “c” or “č”) then replace by  “k” return;
  if (word ends with “z” or “ž”) then replace by  “h” return;
  if (word ends with “e*”) then replace by  “*” return;
  if (word ends with “*ů*”) then replace by “*o*” return;
  return;
}
```

Figure A1. Our Czech light stemmer

Indexing and Searching Strategies for the Russian Language

Ljiljana Dolamic, Jacques Savoy

Computer Science Dept., University of Neuchatel,
Rue Emile Argand 11, 2009 Neuchâtel, Switzerland
{Ljiljana Dolamic, Jacques.Savoy}@unine.ch

Abstract

This paper describes and evaluates various stemming and indexing strategies for the Russian language. We design and evaluate two stemming approaches, a light and a more aggressive one, and compare these stemmers to the Snowball stemmer, to no stemming and also to a language-independent approach (*n*-gram). To evaluate the suggested stemming strategies we apply various probabilistic IR models, including the Okapi, the *Divergence from Randomness* (DFR), a statistical language model (LM), as well as two vector-space approaches, namely the classical *tf idf* scheme and the *dtu-dtn* model. We find that the vector-space *dtu-dtn* and the *Divergence from Randomness* models tend to result in better retrieval effectiveness than the Okapi, LM or *tf idf* models, while only the latter two IR approaches result in statistically significant performance differences. Ignoring stemming generally reduces the MAP by more than 50%, and these differences are always significant. When applying an *n*-gram approach, performance differences are usually lower than an approach involving stemming. Finally, our light stemmer tends to perform best, although performance differences between the light, aggressive and Snowball stemmers are not statistically significant.

1 Introduction

Russian belongs to the Indo-European language family and it is the most widely spoken among Slavic languages. Russian is one of three contemporary East Slavic languages, (the others being Ukrainian and Belorussian). With 165 million native speakers and 110 million second-language speakers, Russian is among the world's top 10 most spoken languages (Malherbe, 1995), and in Central and Eastern Europe it ranks at the very top. Even though in this region Slavic languages dominate, only a rather small number of document collections are available. For the Bulgarian (a South Slavic language), a fairly large collection was created during the 2006 (Peters *et al*., 2007) and 2007 CLEF campaigns (Peters *et al*., 2008), while for the Czech language (West Slavic), a test-collection was created during the CLEF-2007 campaign (Peters *et al*., 2008).

In this paper the main objective is to describe the most significant morphological difficulties encountered when applying IR techniques to the Russian language. Unlike English where few inflectional suffixes are used to denote number or person variations, Russian makes use of a larger number of them, partly because they are also used to denote grammatical cases (Sproat, 1992). Given the importance of this Slavic language, our goal is to propose, compare and evaluate various stemming, indexing and search strategies. Our evaluation will be based on the document collections made available through the 2005 to 2008 CLEF domain-specific tasks. In this case, the main objective is to study information retrieval on domain-specific corpus using both full-text and manual indexing as well the possible usefulness of specialized thesaurus for improving the retrieval effectiveness.

2 Related Work

In the IR domain (Manning *et al*. 2008), it is usually assumed that stemming is an effective mean of enhancing retrieval efficiency by conflating several different word variants into a common form. Most stemming approaches achieve this through applying morphological rules for the language involved (for English see (Lovins, 1968) and (Porter, 1980)). In such cases suffix removal is also controlled through the adjunct of quantitative restrictions (e.g., '-ing' would be removed if the resulting stem consisted of more than three letters as in "running", but not in "king") or qualitative restrictions (e.g., '-ize' would be removed if the resulting stem did not end with 'e' as in "seize"). To improve conflation accuracy, certain *ad hoc* spelling correction rules are also applied (e.g., "running" becomes "run" and not "runn"), due to certain irregular grammar rules, usually applied to facilitate pronunciation.

Compared to other languages having more complex morphologies (Sproat, 1992), English is considered quite simple, while for other languages such as French simply applying a dictionary to correct stemming procedures could be more helpful (Savoy, 1993). For those languages having a more complex morphology, deeper analyses could be required (e.g., for Finnish (Korenius *et al.*, 2004)), and their corresponding lexical stemmers would clearly be more elaborate but they are not always freely available (e.g., Xelda system at Xerox). Not only would their implementation be more labor intensive and complex, their use would depend on a large lexicon and a complete set of grammar rules for each language involved. This could lead to more processing time and would thus be problematic, especially when document collections are very large and dynamic (e.g., within a commercial search engine on the Web). Additionally, lexical stemmers must be capable of handling unknown words such as geographical, product or proper names, or acronyms (out-of-vocabulary problem). Lexical stemmers thus cannot be viewed as error-free approaches. It must also be recognized that when inspecting language usage and real corpora, the morphological variations observed are less extreme than those involved in grammar. According to Kettunen & Airo (2006) for example, while in theory Finnish nouns have around 2,000 different forms, in current collections most of these forms rarely occur. In fact 84% to 88% of inflected noun occurrences in Finish are generated by only six out of a possible 14 cases.

While stemming schemes are normally designed to work with general texts, some could also be designed especially for a specific domain (e.g., in medicine) or a given document collection, such as that developed for a corpus-based approach by Xu & Croft (1998). This would more closely reflect language usage (including word frequencies and other co-occurrence statistics) than a set of morphological rules where the frequency of each rule (and therefore its underlying importance) is not precisely known.

Other than English, few stemming procedures have been suggested for European languages (some of them are freely available at snowball.tartarus.org/ or at the web site www.unine.ch/info/clef/). These proposed stemmers usually pertain to the most popular languages and some of them, like the Finnish language, seem to require a deeper morphological analysis (Korenius *et al*., 2004) to provide adequate retrieval performances.

Algorithmic stemmers ignore word meanings and also tend to make errors due to over-stemming (e.g., "organization" is reduced to "organ") or to under-stemming (e.g., "European" and "Europe" do not conflate to the same root). Most studies carried out so far involved IR performance evaluations for the English language, while for the less popular languages, fewer studies are available. For various European languages, Tomlinson (2004) for example evaluated the differences between Porter's stemmer (1980) and lexical stemmers (based on a dictionary of the corresponding language). For Finnish and German, the lexical stemmers tend to produce statistically better results, yet for seven other languages the performance differences were insignificant.

Finally we could also mention the ROMIP evaluation campaigns producing test-collections mainly extracted from the Web in the Russian language. However, it was not possible to obtain freely this test-collection, and all pertinent information about these corpora, evaluation methodology and linguistic tools are written in Russian. After analyzing the more recent results, we found that the retrieval performance of the Snowball stemmer tends to reflect the best practice in this field.

Based on these facts, in the rest of this paper we analyze stemmer effectiveness for Russian, and suggest which one would be the most effective. We also address the comparative retrieval effectiveness of an *n*-gram scheme, a language-independent approach and compared them to a word-based scheme.

3 Morphology of the Russian Language

When creating stemmers for Russian we started from the same point as for the other languages we have worked with over the past years. We found that the best way to develop effective stemming procedures was to focus mainly on nouns and adjectives (Savoy, 2006), and to avoid verb forms, which are usually too numerous and can lead to a large number of errors.

Russian is a member of the Slavic language family and like many in this family including Bulgarian, Ukrainian or Serbian, it is written in the Cyrillic script and uses 33 letters. Other Slavic languages from areas in which the Roman Catholicism is the dominant religion, such as Polish and Czech are written with the Latin alphabet, with various diacritics being added to represent their particular pronunciations.

All Russian nouns have one of three distinct genders (masculine, feminine, or neutral). As in English, all nouns are declined according to number (singular, plural) but some may only have singular or plural forms, as in the English word "scissors". Like most other Slavic languages (except for Bulgarian), all Russian nouns (common or proper nouns) are also declined according to different grammatical cases and we can find six cases in Russian including nominative, genitive, dative, accusative, instrumental and locative. Each gender-case combination has its own set of characteristic paradigms, including hard-stem types, soft-stem types and special types. Note however that each gender-case combination does not require a distinct suffix. In the first declination for example, the accusative and the genitive have the same ending, or as shown in Table 1 dative and locative case endings are the same.

Suffixes are not always present and in some cases there are none at all. For example, the feminine noun "book" is written as "книг" in genitive plural and takes the form "книга" in the nominative singular form. The stem is therefore not always the nominative singular (for other examples see Table 1, showing the declension of feminine nouns ending in '–a' in the nominative singular). Usually the stem does not change after adding the required suffix (see Table 1 for a few examples). However, as with other Slavic languages, the presence of a suffix may imply a stem modification, as for example, in the elision of the vowel 'o' in the neutral noun "window," which takes the form "окно" in the nominative singular (instead of "оконо," (an incorrect form in Russian) with the final 'o' being a suffix) and "окон" in the genitive plural. This phenomenon is known as the fleeting vowel. Another example we should mention is the vowel 'e' in the noun "father," which takes the form "отец" in the nominative singular and "отцу" in the dative singular (or the noun "ice," taking the forms "лёд" (nominative singular) and "льду" (dative singular)). Table 1 shows another example of the feminine noun "sister," written as "сестра" in the nominative singular and as "сестёр" in the genitive plural. Variations in stem spelling are however not as important as in other languages, such as Finnish. Finally, to complete our description, we should mention that a limited number of nouns, mainly those borrowed from other languages, are not declined.

Case	Moscow	Sister	
		Singular	Plural
Nominative	Москва	сестра	сёстры
Genitive	Москвы	сестры	сестёр
Dative	Москве	сестре	сёстрам
Accusative	Москву	сестру	сёстры
Instrumental	Москвой	сестрой	сёстрами
Locative	Москве	сестре	сёстрах

Table 1. Examples of Russian feminine noun declensions

Inflectional suffixes may also be attached to particles, numerals and adjectives. According to Russian grammar rules, adjectives agree in gender, number and case with the noun they modify. The adjective forms may be one of two major types: long adjectives, inflected for case, gender and number (e.g., as in "John put the red hat"), and the short form, existing only in the nominative predicate form (e.g., "the hat is

red") and inflect only for gender and number. In Russian, indeclinable forms include adverbs, prepositions, conjunctions, plus a limited number of borrowed substantives.

In our experiments we make use of "light" stemmers which apply 57 rules in order to remove only the inflectional suffixes from nouns and adjectives (to normalize the resulting stems, we added 4 more rules).

Suffixes may also be used to derive new words from a stem, usually by changing the word's part-of-speech (e.g., "care" and "careful" or "carefulness"). Primarily, Russian derivations are formed through the use of prefixes and suffixes (e.g., "спутник" (spoutnik) = "с" (prefix) + "пут" (stem, path) + "ник" (suffix)). Forming these words is not always simple, especially without modifying the base form, as in "admit" and "admittance". Just as with English words, Russian consonants and vowels may be shifted, mutated or dropped. The root serves as the derivation's base and center, and it may or may not occur without the use of word-formative components. In developing aggressive stemmers we concentrated primarily on removing adjectival qualitative and relational suffixes (e.g., "кровь" (blood) and "кровавый" (bloody)). We thus completely ignored any prefixes we thought might change the base word's meaning (e.g., "prehistory" and "history"). Their removal may end up with a base form having unrelated or no meaning, thus diminishing retrieval performance (e.g., "закат" means "sunset" but it could be erroneously interpreted as "за" + "кат" where "за" means "after, behind", and "кат" means "kath" (*catha edulis*), bushman's tea). To develop our light stemmer and to remove certain derivational suffixes, 40 rules were added to the light stemmer version.

Compound word construction (e.g., handgun, viewfinder) is another morphological characteristic that might impact retrieval effectiveness. Most European languages use some form of compound construction, indicated either by a hyphen (e.g, in French "porte-clefs" (key ring)) or by a suffix attached to the genitive case (e.g., in German with the "-s" suffix in "Produktionsmethode" = "Produktion" + "-s" + "Methode"). In general however no particular "glue" is used to build a compound from two or more words, as in English ("viewpoint") or German ("Bankgesellschaft"). Compound constructions are also possible in Finnish, such as "rakkauskirje" = "rakkaus" (love) and "kirje" (letter). In Russian also, frequently encountered word forms include "радиоприёмник" (radio-receiver) = "радио" (radio) + "приёмник" (receiver), or "микроволновой" (adjective) = "микро" (micro) + "волновой" (wave) (with "волновой" = "волна" (stem, noun wave) + "ов" (suffix used to form an adjective from a noun) + "ой" (inflection denoting the masculine, nominative, singular case).

In our efforts to improve pertinent matches between topics and documents written in Russian, we have also created a stopword list, which includes 412 most commonly used terms such as pronouns (e.g., "мы" (we)), prepositions (e.g, "в" (in), "на" (on)), conjunctions (e.g., "и" (and), "или" (or)), or other forms (e.g., "да" (yes), "буду" (will)), etc. Both our stemmers and stopword lists are freely available (http://www.unine.ch/info/clef/).

4 Test-Collection

The Russian test-collection used in our experiments was built during the domain-specific tracks at CLEF 2005 to 2008. The main objective of this track is to evaluate the relative performance of various retrieval models for structured scientific bibliographic collections written in English, German and Russian language. In this case, documents contain textual elements (title, abstracts) as well as subject keywords from controlled vocabularies. The main focus is on the evaluation of IR models with short description of information items on the one hand, and on the other the leveraging of controlled vocabularies and other structured metadata entities to hopefully improve monolingual and bilingual information retrieval.

From this test-suite, we have extracted the Russian test-collection consisting of the Russian Social science corpus (RSSC) comprising 94,581 documents, and the INION corpus covering Russian social science and economics bibliographic data (145,802 articles). Document length in each corpus is rather short, being 19 and 15 distinct indexing terms respectively. Some statistics about this test-collection are given in Table 2.

Typical documents from each collection are listed in Figure 1 (RSSC corpus) and Figure 2 (INION corpus). To build document representatives during the indexing process, we retained pertinent sections only. These included the <TITLE> and <TEXT> for the RSSC collection and the <TITLE-RU>, <KEYWORD-RU> (terms extracted manually from INION Thesaurus) and <ABSTRACT-RU> segments (available for around 27% of the documents) in the INION corpus. In our experiments we ignored additional information such as author name (<AUTHOR> or <AUTHOR-RU>) or classification tags (e.g., <HANDLE>).

	2005	2006	2007	2008
Source	RSSC	RSSC INION	INION	INION
Size	64.6 MB	145.5 MB	80.9 MB	80.9 MB
Number of documents	94,581	240,383	145,802	145,802
Number of topics	25	25	25	25
Topics	#126 - #150	#151 - #175	#176 - #200	#201 - #225

Table 2. Test-collections statistics (CLEF)

<DOCNO> RSSC-SOCIONET-RU-20050228-001018 </DOCNO>
<HANDLE> RePEc:rus:cemicf:704 </HANDLE>
<AUTHOR> Зеликина Л.Ф.; Зеликин М.И. </AUTHOR>
<TITLE>
Многофакторные модели экономического роста переходного периода структуры магистральных многообразий. </TITLE>
<CLASSIFICATION>
экономика </CLASSIFICATION>
<TEXT>
Тез. IV Международного семинара "Комплекс ные исследования перехода России и других стран к устойчивому развитию с использованием математического моделирования" - Москва, Ин-т социально-политических исследований РАН, сентябрь 1998. </TEXT>

Figure 1. Example of document from RSSC collection

Created for domain-specific tasks during CLEF campaigns held during the years 2005-2008, the test-collection contains 100 topics. The relevance judgments were made by human assessors, and for 6 topics no relevant document can be found leaving 94 topics for the evaluation. These topics covered various subjects (e.g., “Health risks at work,” “Doping and sports,” “Value change in Eastern Europe”), including both regional (“The German school system”) and international topics (“Poverty”).

Based on the TREC model, each topic was divided into three logical sections. First we can find a brief title (under the tag <EN-TITLE> in Figure 3) follows by a one-sentence description (e.g., <EN-DESC> in Figure 3) and a narrative part specifying the relevance assessment criteria (e.g., <EN-NARR> in Figure 3). Full examples written in the Russian and English languages are depicted in Figure 3. In order to more closely reflect queries sent to commercial search engines in our experiments we used only the title part of the topic formulations. When using only the title section, our queries had a mean size of 3.25 search terms.

<DOCNO> ISISS-RAS-ECOSOC-20060324-45953 </DOCNO>
<AUTHOR-RU> Орлов, Г.М.; Кондратенко, А.И. </AUTHOR-RU>
<TITLE-RU>
Социальное партнерство или усиление экономической зависимости редакционных коллективов </TITLE-RU>
<KEYWORDS-RU>
пресса; социальное партнерство; Россия </KEYWORDS-RU>
<ABSTRACT-RU>
По данным анализа деятельности редакционного коллектива газеты "Орловская правда". </ABSTRACT-RU>

Figure 2. Example of article extracted from INION corpus

<TOP>
<NUM> 160 </NUM>
<EN-TITLE> Precarious working conditions </EN-TITLE>
<EN-DESC> Research papers and publications on types of work that deviate from normal working conditions </EN-DESC>
<EN-NARR> What "atypical" types of work conditions have developed? What "precarious" consequences are there for effected workers? What improvements to healthcare, social security and unemployment insurance status are being discussed? Are there factors that may halt this development? </EN-NARR> </TOP>
…
<TOP>
<NUM> 160 </NUM>
<RU-TITLE> Опасные условия труда </RU-TITLE>
<RU-DESC> Найти научные статьи и публикации в прессе о видах работ, при которых условия труда отличаются от нормальных. </RU-DESC>
<RU-NARR> Какие существуют «нетипичные» виды условий труда? Каковы могут быть опасные последствия для работников? Какие обсуждаются возможные улучшения в здравоохранении, социальном обеспечении и страховании от безработицы? Существуют ли факторы, способные остановить развитие ситуации? </RU-NARR>
</TOP>

Figure 3. Example of topic description in English and Russian languages

5 IR Models

In order to obtain a broader perspective on the relative merit of the various retrieval models and stemming approaches, we applied two vector-space schemes and three probabilistic models. First we adopted the classical *tf idf* model, wherein the weight attached to each indexing term was the product of its term occurrence frequency (or tf_{ij} for indexing term t_j in document d_i) and its inverse document frequency (or idf_j). To measure similarities between documents and requests, after normalizing (cosine) the indexing weights we computed the inner product (for more information, see Chapter 6 in (Manning *et al*., 2008)).

For the vector-space model better weighting schemes have been suggested, especially in cases where the occurrence of a term in a document is viewed as a rare event. Thus, a good practice may be to give more importance to the first occurrence of a term, as compared to its successive and repeating occurrences, with the the *tf* component being computed as the $\ln(tf) + 1$ or as $\ln(\ln(tf)+1)+1$. A term's presence in a shorter document might also provide stronger evidence than it would in a longer document. In order to take document length into account, we could make use of more complex IR models, including the "*dtu-dtn*" IR model suggested by Singhal *et al*., (1999). In this case Equation 1 calculates the indexing weight assigned to document term (*dtu*) and Equation 2 the indexing weight assigned to query term (*dtn*).

$$w_{ij} = [[\ \ln(\ln(tf_{ij})+1) + 1] \cdot idf_j] \,/\, [(1\text{-}slope) \cdot pivot + (slope \cdot nt_i)] \quad (1)$$

$$w_{qj} = [[\ \ln(\ln(tf_{qj})+1) + 1] \cdot idf_j] \quad (2)$$

where nt_i is the number of distinct indexing term in document d_i and *pivot* and *slope* are used for adjusting term weight normalization value according to document length.

This formulation prevents the retrieval system from overfavoring short documents compared to articles longer than the mean corresponding to the *pivot* value. For all our experiments, the constant *slope* was fixed at 0.25 and *pivot* at 15 corresponding to the average document length.

In addition to these two vector-space schemes, we also considered Okapi probabilistic models (Robertson *et al*., 2000), as well as two models derived from the *Divergence from Randomness* (DFR) paradigm (Amati & van Rijsbergen, 2002) wherein the two information measures formulated below were combined:

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = -\log_2[Prob^1_{ij}] \cdot (1–Prob^2_{ij}) \quad (3)$$

where $Prob^1_{ij}$ is the pure chance probability of finding tf_{ij} occurrences of the term t_j in document d_i. On the other hand, $Prob^2_{ij}$ is the probability of encountering a new occurrence of term t_j in the document, given that tf_{ij} occurrences of this term had already been found. To estimate these probabilities, we might instead use the DFR-GL2 model based on the following formulae:

$$Prob^1_{ij} = [1/(1+\lambda_j)] \cdot [\lambda_j /(1+\lambda_j)]^{tf_{ij}} \quad \text{with } \lambda_j = tc_j/n \quad (4)$$

$$Prob^2_{ij} = tfn_{ij}/(tfn_{ij} + 1) \quad \text{with } tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((c \cdot mean\ dl)/l_i)] \quad (5)$$

where tc_j is the number of occurrences of term t_j in the collection, n the number of documents in the corpus, l_i the length of document d_i, *mean dl* (fixed at 15) the average document length, and c a constant (fixed empirically at 1.5).

In our second DFR model, DFR-I(n_e)B2, Equation 6 is used to calculate Inf^1_{ij}, and Equation 7 to calculate $Prob^2_{ij}$, as shown below:

$$Inf^1_{ij} = tfn_{ij} \cdot \log_2[(n+1) / (n_e+0{,}5)] \quad \text{with } n_e = n \cdot [1 – [(n-1)/n]^{tc_j}] \quad (6)$$
$$\text{and } tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((c \cdot mean\ dl)/l_i)]$$

$$Prob^2_{ij} = 1 - [(tc_j +1) / (df_j \cdot (tfn_{ij}+1))] \quad (7)$$

Finally, we also considered an approach based on a language model (LM) (Hiemstra, 2000), known as a non-parametric probabilistic model. Probability estimates would not be based on any known distribution (as in Equation 4), but rather be estimated directly and based on occurrence frequencies in document d_i or the entire C corpus. Within this language model paradigm, various implementations and smoothing methods (Zhai & Lafferty, 2004) might also be considered, and in this study we adopted a model proposed by Hiemstra (2000) as described in Equation 8, which combines an estimate based on document ($P[t_j | d_i]$) and corpus ($P[t_j | C]$).

$$P[d_i | q] = P[d_i] \cdot \prod_{t_j \in Q} [\lambda_j \cdot P[t_j | d_i] + (1-\lambda_j) \cdot P[t_j | C]]$$
$$\text{with } P[t_j | d_i] = tf_{ij}/l_i \text{ and } P[t_j | C] = df_j/lc \quad \text{with } lc = \sum_k df_k \quad (8)$$

where λ_j is a smoothing factor (fixed at 0.25 for all indexing terms t_j), df_j indicates the number of documents indexed with the terms t_j, and lc are constants related to the underlying corpus C.

In Equation 8, $P[d_i]$ is the previously calculated probability that the document d_i is pertinent. We ignored this value in our experiments because it did not vary across the documents and thus did not change the final ranking. For web searches however this probability may vary across different web pages, depending, on the number of incoming links, page lengths or other factors such as page popularity measures within the web site (Kraaij *et al*. 2002).

6 Evaluation

To evaluate the retrieval performance of the various IR schemes, we used the mean average precision (MAP) a performance measure that has been used by all evaluation campaigns for more than 15 years in order to objectively compare various IR strategies, particularly regarding their ability to retrieve relevant items (ad hoc tasks) (Buckley & Voorhees, 2005). The MAP value does not have a direct interpretation for the final user. It is computed as the mean of the precision scores obtained after each relevant document is retrieved, using zero as the precision for relevant documents that are not retrieved (Buckley & Voorhees, 2000). The MAP values were computed by TREC_EVAL software, based on a maximum of 1,000 retrieved records. By using a mean to measure performance we give equal importance to all queries. We also combined the topic descriptions from the 2005 to 2008 CLEF evaluation campaigns in order to base our results on relatively large number of topics (94 in this case), believing that it is important to perform experiments involving the largest possible number of observations.

In order to statistically determine whether one strategy was better than another, we used the two sided *t*-test (Buckley & Voorhees, 2005), with the null hypothesis H_0 stating that both retrieval strategies produce a similar MAP. This null hypothesis is accepted if two retrieval schemes returned statistically similar MAP, otherwise it is rejected. In the experiments presented in this paper, statistically significant differences were detected by a two-sided *t*-test with a significance level of $\alpha = 5\%$.

Finally, it is also well known that the basis for comparisons between two (or more) IR strategies must be similar, using the same document collection and the same topics, as was mentioned by (Buckley & Voorhees, 2005).

> "The primary consequence of the noise is the fact that evaluation scores computed from a test collection are *relative* scores only. The only valid use for such scores is to compare them to scores computed for other runs using the exact same collection." (Buckley & Voorhees, 2005, p. 73).

Thus, it is clearly impossible to compare the performance obtained using a test collection with that achieved based on another document collection or directly performances obtained from the CLEF 2007 topics with those of CLEF 2008.

6.1 IR Models Evaluation

Table 3 depicts the MAP based on the methodology mentioned above and using four different stemming approaches and six IR models. The last column lists a 4-gram language-independent indexing approach (McNamee & Mayfield, 2004). In this indexing scheme, words are decomposed by overlapping 4 letter sequences (the value 4 was selected because it produced the best IR performance). For example, the sequence "prime minister" generates the following 4-grams {"prim," "rime," "mini," "inis," … and "ster"}.

	MAP (Mean Average Precision)				
	None	Light	Aggressive	Snowball	4-gram
tf idf	0.0739*	0.1302*	0.1328*	0.1282*	0.1381*
dtu-dtn	**0.0999**	**0.1892**	0.1749	**0.1847**	0.1708
Okapi	0.0881*	0.1734	0.1735	0.1648	0.1710
DFR-I(n_e)B2	0.0928	0.1802	**0.1812**	0.1734	**0.1741**
DFR-GL2	0.0879*	0.1708	0.1688	0.1624	0.1712
LM	0.0964*	0.1821	0.1793	0.1762	0.1613*
mean	0.0898	0.1710	0.1684	0.1650	0.1644
% change		+90.3%	+87.5%	+83.6%	+83.0%

Table 3. MAP of various stemming strategies and IR models

In Table 3, the best performance obtained for each stemming approach is shown in bold, indicating that either the vector-space model *dtu-dtn* or the probabilistic model DFR-I(n_e)B2 would always prove to be the best IR model. We then used these best performances as a baseline for statistical testing. Any performance differences that were statistically significant when compared to the best IR model are indicated with a star ("*"). We can thus see that when compared to *tf idf*, the differences were always statistically significant. For the other models, the performance differences were usually not statistically significant (except for the LM model with the 4-gram indexing or those listed in the "None" column).

In addition to the indexing strategies shown in Table 3, we also tested one where stemming was combined with decompounding procedure. Even though decompounding may be effective for some languages (e.g., German and Finnish), for Russian it resulted in lower MAP than it would have in the strategy not using decompounding (around 5% in average).

Finally, for all experiments listed in Table 3 we used our stopword list to remove very frequent and non-content bearing terms. We also compared retrieval effectiveness of different IR models with and without this list, discovering that performance differences were rather small (around 2% on average), thus showing no evidence that removing the stopword list had any important impact on the MAP.

6.2 Stemming Strategies Evaluation

As shown in Table 3, we first evaluated the retrieval performance without any stemmer, listing the MAP values in the "None" column. We then reported the retrieval performance obtained by our "Light" and "Aggressive" stemmers. In the "Snowball," column we listed the MAP obtained using the available Snowball stemmer (http://snowball.tartarus.org/) and in the last column lists we listed the results of the language independent 4-gram indexing strategies. As shown in the second to last row of Table 3, we computed the average performance achieved by each of the six

retrieval models in order to obtain an overview of the performance of each stemming approach.

As shown by the values listed in Table 3, all approaches using stemming performed much effectively than those that did not use stemming. When compared to an approach without stemming (the "None" column in Table 3), averaging the performance over six given models showed relative increases ranging from 83% with the 4-gram indexing scheme to 90.3% with our "light" stemmer (percentage depicted in the last row of Table 3). These relative improvements were clearly quite large and more significant than those found for other European languages (e.g., +4% with the English language, +4.1% Dutch, +7% Spanish, +9% French, +15% Italian, +19% German, +29% Swedish, or +40% Finnish (Tomlinson, 2004)).

After applying our statistical tests, we found that the performance differences between stemming and no stemming schemes were always statistically significant. Finally when comparing the 4-gram to the word-based indexing strategies (other than those listed under "None"), performance differences were rather small (e.g., -3.8% over the "light" stemmer), and these performance differences were never statistically significant.

To analyze the effect of applying a stemming, we performed a query-by-query analysis, concentrating only on a single retrieval model DFR-I(n_e)B2, one of the best performing models for any of the indexing strategies used. In this study we thus showed that by applying a stemmer could increase the performance for more than 60 topics (61 with "light", 67 with "aggressive") over a no stemming scheme, and in both cases it was observed that a decrease occurred in average precision (AP) for only 18 topics.

When using the light stemmer the greatest improvement was obtained by Topic #223 "Media in the preschool age," with an AP of 0.6607 compared to 0.01 without stemming. This improvement can be explained by the fact that the term "детьми" (children, instrumental) is found in the topic while terms "детей" (children, accusative or getitive) or "дети" (children, nominative) can be found in the relevant documents. These variants are conflated to the same stem with both our light or aggressive stemmers, but do not with Snowball stemmer (AP of 0.0227), nor do they yield the same 4-gram (AP: 0.0598).

We found a somewhat similar situation with Topic #160 "Precarious working conditions" when the terms "опасные" and "опасньх" were conflated to the same stem and significantly improved the performance for all stemming procedures (e.g., AP of 0.0093 with "None" vs. 0.6165 with "Light"). At times of course stemming can diminish retrieval performance, usually through conflating non-related terms into the same stem.

We also found that Topic #146 was the most difficult topic in this "Diabetes Mellitus" ("Диабет меллитус") collection. It did not retrieve any items, relevant or not, since none of the terms in the topic appeared in the collection.

7 Conclusion

In this paper, we have presented the main aspects of Russian morphology and suggested two stemmers for this Slavic language, one removing only inflectional suffixes (denoted "light") and a second removing certain frequent derivational suffixes (denoted "aggressive"). Both approaches apply a few rules to correct orthographic irregularities. We have also suggested a stopword list containing 412 word forms. These linguistic tools are freely available on the Internet (www.unine.ch/info/clef/).

To evaluate our stemming approaches, we use the most effective current IR models, finding that those IR models derived from *Divergence from Randomness* (DFR) paradigm or the vector-space model *dtu-dtn* perform best, depending of the underlying indexing and stemming strategy. Statistically speaking, these approaches perform better than the classical *tf idf* model or in some cases than a language model, while for the Okapi model there are no significant statistical differences.

When applied to the Russian language, our various experiments clearly show that a stemming procedure improves retrieval effectiveness, especially in the case of the collection containing short documents (e.g. bibliographical records, tables or pictures captions, statistical tables, etc.). From a statistical point of view, the differences are always significant when compared to an approach ignoring stemming. When comparing different stemming strategies, for most IR models we observe that even though our light stemming tends to perform better than other stemming strategies, performance differences among these different stemmers are never statistically significant. Based on our various examples, we also show that stemming can have a concrete effect on various topic formulations.

In our opinion when comparing stemming procedures, it is also important to consider the final user. A non-stemming or a light stemming approach is better understood than a more aggressive approach that might return unexpected results. For this same reason for the Russian language we suggest applying a light stemmer, only removing the plural and grammatical cases associated with nouns or adjectives.

Acknowledgments

This research was supported in part by the Swiss NSF under Grant #200021-113273.

References

Amati, G., & van Rijsbergen, C.J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM-Transactions on Information Systems*, 20, 357-389.

Buckley, C., & Voorhees, E. M. (2000). Evaluating evaluation measure stability. In *Proceedings ACM SIGIR*, (pp. 33-40). New York, NY, The ACM Press.

Buckley, C., & Voorhees, E. M. (2005). Retrieval system evaluation. In *TREC Experiment and Evaluation in Information Retrieval*, E.M. Voorhees, D.K. Harman (Eds), (pp. 53-78). Cambridge (MA): The MIT Press.

Hiemstra, D. (2000). *Using Language Models for Information Retrieval*. CTIT Ph.D. Thesis.

Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2004). Stemming and lemmatization in the clustering of Finnish text documents. In *Proceedings of the ACM-CIKM*, (pp. 625-633). New York, NY, The ACM Press.

Kraaij, W., Westerveld, T., & Hiemstra, D. (2002). The importance of prior probabilities for entry page search. In *Proceedings of the ACM-SIGIR*, (pp. 27-34). New York, NY, The ACM Press.

Lovins, J.B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11, 22-31.

Malherbe, M. (1995). *Les langues de l'humanité*. Paris: Robert Laffont.

Manning, C., Raghavan, P., Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge (UK): Cambridge University Press.

McNamee, P., & Mayfield, J. (2004). Character *n*-gram tokenization for European language text retrieval. *IR Journal*, 7, 73-97.

Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M. & Stempfhuber, M. (Ed). (2007). *Evaluation of Multilingual and Multi-modal Information Retrieval*. LNCS #4730, Berlin: Springer-Verlag.

Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A. & Santos, D. (Ed). (2008). *Advances in Multilingual and Multimodal Information Retrieval*. LNCS #5152, Berlin: Springer-Verlag.

Petras, V., Baerisch, S., & Stempfhuber, M. (2008). The domain-specific track at CLEF 2007. In C. Peters, V. Jijkoun, T. Mandl, H. Müller, D.W. Oard, A. Peñas, D. Santos (Ed). *Advances in Multilingual and Multimodal Information Retrieval*, (pp. 160-173). LNCS #5152, Berlin: Springer-Verlag.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, 14, 130-137.

Robertson, S.E., Walker, S., & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. *Information Processing & Management*, 36, 95-108.

Savoy, J. (1993). Stemming of French words based on grammatical category. *Journal of the American Society for Information Science*, 44, 1-9.

Savoy, J. (1997). Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management*, 33, 495-512.

Savoy, J. (2006). Light stemming approaches for the French, Portuguese, German and Hungarian languages. In *Proceedings ACM-SAC*, (pp. 1031-1035). New York, NY, The ACM Press.

Savoy J. (2007). Searching strategies for the Bulgarian language. *Information Retrieval*, 10, 509-529.

Singhal, A., Choi, J.,Hindle, D., Lewis, D.D., & Pereira, F. (1999). AT&T at TREC-7. In *Proceedings TREC-7*, (pp. 239-251). Gaithersburg (MA): NIST.

Sproat, R. (1992). *Morphology and Computation.* Cambridge: The MIT Press.

Tomlinson, S. (2004). Lexical and algorithmic stemming compared for 9 European languages with Hummingbird SearchServer™ at CLEF 2003. In C. Peters, J. Gonzalo, M. Braschler, M. Kluck (Ed). *Comparative Evaluation of Multilingual Information Access Systems*, (pp. 286-300). LNCS #3237, Berlin: Springer-Verlag.

Xu, J., & Croft, B. (1998). Corpus-based stemming using cooccurrence of word variants. *ACM-Transactions on Information Systems*, 16, 61-81.

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22, 179-214.

Comparative Study of Indexing and Search Strategies for the Hindi, Marathi and Bengali Languages

Ljiljana Dolamic, Jacques Savoy
Computer Science Dept., University of Neuchatel,
Rue Emile Argand 11, 2009 Neuchâtel, Switzerland
{Ljiljana.Dolamic, Jacques.Savoy} @unine.ch

Abstract

The main goal of this paper is to describe and evaluate various indexing and search strategies for the Hindi, Bengali and Marathi languages. These three languages are ranked among the world's 20 most spoken languages and they share similar syntax, morphology and writing systems. In this paper we examine these languages from an IR perspective through describing the key elements of their inflectional and derivational morphologies and on this basis suggest a light and more aggressive stemming approach.

To evaluate these stemming strategies we apply them to the FIRE 2008 test-collections. To extend our comparisons we also implement and evaluate two language independent indexing methods: the *n*-gram and trunc-*n* (truncation of the first *n* letters). We evaluate these solutions by applying our various IR models, including the Okapi, *Divergence from Randomness* (DFR) and statistical language model (LM) together with two classical vector-space approaches: *tf idf* and *Lnu-ltc*.

Experiments performed with all three languages tend to demonstrate that the I(n_e)C2 model derived from *Divergence from Randomness* paradigm results in the best mean average precision (MAP). Our tests suggest that

better retrieval effectiveness would be obtained with a more aggressive stemmer accounting for certain derivational suffixes than one involving a light stemmer or one that ignores this type of word normalization procedure. When comparing a no stemming with a stemming indexing scheme, performance differences are almost always statistically significant. When an aggressive stemmer was applied the relative improvements obtained are ~28% for the Hindi language, ~42% for Marathi and ~20% for Bengali over a no stemming approach. Based on a comparison of word-based and language independent approaches we find that the trunc-4 indexing scheme would tend to result in performance levels that are statistically similar to an aggressive stemmer, yet better than the 4-gram indexing scheme. A query-by-query analysis reveals the reasons for this, thus demonstrating the advantage of applying a stemming or a trunc-4 indexing scheme.

Categories and Subject Descriptors
H.3.1 [**Content Analysis and Indexing**]: *Indexing methods; Linguistic processing*. H.3.3 [**Information Search and Retrieval**]: *Retrieval models*. H.3.4 [**Systems and Software**]: *Performance evaluation*.

General Terms
Algorithms, Measurement, Performance.

Keywords
Indic languages, Stemmer; Natural Language Processing with Indo-European Languages, Search Engines with Asian Languages; Hindi Language, Marathi Language, Bengali Language.

1. Introduction

Over the last few years there has been an increasing interest in Asian languages, focusing mainly on those of the Far-East (e.g., the Chinese, Japanese and Korean languages) and on the Indian subcontinent. Given the increasing volume of sites available in these languages and of Internet pages

in general, plus the online users working with them, a better understanding of the automated procedures used to process them is clearly needed.

As in Europe, the Indian subcontinent can be characterized by the use of many languages. With 23 official languages being spoken in the European Union the situation there would seem to be more complex than in the Republic of India, where there are only two official languages (Hindi and English). This general view however hides the fact that around 29 languages are spoken by more than one billion native speakers there, and in various Indian states, many of these have an official status. From a linguistic perspective, the situation in India is thus slightly more complex than in Europe because the various languages are grouped into four main families: the Indo-European (more precisely the Indo-Aryan branch [Massica 1991] with for example the Bengali, Hindi, Marathi, Pandjabi) located mainly in the north, the Dravidian family (e.g., Kannada, Malayalam, Tamil, Telugu) in the southern part, the Sino-Tibetan (e.g., Bodo, Manipur) in the north-eastern part, and the Austra-Asiatic group (Santali) in the eastern part of the subcontinent. While Europe is also made up of various language families (e.g., with the Finnish and Hungarian belonging to the Finno-Ugric branch), in India the non Indo-European languages make up a larger proportion than in Europe. Compared to the two alphabets used in Europe (Latin and Cyrillic), at least seven different writing systems are used in the various Indian languages.

Based on the test-collections made available through the first Forum for Information Retrieval Evaluation (FIRE[1]) campaign, this paper focuses on three of the most popular Indian languages, namely Hindi (the native language of ~180 million speakers), Marathi (~65 million) and Bengali (~190 million). This paper also describes the main morphological variations and constructions of these languages, particularly from an IR perspective. For these three languages we propose and evaluate stopword lists and various stemmers, and then compare them by applying various indexing and search strategies.

[1] More information available in the FIRE Web site, see http://www.isical.ac.in/~clia/

The rest of this paper is organized as follows. Section 2 presents some related work on stemming. Section 3 describes the main morphological aspects of the three selected languages. Section 4 reveals various stemming approaches while Section 5 depicts the main characteristics of our test-collections. Section 6 briefly describes the different IR models used during our experiments. Section 7 evaluates and analyzes the performance of the various indexing and search strategies, and the key findings are presented in the last section.

2. Related Work

In the IR domain it is usually assumed that stemming is an effective mean of enhancing retrieval efficiency through conflating several different word variants into a common form or stem [Manning *et al*. 2008]. This efficiency is achieved through applying morphological rules specific to each language involved. For English Lovins [1968] and Porter [1980] are typical examples. This suffix removal process is also controlled through the adjunct of quantitative restrictions (e.g., '-ing' would be removed if the resulting stem had more than three letters as in "hopping", but not in "ring") or qualitative restrictions (e.g., '-ize' would be removed if the resulting stem did not end with 'e' as in "seize"). Certain *ad hoc* spelling correction rules are also applied to improve conflation accuracy (e.g., "hopping" gives "hop" and not "hopp"), through applying irregular grammar rules usually designed to facilitate pronunciation.

Simple algorithmic stemming approaches ignore word meanings and tend to make errors, often caused by over-stemming (e.g., "general" becomes "gener", and "organization" is reduced to "organ") or to under-stemming (e.g., with Porter's stemmer, the words "create" and "creation" do not conflate to the same root. This is also the case with the words "European" and "Europe"). For this reason the use of an on-line dictionary means of obtaining better conflation has been suggested [Krovetz 1993].

Compared to other languages (such as French) having more complex morphologies [Sproat 1992], English could be considered quite simple and the

use of a dictionary to correct stemming procedures would be more helpful than for other languages [Savoy 1993]. For languages with even more complex morphologies deeper analyses would however be required (e.g., for Finnish [Korenius *et al*. 2004]) and the lexical stemmers [Tomlinson 2004] are not always freely available (e.g., Xelda system at Xerox). Their design and elaboration is more labor intensive and complex. Moreover, their use would involve a large lexicon as well as a complete grammar, and thus their application would be problematic and require more processing time, especially with very large and dynamic document collections (e.g., within a commercial search engine on the Web). Additionally, lexical stemmers must be capable of handling unknown words such as geographical, product and proper names, as well as acronyms (out-of-vocabulary problem). Lexical stemmers could not therefore be as considered error-free approaches. Finally it must be recognized that based on language usage and real corpora, the morphological variations observed would be less extreme than those imaginable when inspecting grammar as well. For example in theory Finnish nouns have around 2,000 different forms, yet in actual collections most of these forms occur very rarely [Kettunen & Airo 2006]. As a matter of fact 84% to 88% of the occurrences of inflected nouns in Finnish are generated by only six of a possible 14 grammatical cases.

While as a rule stemming schemes are designed to work with general texts, some are specifically designed for a given domain (e.g., in medicine) or document collection (such as that developed by Xu & Croft [1998] for use in a corpus-based approach). This in fact more closely reflects general language usage (including word frequencies and other co-occurrence statistics), rather than a set of morphological rules in which the frequency of each rule (and thus its underlying importance) is not precisely known.

While studies of the English language have been more inventive, various algorithmic stemmers have indeed been suggested for the most popular European languages, especially in conjunction with CLEF[2] evaluation campaigns [Peters *et al*. 2008], while within previous NTCIR[3] evaluation

[2] See the Web site http://clef.iei.pi.cnr.it/

[3] For more information, see the Web site at http://research.nii.ac.jp/ntcir/

campaigns Far-East languages such as Chinese, Japanese or Korean have been evaluated. Although stemming procedures have been proposed for Hindi [Ramanathan & Rao 2003] and Bengali [Sakar & Bandyopadhyay 2008] as well as morphological analyzers[4] for Hindi and Marathi, there have been no reports of any comparative evaluations of these propositions based on test-collections.

Most evaluation studies done to date involved IR performance evaluations for the English language, while stemmer performance studies for other languages are found much less frequently. In their evaluations of retrieval performances resulting from the application of two statistical stemmers to five languages, Di Nunzio *et al*. [2004] demonstrated that wide variations across these languages could be found. Compared to statistical stemmers, Porter's stemmers seem to work slightly better, while for the German language, Braschler & Ripplinger [2004] showed that for short queries stemming may enhance mean average precision by 23%, compared to 11% for longer queries. Finally, Tomlinson [2004] evaluated the differences between Porter's stemmer and the lexical stemmer. For Finnish and German, Tomlinson [2004] found that lexical stemmer tends to produce better results statistically, while for Dutch, Russian, Spanish, French and English performance differences are small and insignificant. For Swedish, the algorithmic stemmer results in statistically superior mean average precision (MAP) when compared to a lexical stemming approach.

3. Morphology

Given that the Sanskrit language is their common root, Hindi, Marathi and Bengali are clearly related and thus their sentence structure follows the same Subject - Object - Verb (or SOV) pattern. From an IR point of view however this aspect is not of primary importance, since in this paper the IR models used are based on the bag-of-words assumption wherein the absolute and relative position of words within a sentence are ignored.

[4] http://ltrc.iiit.net/showfile.php?filename=onlineServices/morph/index.htm

Indeed in these languages a closer inspection of their lexicons reveals that words with similar meanings may have similar spellings. As examples the word "king" can be taken spelled as "राजा" in Hindi and Marathi as "রাজা" in Bengali, while for other terms spellings may be similar for only two of these languages, and for some words the spelling is completely different in all three languages (e.g., "God" is written as "ईश्वर" in Marathi, "ख़ुदा" in Hindi and "ঈশ্বর" Bengali). As with other languages, lexicons in these languages are never free from the influence of others, and the same is true in the other direction. English borrows some words from the Indian languages, such as "jungle" (from a Sanskrit stem), "punch" (drink, from Hindi or Marathi), "jute" (vegetable fiber, from Bengali) or "curry" (from the Tamil). Similarly and to a larger extent, many word forms in Indian languages are borrowed from English, especially given its dominant presence in commerce (e.g., taxi, company, bank, budget, ice cream, gasoline) as well as in technology (e.g., computer, internet).

In their written forms, Hindi and Marathi employ the Devanagari script while the Bengali alphabet belongs to the Brahma family. The two scripts are however clearly related and share certain characteristics. All vowels except the short 'a' (written as 'अ' in Devanagari and 'অ' in Bengali) have two forms: first as an initial or syllable ('-आ', '-আ') and the second as a medial or final vowel (e.g., 'क' + 'आ' = "का" in Devanagari, 'ক' + 'আ' = "কা" in Bengali). Consonants appearing together in special clusters form conjunct letters (ligature) such as ('क' + 'क' = "क्क", 'क' + 'स' = "क्स" (Devanagari) and 'ক' + 'ক' = "ক্ক", 'ক' + 'ম' = "ক্ম" (Bengali)).

In the rest of this section we describe the key morphological characteristics of these three languages that how they impact on IR design and performance.

3.1 Key Features of Hindi Morphology

Hindi is spoken by about 500 million people and ranks second among the world's most spoken languages (Chinese is the first while English and Hindi have the same ranking). The term "Hindi language" however does

not refer to a well-defined and clearly standardized language but rather to a relatively large group of dialects wherein inter-lingual understanding is always possible (just as English in the UK and the US).

Written using the Devanagari script, Hindi contains eleven vowels and 33 simple consonants, and along with nasal symbols such as anusvar (‘ं’) and anurasik (‘ँ’), and a symbol for weak aspiration visgar (‘ः’) (although very rare in this language). Generally no distinction is made between uppercase and lowercase letters.

In Hindi grammar [Kellogg 1938] there are only two genders, masculine and feminine, while the neuter found in Sanskrit has disappeared. Feminine nouns are usually formed from the masculine, either by replacing the final ‘-आ’(‘ा’) by ‘-ई’(‘ी’) (e.g., “घोड़ा” (horse), “घोड़ी” (mare)) or by adding ‘-ई’ for nouns ending with a consonant (“बंदर” (monkey), “बंदरी” (female monkey)). As for number, a distinction is made between singular and plural.

This language makes no use of a definite article (the), and instead placing prepositions before the noun, it positions them after (e.g., “on the table” → “table on”) in the form of postpositions. These are used in certain Western European languages such as German in the expression “den Fluss *entlang*” (“*along* the river”)), while in other European languages of Europe the use of this linguistic construction is clearly the exception.

Nouns and adjectives may have also two distinct grammatical cases, direct and oblique. The direct case normally case indicates the subject of a verb, while the oblique case might be combined with postpositions to form other object or adverbials complements (e.g., “John gives a *bone* to *Fido* in the *garden*”).

Number and case are expressed by inflectional suffixes and in part by adding certain particles to the stem or base form. To obtain the oblique singular form, most masculine nouns ending in ‘-आ’ (written as ‘ा’ in medial or final form) inflect their final vowel to ‘-ए’ (‘े’), and those in ‘-आं’ to ‘-एं’ or into ‘-ए’. All such nouns inflected in the oblique singular retain the same form in the nominative plural, while for all other masculine nouns the nominative singular and plural have the same form.

As an example, the masculine noun "horse" is written as "घोड़ा" in the direct singular while its oblique singular is "घोड़े" and as a rule it is used in conjunction with postpositions to designate other complements, as in "घोड़े को" (dative singular). As for plural forms, the direct case is written as "घोड़े" or the oblique case as "घोड़ों".

Hindi adjectives may be either inflected or uninflected. Uninflected adjectives remain unchanged before all nouns and under all circumstances, the same as with English adjectives (e.g., "सुंदर" (beautiful)). All inflected adjectives usually end in '-आ' (e.g., "काला" (black)) and their inflection depends on the gender and case of the noun they alter (e.g., as for masculine noun "काला घोड़ा" (black horse), "काले घोड़े" (black horses) or with feminine noun in "काली बिल्ली" (black cat), "काली बिल्लिज़ां" (black cats)) [Kellogg 1938].

Derivational morphology in Hindi takes place through adding a suffix to the stem, and typically the stem's part-of-speech (POS) changes once the suffix is added (e.g., '-ial' in "commerce" and "commercial"). In most cases the derivation is performed without modifying the stem itself, as in "लघिमा" (lightness) from "लघ" (light), although some changes do occur when forming adjectives, such as "सांसारिक" (worldly) derived from "संसार" (the world).

The Hindi vocabulary is borrowed from both the Sanskrit and the Persian languages (with many terms also borrowed from Arabic via Persian), and as such Hindi may thus have two distinct words denoting the same item or a similar object (e.g., "पुस्तक" from Sanskrit or "किताब" from Persian). In these cases one is usually reserved as a technical term and the other for ordinary language. While this phenomenon is not unknown in English (e.g., "car" and "automobile" or "film" and "movie"), it occurs more frequently in Hindi and thus may impact retrieval effectiveness.

3.2 Key Features of Marathi Morphology

Marathi is spoken in western India by about 70 million people, and thus ranks fourth among the languages spoken there. As in other languages it

may include various dialects, along with certain spelling and phonological variations.

Marathi is written in the Devanagari script as well as another variant, the Balbodh script. Marathi contains 52 letters, of which only 50 represent distinct sounds. These sounds are expressed by 14 vowels having different initial-leading forms and also different shapes when following consonants. There are 36 consonants in all, including two compound consonants as well as nasal symbols.

As in Sanskrit, Marathi nouns may have three possible genders (masculine, feminine and neutral) and be singular or plural in number [Navalkar 2001]. Masculine, feminine or neutral noun forms are derived through applying regular and simple rules (for example, a child "मुलगा" (masculine), "मुलगी" (feminine), "मुलग" (neutral); or for a dog "कुत्रा" (masculine), "कुत्री" (feminine), "कुत्रें" (neutral)). As in other languages there are certain exceptions, such as the noun "camel" which has two distinct forms ("उंट" (masculine), "मांड" (feminine)).

The plural form of nouns depends on their gender. Masculine nouns ending in '-आ' become plural by changing the final vowel into '-ए', while others normally remain unchanged. The plural form of feminine nouns is usually derived by replacing the tailing '-अ' by '-इ' or by adding '-आ'. Neuter nouns ending in '-ए' usually become '-ईं' in the plural, while the rest become '-एं'. Table 1 depicts few examples.

Marathi is an inflected language with eight grammatical cases (nominative, accusative, instrumental, dative, ablative, genitive, locative, and vocative). A noun's inflectional termination depends on its case, number and gender, thus giving it a complex morpho-syntaxical construction that can be found in other Indo-European languages, such as Czech [Dolamic & Savoy 2010].

Case		"House"		"wise" (masc)	
		Singular	Plural	Singular	Plural
Nominative		घर	घरें	शहाणा	शहाणा
Accusative		घर	घरें	शहाणा	शहाणा
Instru-mental	by	घरानें	घरांशीं	शहाण्यानें	शहाण्यानें
	with	घराशीं	घरांस	शहाण्यानें	शहाण्यानें
Dative		घराला – स	घरांला – स – ना	शहाण्याला – म	शहाण्याला – म
Ablative		घराहून	घरांहूम	शहाण्याहून	शहाण्याहून
Genitive		घराचा	घरांचा	शहाण्याचा – ची – चें	शहाण्याचे – च्या – चीं
Locative		घरीं	घरीं	शहाण्यांत	शहाण्यांत
Vocative		घरा	घरांनों	शहाण्या	शहाण्या

Table 1. Examples of Marathi noun and adjective declination

The examples shown in Table 1 demonstrate how a noun may change its stem thus forming what is known as a crude (unfinished or imperfect) form in order to accommodate the various case terminations (e.g., the word "घर" (house, nominative singular) becomes "घरा" (instrumental, dative, ablative, genitive and vocative singular)). The crude form is usually formed by union of demonstrative pronouns 'या' (e.g., "आंबा" (mango) + 'या' = "आंब्या") or 'ई' ("भिंत" (wall) + 'ई' = "भिंती") with a noun. In certain declinations, these pronouns may also take on their impure forms 'आ' for 'या' (e.g., "घर" + 'आ' = "घरा") and 'ए' for 'ई' (e.g., "कथा" (tale) + 'ए' = "कथे"). Proper names for persons and certain terms used to express respect reject the 'या' in the crude-form thus thus the name Ravji "रावजी" becomes "रावजीला" (to Ravji) and not "रावज्याला" [Navalkar 2001].

In Marathi an adjective may be inflected according to the noun to which it is attached. When the adjective ends in '-आ' it is generally inflected otherwise it remains unaltered before the noun it qualifies. Finally, when an adjective is used as a substantive it is declined as such (see examples in Table 2).

"Good"	Singular	Plural
masculine	चांगला	चांगलें
feminine	चांगली	चांगल्या
neuter	चांगलें	चांगलीं

Table 2. Examples of gender-number agreement for the noun "good"

In Marathi there are four distinct ways of constructing derivational morphology. First, are the primary derivatives where only the radical vowel and/or consonant are modified (e.g., "डोळा" (an eye) → "डोळू" (an eyelet or a little hole)), and second those derivatives in which a prefix or a suffix is added to a given stem (e.g., "रवोडी" (mischief) → "रवोडकर" (mischievous)). This method is generally applied when in the derivation of new words adapted from the English language (e.g., from "history" we get the adjective "historic" or the related noun "prehistory"). A third method of forming new words involves reduplicates (e.g., "लाललाल", literally "red red", meaning "very red"), and finally when, two (or more) words are concatenated to from a new compound construction (such as "रण" (battle) + "भूमि" (field) = "रणभूमि" (battlefield)).

3.3 Key Features of Bengali Morphology

About 180 million people speak Bengali (or Bangla) in the eastern part of India and in Bangladesh, and thus it ranks second among the languages spoken in India. Although closely related to that used in Hindi, Bengali has its own script and an alphabet consisting of 35 consonants, 11 vowels along with five modifying symbols.

While the adjectival and nominal morphology in Bengali is very light, its verbs are highly inflected. Nouns are inflected according to seven grammatical cases (nominative, accusative, instrumental, ablative, genitive, locative and vocative), number (singular, plural) and determiners. The vocative is included in this list, yet strictly speaking it is not a case because it is identical in form to the nominative and can be distinguished by various

prefixes. Note that adjectives are invariable, and this simplifies the automatic processing of Bengali texts.

Bengali makes no use of gender distinction and thus all nouns are declined using the same terminations. Stems are usually not affected by the application of inflections and case-marking patterns may depend on a noun's degree of animacy (e.g., human beings, living beings other than human or inanimate objects). The noun "সন্তান" for example appears as such in the singular nominative, instrumental and ablative cases, but varies in other cases "সন্তানকে" (accusative, dative), "সন্তানেক" (genitive) and "সন্তানে" or "সন্তানেতে" (locative) while the plural forms of this noun are "সন্তানেরা", "সন্তান" and "সন্তানদের" [Beames 1891]. To express correct meaning more precisely, Bengali makes use of postpositions rather than the prepositions found in English.

The determiner in Bengali is attached to the noun as a suffix. The definite article for example adds the suffix '-টা' or '-টি' in the singular or adds the suffix '-দের' (animate) or the suffixes '-গুলা' or '-গুলি' (inanimate) for plural. Particles representing determiners must be placed before the case ending (e.g., "ছাত্র" (student) gives "ছাত্রটার" (the student's) and "ছাত্রদের" for the plural (the students')).

Note that additional suffixes may be found, such as those added to indicate measure, words added after the numeral and those that normally precede the noun. This suffix then becomes '-টা' (the same as the definite article) or '-জোণ' (reserved for persons) ("অনেকজন লোক" → "many people").

4. Suggested Stemming Strategies

In creating stemming procedures for the three Indian languages we adopted light stemmers, the same strategy we have suggested for other European languages over the past few years [Savoy 2006; Dolamic & Savoy 2010]. In our opinion effective stemming should focus mainly on inflectional suffixes attached to nouns and adjectives (sustaining most of a document's meaning) and ignoring numerous verb forms. Attempting to conflate all verb forms under a common stem tends to generate more stemming errors

than benefits. Moreover, stemmed forms obtained by the removal of suffixes related to number, gender and case variations tend to contain less erronenous forms and preserve more often the correct meaning of word involved. Additionally, most users are more capable of understanding the results of a light stemming procedure returning the dictionary entries ("initiatives" → "initiative") than a more aggressive procedure returning obscure terms ("initiatives" → "initi").

The "light" stemmers used in our experiments remove only the inflectional suffixes from nouns and adjectives, not taking into account for exceptions present in all natural languages (e.g., "mice" and "mouse). For the Hindi language, we suggested a light stemmer based on 20 rules, while for Marathi we created 51 rules and for Bengali 70 rules.

Suffixes may also be used to derive new words from a stem, usually by changing its part-of-speech (POS) (e.g., "care" and "careful" or "carefulness"). Thus for each language studied we also proposed and evaluated a more aggressive stemmer that not only removed the inflectional suffixes from nouns and adjectives, but also removed a limited number of derivational suffixes. To develop this more elaborate stemmer (denoted "aggressive" in our experiments), we have designated 49 rules for the Hindi language, 31 rules for Marathi and 85 for Bengali.

Finally, to identify pertinent matches between search keywords and documents we removed very frequently occurring insignificant terms such as "the", "but", "some", "we", "that" and "have". Following the guidelines provided by Fox [1990], we proposed a stopword list containing 165 Hindi, 114 Bengali and 99 Marathi terms. These lists were rather conservative and thus mainly included only determinants (e.g., "the", "this"), postpositions ("in", "near"), various pronouns ("we", "my") and conjunctions ("and", "while", "because"). They were also rather short compared to other Indo-European languages, (e.g. for the English language the SMART system [Salton 1971] suggests 571 words).

5. Test-Collections

The evaluations reported in this paper were based on the test-collections built for the Hindi, Marathi and Bengali languages during the first FIRE 2008 evaluation campaign. The corpora consist of newspaper articles extracted from the *Jagran* newspaper for the Hindi language, from the *Maharashtra Times* and *Sakal* for Marathi (articles spanning the period April 2004-September 2007) and from CRI & *Anandabazar Patrika* (a newspaper edited by ABP Ltd) for Bengali. The encoding system used for both documents and topic formulation is UTF-8.

	Hindi (HI)	Marathi (MR)	Bengali (BN)
Size (in MB)	718 MB	487 MB	732 MB
# of documents	95,215	99,357	123,047
# of distinct terms	127,658	511,550	249,215
Number of indexing terms per document			
Mean	356.2	264.6	291.88
Standard deviation	400.43	188.96	180.62
Median	256	222	265
Maximum	6,998	5,077	2,928
Minimum	0	28	0
Number of topics	45	73	75
Number rel. items	3,436	1,534	2,610
Mean rel./topic	76.4	21.0	34.8
Median	67	16	28
Maximum	194 (T#60)	123 (T#4)	149 (T#32)
Minimum	1 (T#59,T#66)	1 (T#12, #23) 1 (T#47, #50, #72)	4 (T#23)

Table 3. FIRE 2008 test-collection statistics

Table 3 lists statistics on the three corpora, showing that the Hindi and Bengali collections are similar in size (in MB) while the Marathi is smaller. In terms of numbers, the Bengali corpus contains the largest number of documents, while the Hindi or Marathi collections contain a relatively similar number. The Hindi corpus has a greater mean document length (based on the mean number of indexing terms per article, following stopword removal). The Bengali and Marathi corpora have similar mean

document lengths (about 275 indexing terms/article), based on the same measuring technique.

The Hindi, Marathi and Bengali language test-collections used in this study contain 45, 73 and 75 topics respectively. The available topics cover various subjects (e.g., Topic #028: "Iran's Nuclear Programme," Topic #034: "Jessica Lall Murder,") covering cultural issues (Topic #041: "Kolkata Book Fair 2007" or Topic #070: "Remake in Bollywood"), scientific problems (Topic #045: "Global Warming") or sports (Topic #073: "Zinedine Zidane's headbutting incident at the World Cup"). Certain topics seem to be more national in coverage (Topic #041: "New Labour Laws in France," Topic #058: "Thailand Coup"), while for others the real subject being covered is sometimes difficult to determine, at least based on the title section (Topic #049: "World wide natural calamities," Topic #052: "Budget 2006-2007"). Topic descriptions tend to contain many proper names (e.g., geographical with "Singur," "China," "Kolkata", personal names such as "Bush," "Sania Mirza," or products such as "Prince" and "Bofors"), as well as acronyms ("ULFA," "CBI," "HIV," "LOC").

Based on the TREC model, each topic formulation was divided into three logical sections. First a brief title (under the tag <TITLE>, see Figure 1) containing between 2 and 4 words, followed by a one-sentence description (tag <DESC>) the user's information need, and finally, a narrative part specifying the relevance assessment criteria (tag <NARR>). Full examples written in the Hindi, Marathi, Bengali and English languages are depicted in Figure 1. In our experiments and in order to closely reflect request sent to commercial search engines we used only the title part of topic description. This resulted in a mean query size 3.8 search terms for Hindi, 3.79 for Marathi and 3.65 for Bengali (following removal of stoplist words).

The bottom rows of Table 3 also compare the number of relevant documents per request, showing that the mean was always greater than the median (e.g., for Marathi, the average number of relevant documents per query was 21.0, and its corresponding median was 16). These findings indicate that only a relatively small number of relevant items were found per

<TOP lang="hi">
<NUM> 30 </NUM>
<TITLE> भारत के रेल मंत्री के रूप में लालू प्रसाद यादव </TITLE>
<DESC> रेलमंत्री के रूप में लालू प्रसाद यादव की भूमिका </DESC>
<NARR> रेलमंत्री के रूप में लालू की भूमिका, आधार संरचना के रूप में रेलवे का उन्नयन, अपने कार्यकाल के दौरान उसके द्वारा प्रस्तुत बजट के गुणदोष, उच्च शक्ति जाँच आयोग बिठा कर गोधरा रेल दुर्घटना की गुत्थी सुलझाने में लालू की भूमिका आदि सूचनाओं को संबद्ध प्रलेख में शामिल किया जाए। इनके अलावा अन्य सूचनाएँ यहाँ संगत नहीं हैं। </NARR> </TOP >

<TOP lang="mar">
<NUM> 30 </NUM>
<TITLE> भारतीय रेल्वे मंत्री म्हणून लालू प्रसाद यादव. </TITLE>
<DESC> भारतीय रेल्वे मंत्री म्हणून लालू प्रसाद यादवांची भूमिका. </DESC>
<NARR> एक रेल्वेमंत्री म्हणून लालूंची भूमिका, पायाभूत सोयींच्या बाबतीत रेल्वेच्या दर्जामध्ये सुधारणा, त्यांच्या कारकिर्दीत त्यांनी तयार केलेल्या रेल्वे अंदाजपत्रकातील सर्व लहान मोठ्या बाबी, गोध्रा येथील रेल्वेगाडी घटनेचा उलगडा करण्या-मध्ये उच्चाधिकार चौकशी आयोग बोलाविण्यामधील लालूंची भूमिका या संबंधीची माहिती संबंधित कागदपत्रात असली पाहिजे. या व्यतिरिक्त इतर माहिती येथे सुसंगत नाही. </NARR>
</TOP >

<TOP lang="bn">
<NUM> 30 </NUM>
<TITLE> রেল মন্ত্রী হিসেবে লালু প্রসাদ যাদব </TITLE>
<DESC> রেল মন্ত্রী লালু প্রসাদ যাদব এবং তাঁর আমলে ভারতীয় রেল সম্বন্ধে নথি খুঁজে বার করো। </DESC>
<NARR> রেল মন্ত্রী লালু প্রসাদ যাদবের সময়কালে ভারতীয় রেলের নিরাপত্তা, পরিকাঠামোগত উন্নতি এবং বিভিন্ন বিতর্ক প্রাসঙ্গিক নথিতে থাকা চাই। </NARR> </TOP >

<TOP lang="en">
<NUM> 30 </NUM>
<TITLE> Laloo Prasad Yadav as the Railway Minister </TITLE>
<DESC> The performance of Laloo Prasad Yadav and the Indian rail in his tenure.
</DESC>
<NARR> A relevant document should contain information about the safety measures taken by the Indian Railways, or infrastructural improvements planned or undertaken during the tenure of Laloo Prasad Yadav. Information about disputes / controversies surrounding Laloo are only relevant if they pertain to the Railways. </NARR> </TOP >

Figure 1. Topic description examples for the Hindi, Marathi, Bengali and English languages

topic. For Hindi no relevant records were found in the collection for five topics (#40, #43, #47, #48, and #50) while for Marathi Topic #70 ("Remake in Bollywood") did not have any relevant items.

Topic #32 ("Relations between Congress and its allies") returned the largest number of relevant articles in the Bengali collection (149), while for

example, Topic #59 (“Protests by American citizens against Iraq War”) and Topic #66 (“Khadim owner abduction case”) returned the smallest number of relevant documents (1 in this case and only for the Hindi corpus).

6. Information Retrieval Models

In order to ensure that useful conclusions would be obtained when analyzing new test-collections, we considered it important to evaluate retrieval performance under varying conditions to form a broad perspective. We thus evaluated a variety of indexing and search models, ranging from classical *tf idf* indexing schemes to more complex probabilistic models.

To evaluate and analyze different stemming approaches with respect to various IR models, we first used the classical *tf idf* vector-space model wherein the weight attached to each indexing term was the product of its term occurrence frequency (tf_{ij} for indexing term t_j in document d_i) and the logarithm of its inverse document frequency (idf_j). To measure similarities between documents and requests, we computed the inner product after normalizing (cosine) the indexing weights [Manning *et al*. 2008].

Better weighting schemes have been suggested for the vector-space model, especially in cases where the occurrence of a term in a document might be viewed as a rare event. A good practice may thus be to assign more importance to the first occurrence of a term compared to its successive and repeating occurrences, where the *tf* component is computed as the $\ln(tf) + 1$ or as $\ln(\ln(tf)+1)+1$. A term's presence in a shorter document might also provide stronger evidence than in a longer document. To take document length into account we could make use of more complex IR models, including the *Lnu-ltc* forms suggested [Buckley et al. 1996]. In this case Equation 1 calculates the indexing weight assigned to document term (*Lnu*) while Formula 2 gives the indexing weight assigned to query term (*ltc*).

$$w_{ij} = \frac{\left(\left(\ln(tf_{ij}) + 1\right) \Big/ \left(\ln(\text{mean tf}) + 1\right) \right)}{(1 - slope) \cdot pivot + slope \cdot nt_i} \quad (1)$$

$$w_{qj} = \left(\ln(tf_{qj}) + 1\right) \cdot idf_j \Big/ \sqrt{\sum_{k=1}^{t} \left(\left(\ln(tf_{qk}) + 1\right) \cdot idf_k \right)^2} \quad (2)$$

where nt_i is the number of distinct indexing terms in document d_i and *pivot* and *slope* are two constants used to adjust term weight normalization values, according to document length. This formulation prevents the retrieval system from overly favoring short documents compared to articles longer than the mean, depending on the *pivot* value.

To complement this vector-space model, we implemented three probabilistic models representing three different paradigms. First, we implemented the well-known Okapi (or BM25) approach [Robertson et al. 2000], regularly producing high retrieval effectiveness on various test-collections. Second, we included a model derived from *Divergence from Randomness* (DFR) paradigm [Amati & van Rijsbergen 2002], combining two information measures formulated as:

$$w_{ij} = Inf^1_{ij}(tf) \cdot Inf^2_{ij}(tf) = -\log_2[Prob^1_{ij}(tf)] \cdot (1 - Prob^2_{ij}(tf)) \quad (3)$$

where for the first information factor, $Prob^1_{ij}(tf)$ represents the pure chance probability of finding tf_{ij} occurrences of the term t_j in a document. If this probability is high, term t_j may correspond to a non-content bearing word within the context of the entire collection [Harter 1975] and otherwise if $Prob^1_{ij}(tf)$ is small (or if $-\log_2[Prob^1(tf)]$ is high), the term t_j would provide important information regarding the content of the document d_i. The second information measure depends on $Prob^2_{ij}(tf)$, the probability of having tf+1 occurrences of the term t_j, knowing that tf occurrences of this term have already been found in document d_i. To implement these two underlying probabilities, we selected the $I(n_e)C2$ model based on the following formulae:

$$Inf^{1}_{ij} = tfn_{ij} \cdot \log_2[(n+1) / (n_e+0.5)] \quad (4)$$

with $n_e = n \cdot [1 - [(n-1)/n]^{tc_j}]$

and $tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((c \cdot mean\ dl)/l_i)]$

$$Prob^{2}_{ij} = 1 - [(tc_j + 1) / (df_j \cdot (tfn_{ij} + 1))] \quad (5)$$

where tc_j indicates the number of occurrences of term t_j in the collection, n the number of documents in the corpus, *mean dl* the mean length of a document and l_i the length of document d_i.

Finally we also used an approach based on a language model (LM) [Hiemstra 2000], known as a non-parametric probabilistic model. Within this language model paradigm various implementations and smoothing methods [Zhai & Lafferty 2004] might also be considered, and in this paper we adopted the model proposed by Hiemstra [2000] as described in Equation 6, using the Jelinek-Mercer smoothing and combining an estimate based on document ($P[t_j | d_i]$) and one based on the entire corpus ($P[t_j | C]$).

$$Prob[q_i | q] = Prob[d_i] \cdot \prod_{tj \in Q} [\lambda_j \cdot Prob[t_j | d_i] + (1-\lambda_j) \cdot Prob[t_j | C]]$$

$$\text{with } Prob[t_j | d_i] = \left(\frac{tf_{ij}}{l_i}\right) \text{ and } Prob[t_j | C] = \left(\frac{df_j}{lc}\right) \text{ with } lc = \sum_{k=1}^{t} df_k \quad (6)$$

where λ_j is a smoothing factor (fixed at 0.35 for all indexing terms t_j), df_j indicates the number of documents indexed with the term t_j, and lc is a constant related to the underlying corpus C.

7. Evaluation

To evaluate the various indexing and search strategies, we adopted mean average precision (MAP) method of measuring retrieval performance (computed by the TREC_EVAL software based on a maximum of 1,000 retrieved records). Used by all evaluation campaigns for around 20 years, this performance measure is able to objectively compare various IR models, especially their ability to retrieve relevant items (*ad hoc* tasks) [Buckley & Voorhees 2005].

Using MAP to measure a system's performance signifies that we attached equal importance to all queries. Comparisons between two IR strategies would therefore not be based on a single query with respect to those available in the underlying test-collection or specifically created to demonstrate that a given IR approach must be rejected. Thus we believe that it is important to conduct experiments involving the largest possible number of observations (between 45 and 75 queries in our evaluations, depending on the language).

To statistically determine whether or not a given search strategy would be better than another, we applied the bootstrap methodology [Savoy 1997]. This led to a conclusion very similar to that of the *t*-test method but did not require parametric assumptions [Abdou & Savoy 2006]. In our statistical tests, the null hypothesis H_0 stated that both retrieval schemes produce similar MAP performance. This null hypothesis would be accepted if two retrieval schemes returned statistically similar MAP, otherwise it would be rejected. Thus, in the experiments presented in this paper, statistically significant differences were detected by a two-sided test (significance level $\alpha = 5\%$).

7.1 IR Model Evaluation

We evaluated the various IR models described in the previous section, applying them to the Hindi (see Table 4), Marathi (Table 5) and Bengali test-collections (Table 6). These report the MAP achieved by the IR models when applying the three different stemming strategies for the Hindi and Marathi (e.g. "None", "Light" and "Aggress"), and four for Bengali (e.g. "GM" was added to the other three stemming schemes). In each table, the last two columns list the retrieval performances produced by two language independent indexing strategies. Listed under the heading "Trunc-4" in these tables are the results of simply truncating the term into its first few letters (e.g., "goodness" generates "good"), while listed in the last column are the results of evaluations obtained by applying the 4-gram indexing approach (e.g., "minister" gives "mini", "inis", …, "ster") [McNamee & Mayfield 2004; McNamee *et al*. 2009]. The fixed length of 4 was selected

for both the truncating and *n*-gram methods because it produced the best IR performance for all three languages.

Table 4 lists the best performance results for the Hindi language, obtained by either the I(n_e)C2 model derived from *Divergence from Randomness* paradigm or the vector-space *Lnu-ltc*. As shown in bold in Table 4, this latter scheme performed best when we applied an aggressive stemming or when we ignored this word normalization procedure. Tables 5 and 6 show that for both Bengali and Marathi the best performing model was always the I(n_e)C2 approach.

	Mean Average Precision					
	None	Light	Aggress	YASS	Trunc-4	4-gram
tf idf	0.1548†	0.1756†*	0.1748†	0.1588†	0.1987†	0.1750†
Lnu-ltc	**0.2368**	0.2844*	**0.2981***	**0.2843**	0.2852	0.2516
Okapi	0.2179	0.2601*	0.2811*	0.2598	0.2867*	0.2495†
I(n_e)C2	0.2311	**0.2692***	0.2936*	0.2753	**0.2966***	**0.2629**
LM	0.1872†	0.2369†*	0.2640†*	0.2368†	0.2730†*	0.2199†
Average	0.2056	0.2452	0.2623	0.2430	0.2680	0.2318
% change		+19.3%	+27.6%	+18.2	+30.4%	+12.8%

Table 4. MAP of various indexing strategies and IR models for the Hindi language (45 queries)

A difference in mean performance, particularly when small, did not always indicate differences that might be clearly perceived by the final user. A cross ("†") in these tables indicates which retrieval models that resulted in statistically significant performance differences, compared to the best performing models. In this case, the classical *tf idf* vector-space model and the language model (LM) typically resulted in significantly lower performance levels. For the other models, the outcome varied according to the language and the indexing scheme involved. It is however evident that performance differences between the *Lnu-ltc* and I(n_e)C2 for the Hindi language were never significant, while for the Bengali corpus performance

differences between the I(n_e)C2 and the other approaches always tended to be significant (exceptions can be found only in the "None" column, see Table 6).

	Mean Average Precision					
	None	Light	Aggress	YASS	Trunc-4	4-gram
tf idf	0.1844†	0.1920†	0.2518†*	0.1886†	0.2299†	0.2394†
Lnu-ltc	0.2152†	0.2542†*	0.3085*	0.2507*	0.3137*	0.2929†*
Okapi	0.2359	0.2759*	0.3438*****	0.2626	0.3307*	0.3268*
I(n_e)C2	**0.2416**	**0.2839***	**0.3517***	**0.2770***	**0.3368***	**0.3418***
LM	0.2232	0.2480†	0.3027*	0.2472†	0.3102*	0.2929†*
Average	0.2201	0.2508	0.3117	0.2452	0.3043	0.2988
% change		+13.9%	+41.6%	+11.4%	+38.3%	+35.8%

Table 5. MAP of various indexing strategies and IR models for the Marathi language (73 queries)

	Mean Average Precision						
	None	Light	Aggress	YASS	GM	Trunc-4	4-gram
tf idf	0.1876†	0.2015†	0.2144†*	0.2247†*	0.2114†*	0.2102†	0.1987†
Lnu-ltc	0.2539	0.2897†*	0.2979†*	0.3058†*	0.2831†*	0.3242†*	0.2590†
Okapi	0.2662	0.2966†*	0.3066†*	0.3066†*	0.2893†*	0.3310†*	0.2662†
I(n_e)C2	**0.2628**	**0.3064***	**0.3132***	**0.3243***	**0.2990***	**0.3390***	**0.2830**
LM	0.2353†	0.2683†*	0.2780†*	0.2747†*	0.2585†*	0.2947†*	0.2418†
Average	0.2412	0.2725	0.2820	0.2878	0.2683	0.2998	0.2497
% change		+13.7%	+17.7%	20.1%	+11.9%	+25%	+4.2%

Table 6. MAP of various indexing strategies and IR models for the Bengali language (75 queries)

7.2 Stemming Evaluation

The Hindi, Marathi and Bengali morphologies are more complex than that of the English language and thus for the former the MAP could be improved by applying a stemming procedure that would conflate different surface words having similar meanings under the same stem or indexing unit. If this assumption is true, we could then consider a variety of stemming strategies, be they light or more aggressive. The question then arises as to whether stemming would affect IR performances for these various languages and to which extent?

Table 4 (Hindi), 5 (Marathi) and 6 (Bengali) lists the results of our first retrieval performance evaluations in which the stemming was omitted, and lists the MAP values under the "None" column. The "Light" column lists retrieval performances obtained by applying a light stemmer and the "Aggress" column a more aggressive stemmer. We also evaluated the performance of the stemmer proposed by Gungaly & Mitra [2008] (GM) for the Bengali language only.

We listed in the last but one line the average retrieval performance for all five IR models (to obtain an MAP overview for each of these stemming strategies). Finally the last row (labeled "% change") lists the results computed based on comparing the percentage improvement to the mean performance, obtained when we ignored the stemming stage (listed in the "None" column).

The last two rows in the above-mentioned tables show how all approaches applying stemming performed more effectively than those indexing strategies that omitted stemming, and this finding holds for all three languages studied. More precisely, for the Hindi language there were relative increases ranging from 19.3% with a light stemmer to 27.6% with the more aggressive approach. For Marathi this increase was from 13.9% with the light stemming approach to 41.6% when the aggressive stemming method was performed. Finally for Bengali the range of improvement was relatively similar across all three stemmers, ranging from 11.9% with the GM stemmer to 17.7% for the aggressive stemmer. Based on this data, we found that a more aggressive stemmer tended to result in better MAP while

for some languages (e.g., Bengali) the performance difference between a light and an aggressive stemmer was not significant. Moreover, when compared to those found for certain European languages, these relative improvements were quite large (e.g., 4% for English, 4.1% for Dutch, 7% for Spanish, 9% for French, 15% for Italian, 19% for German, 29% for Swedish and 40% for Finnish [Tomlinson 2004] and 45% for Czech [Dolamic & Savoy 2010]).

When the no stemming approach was taken as a baseline and after applying statistical tests the differences between this approach and the other word-based approaches were very often statistically significant (marked with "*" in the tables) for all three languages tested. For Marathi there were only two exceptions found (see Table 5), where performance differences resulting from the classical *tf idf* or the LM models cannot be viewed as significant (*tf idf*: 0.1844 vs. 0.1920; LM: 0.2232 vs. 0.2480).

To analyze the effect of applying a stemming, we performed a query-by-query analysis, concentrating on DFR-I(n_e)C2, the best performing retrieval model, and for each query comparing average precision (AP) before and after applying a stemmer.

For Hindi we found the first explanation for this improved performance resulting from applying the stemming approach. The topics and their corresponding relevant documents contained the same word but were expressed in different grammatical cases. Even though the two strings were not identical before stemming, they were indeed conflated to the same stem, thus resulting in the best performance. As an example, the title of Topic #33 ("President Bush visits India") contained "बुश" and "भारत" ("Bush" and "India" respectively, in direct case), while a large number of relevant documents contained "बुशा" ("Bush") or "भारतीय" ("India") in the oblique cases.

In Marathi case, Topic #41 ("New labor laws in France") could serve as an example demonstrating the second advantage of applying a stemming procedure. The topic title contains the term "फ्रान्स" ("France") while the relevant documents contain the following terms "फ्रान्सचे", "फ्रान्सचा", "फ्रान्सचया" all of which are conflated to the same stem by the aggressive stemmer and thus resulting in average precision changes from 0.0111 ("none") to 0.6389

("Aggress"). The topic title also contains the noun "कायदे" (law) while some relevant documents contain only the derivational term "कायदयावर" (legal). With the light stemmer, the various surface forms were not conflated under the same stem.

For Bengali the largest AP differences between the different indexing strategies were observed with Topic #58 ("Thailand Cup"). The term "থাইল্যাণ্ডের" ("Thailand" in the genitive case) was only found in the topic formulation, while the terms "থাইল্যাণ্ড" and "থাইল্যাণ্ডের" found in relevant documents had been conflated to the same stem for all stemming strategies applied, thus resulting in much better AP for that particular stemming method. Certain relevant documents contained however the form "তাইল্যাণ্ড", a different spelling of the country name.

7.3 Word-Based vs. *n*-gram Indexing Strategies

The last two columns of previous tables report the results of our tests on the Indian languages: Hindi, Marathi and Bengali showing the retrieval performances obtained by "Trunc-4" and "4-gram", the two language independent approaches we applied. For these indexing strategies we ignored the each language's underlying morphology and syntax, assuming that the first part of the word (trunc-*n*) or character sequence (*n*-gram) would provide the information needed to obtain a pertinent match between the search keywords and document surrogates.

Based on the retrieval performances obtained for the Hindi (Table 4), Marathi (Table 5) or Bengali (Table 6) languages, the simple trunc-*n* (or trunc-4 in our evaluation) resulted in high performance levels. On average for both Hindi and Bengali this indexing approach led to the best retrieval effectiveness. Moreover with the *n*-gram approach the mean performance differences were relatively large (Hindi: 0.2680 vs. 0.2318, -13.6%; Bengali: 0.2998 vs. 0.2497, -16.7%) while for Marathi the average retrieval performances were similar (0.3043 vs. 0.2988, -1.8%).

To verify whether these differences might be considered significant we performed a statistical test using the "trunc-4" indexing approach as a baseline

with both the Hindi and Bengali languages. For Hindi (see Table 4) performance differences were always statistically significant when compared to "None" and "4-gram" (except for *tf idf* model). When compared to light and aggressive stemming approaches, the differences with the trunc-4 were never statistically significant. For Bengali (see Table 6) however the performance differences were never statistically significant when compared to the aggressive stemming approach, yet for other approaches they were mostly significant (except for *tf idf* model).

For Marathi (Table 5) the best overall performance was achieved using the aggressive stemming approach. Using this best performance as baseline, the performance differences were always statistically significant when compared to "None" or with the light stemming approach, but they were not significant when compared to "Trunc-4" or "4-gram" strategies.

To provide a more complete picture we should mention that for Bengali and Hindi, the MAP differences between "None" and 4-gram indexing strategies were never statistically significant.

To obtain a better understanding of when word-based or language independent indexing strategies performs best, we analyzed a few specific queries. As a first explanation for performance differences, we noted certain spelling variations in topic formulations and in relevant documents. With the Hindi corpus for example in Topic #44 ("Terrorist attacks in Britain") the term terrorist was spelled "अआतकवादी" while it was spelled "आतकवादी" in the relevant documents. In the Marathi corpus, we found a similar situation with Topic #39 ("Attacks on American soldiers in Iraq") where the corresponding script was "बॉंबस्फोट" while in the relevant documents it was "बॉबस्फोट" or "बॉमबस्फोट". In these cases the 4-gram was the best performing strategy, which produced at least one match for the two terms. In Marathi we also encountered a spelling problem, as in Topic #9 with "Israel" being spelled "ईस्राएल" in the topic formulation and "इस्राईल" in the relevant documents. In this case also the 4-gram approach performed better than a word-based indexing scheme.

For the Marathi case, Topic #41 ("New labor laws in France") can be cited as an example of a improved retrieval effectiveness following the application of a word-based indexing approach. The topic title contains the term

"फ्रान्स" ("France") while the relevant documents contain the following terms "फ्रान्सचे", "फ्रान्सचा", "फ्रान्सचया" all of which were conflated to the same stem by the aggressive stemmer and thus resulted in a change in average precision from 0.1946 with trunc-4 to 0.6389 with the aggressive stemmer, due to over-stemming by trunc-4 strategy. For this query the fixed limit of 4 was clearly too small.

7.4 Stopword List Evaluation

During the indexing of documents or queries, it is assumed that very frequent word forms having no precise meaning (e.g., "the", "you", "of", "is") may be removed. In fact, each match between a query and a document should be based on pertinent terms, rather than retrieving documents simply because they contain words such as "an", "ours" or "but".

In our final experiments we compared the retrieval effectiveness of various IR models, with and without the suggested stopword list, for each of the three languages studied. For the Marathi language (the stopword list contained 99 words) and the mean difference over the five retrieval models tested was around 1%, while for Bengali (114 stopwords) this difference was around 2%. For both these languages the differences were never statistically significant.

For Hindi (165 forms in the stopword list) however the mean differences between various search models with or without applying a stopword list were larger. Table 7 shiws by the results obtained when ignoring the stemming stage (columns labeled "No stemmer") and after applying a light stemmer ("Light stemmer). As we can see in the last two rows, the average performance differences were around 20% for both stemming strategies. Using the retrieval performance with stopword removal as a baseline, any significant MAP differences detected were also listed in Table 7 and marked with the symbol "*". As can be seen the differences were always statistically significant, except for those obtained using the *Lnu-ltc* model, with no stemmer applied (0.2368 vs. 0.2182).

	Mean Average Precision			
	No stemmer		Light stemmer	
	With	Without	With	Without
tf idf	0.1548	0.1024*	0.1756	0.1187*
Lnu-ltc	0.2368	0.2182	0.2844	0.2547*
Okapi	0.2179	0.1593*	0.2601	0.1969*
$I(n_e)C2$	**0.2311**	**0.2020***	**0.2692**	**0.2374***
LM	0.1872	0.1664*	0.2369	0.2138*
Average	0.2056	0.1697	0.2452	0.2043
% change	+21.2%		+20.0%	

Table 7. MAP with and without stopword removal for the Hindi corpus (45 queries)

Based on an analysis of mean query length across the three languages, we found that removing the stopwords only slightly changed the averages for the Marathi (from 4.04 to 3.78 search terms) or Bengali languages (from 3.80 to 3.64 terms). For the Hindi however this difference was somewhat greater, decreasing from 4.80 indexing terms without stopwords removal to 3.8 terms if this step was performed.

This important improvement resulting from stopword removal for Hindi can be partially explained by an analysis of Topic #27 ("Relations between India and China") showing an AP of 0.2690 after removal vs. 0.0532 before stopword removal. The situation was similar with Topic #38 ("Uneasy truce between Greg Chappell and Sourav Ganguly") with an AP of 0.6055 (after) vs. 0.2221 (before) or Topic #54 ("HIV and AIDS epidemic") providing an AP of 0.7271 vs. 0.2929 (before). In these three topics, it is the term "और" meaning "and" that makes the difference. This word did not have high document frequency because other words can be used to express "and" in Hindi (in fact the underlying document frequency is similar to a word like "China"). This resulted in the term being incorectly viewed as pertinent for the given queries and thus hurt the resulting retrieval effectivness.

8. Conclusion

This paper presents the main morphological characteristics of the Hindi, Marathi and Bengali languages. To facilitate IR operations in each of these Indo-Aryan languages we suggest two algorithmic stemmers, one removing only inflectional suffixes (denoted "Light") and a second also removing certain frequently occurring derivational suffixes (listed our tables under the heading "Aggress"). To compare these word-based indexing models with language independent approaches, we also include an *n*-gram and trunc-*n* indexing scheme. We also propose a stoplist for each of these languages which for Hindi contains 165 words, 99 for Marathi and 114 for Bengali.

To evaluate and compare these various indexing approaches, we use five different IR models corresponding to different probabilistic approaches (Okapi, one model derived from the *Divergence from Randomness* (DFR) paradigm, and another a language model) as well at two vector-space approaches, namely the classical *tf idf* and the *Lnu-ltc* weighting schemes.

For all three languages and independently of the indexing approaches, we find that the $I(n_e)C2$ derived from *Divergence from Randomness* (DFR) paradigm tends to produce the best retrieval performance. Only for the Hindi corpus (see Table 4) does the *Lnu-ltc* vector-space model result in better performances in relation to two indexing strategies (applying an aggressive stemmer or ignoring this word normalization procedure). Based on the application of a statistical test for each of these three languages, we conclude that performance differences are statistically significant when comparing the best performing model with both the classical *tf idf* and the language model (LM). For the Bengali corpus however when comparing the best IR model ($I(n_e)C2$) and the others (see Table 6) the differences are always significant (denoted by a "†").

Upon an analysis of performance differences resulting from the application of the various stemming strategies, we find that for all three languages a stemming approach tends to perform significantly better than an indexing scheme without a stemmer. Moreover, an aggressive stemmer usually results in better MAP than a light stemmer, and these performance difference

are even statistically significant although this holds for the Marathi language only (see Table 5).

Language independent indexing strategies such as *n*-gram and trunc-*n* are valid alternatives, especially for unfamiliar languages. For the three languages studied, truncation after the first *n* characters tends to produce better MAP than the *n*-gram scheme. When comparing word-based indexing strategies using an aggressive stemmer, mean differences tend to be relatively small, +2.2% for the Hindi test-collection (0.2680 "Trunc-4" vs. 0.2623 "Aggress") or -2.4% for the Marathi corpus (0.3043 "Trunc-4" vs. 0.3117 "Aggress"). For Bengali however mean performance differences are larger (0.2998 "Trunc-4" vs. 0.2820 "Aggress", -5.9%).

When comparing all indexing schemes, we find that for the Hindi language the best approach is either trunc-4 or word-based with either a light or an aggressive stemmer, even though performance differences between these three schemes are usually not statistically significant. For Marathi our statistical tests detect no significant differences between an aggressive stemmer, the trunc-4 or the 4-gram schemes. According to our statistical tests for Bengali both the trunc-4 method and the aggressive stemmers lead to similar performance levels.

When comparing retrieval performances with or without the removal of stopword, there appear to be no real and significant differences for the Marathi and Bengali languages. For Hindi however the use of a stopword list significantly improves retrieval performances, with average differences being around 20% (see Table 7).

ACKNOWLEDGEMENTS

This research was supported in part by the Swiss National Science Foundation under Grant #200021-113273.

References

Amati, G., & van Rijsbergen, C.J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM-Transactions on Information Systems*, 20(4), 357-389.

Beames, J. (1891). *Grammar of the Bengali Language, Literary and Colloquial*. Clarendon Press, Oxford (UK).

Braschler, M., & Peters, C. (2004). Cross-language evaluation forum: Objective, results, achievements? *IR Journal*, 7(1-2), 7-31.

Braschler, M. & Ripplinger, B. (2004). How Effective is Stemming and Decompounding for German Text Retrieval? *IR Journal*, 7(3-4), 291-316.

Buckley, C., Singhal, A., Mitra, M. & Salton, G. 1996. New retrieval approaches using SMART. In *Proceedings of TREC-4*, Gaithersburg, MD, November 1995. D.K. Harman (Eds), NIST Special Publication 500-236, 25-48.

Buckley, C., & Voorhees, E.M. (2005). Retrieval system evaluation. In E.M. Voorhees, D.K. Harman (Eds): *TREC. Experiment and Evaluation in Information Retrieval* (pp. 53-75). The MIT Press, Cambridge (MA).

Di Nunzio, G.M., Ferro, N., Melucci, M., & Orio, N. (2004). Experiments to evaluate probabilistic models for automatic stemmer generation and query word translation. In *Comparative Evaluation of Multilingual Information Access Systems* (pp. 220-235). LNCS #3237, Berlin: Springer.

Dolamic, L, & Savoy, J. (2010). Indexing and stemming approaches for the Czech language. *Information Processing & Management*, to appear.

Fox, C. (1990). A stop list for general text. *ACM-SIGIR Forum*, 24, 19-35.

Gungaly, D. & Mitra, M. (2008). Using language modeling at FIRE 2008 Bengali monolingual track. In Proceedings FIRE'2008 (available at http://www.isical.ac.in/~fire/paper/lm_at_fire.pdf, visited March 2008).

Harter, S.P. 1975. A probabilistic approach to automatic keyword indexing. Part I: On the distribution of specialty words in a technical literature. *Journal of the American Association for Information Science*, 26, 197-216.

Hiemstra, D. (2000). *Using Language Models for Information Retrieval*. CTIT Ph.D. Thesis.

Kellogg, S.H. (1938). *A Grammar of the Hindi Language*. Kegan Paul, Trench, Trubner & Co. Ltd., London (UK).

Kettunen, K. & Airo, E. (2006). Is a morphologically complex language really that complex in full-text retrieval? In *Advances in Natural Language Processing* (pp. 411-422). LNCS #4139, Berlin: Springer.

Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2004). Stemming and lemmatization in the clustering of Finnish text documents. In *Proceedings of the ACM-CIKM*. The ACM Press, Washington (DC), 625-633.

Krovetz, R. (1993). Viewing morphology as an inference process. In *Proceedings of the ACM-SIGIR'93*, The ACM Press, Pittsburgh (PA), 191-202.

Lovins, J.B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1), 22-31.

Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge (UK).

Masica, C.P. (1991). *The Indo-Aryan Languages*. Cambridge University Press, Cambridge (UK).

McNamee, P., & Mayfield, J. (2004). Character *n*-gram tokenization for European language text retrieval. *IR Journal*, 7(1-2), 73-97.

McNamee, P., Nicholas, C., & Mayfield, J. (2009). Addressing morphological variation in alphabetic languages. In *Proceedings of the ACM-SIGIR'09*, ACM, New York, 75-82.

Navalkar, G.R. (2001). *The Student's Marathi Grammar*. Asian Education Services, New Dehli.

Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A. & Santos, D. (Eds.) (2008). *Advances in Multilingual and Multimodal Information Retrieval*. LNCS #5152, Springer-Verlag, Berlin.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130-137.

Ramanathan, A. & Rao, D. (2003). A Lightweight stemmer for Hindi. In *Proceedings Workshop of Computational Linguistics for the South Asian Languages, EACL-2003* (pp. 42-48). Budapest (Hungary).

Robertson, S.E., Walker, S., & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. *Information Processing & Management*, 36(1), 95-108.

Sakar, S. & Bandyopadhyay, S. (2008). Design of a rule-based stemmer for natural language text in Bengal. In *Proceedings IJCNLP-08 Workshop on NLP for Less Privileged Languages*, (pp. 65 72). Hyderabad (India).

Salton, G. (Ed.) (1971). *The SMART Retrieval System: Experiments in Automatic Document Processing*. Englewood Cliffs (NJ): Prentice-Hall.

Savoy, J. (1993). Stemming of French words based on grammatical category. *Journal of the American Society for Information Science*, 44(1), 1-9.

Savoy, J. (1997). Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management*, 33(4), 495-512.

Savoy, J. (2006). Light stemming approaches for the French, Portuguese, German and Hungarian Languages. In *Proceedings ACM-SAC*, (pp. 1031-1035). Dijon (France).

Sproat, R. (1992). *Morphology and Computation*. Cambridge: The MIT Press.

Tomlinson, S. (2004). Lexical and algorithmic stemming compared for 9 European languages with Humminbird SearchServer™ at CLEF 2003 (2004). In *Comparative Evaluation of Multilingual Information Access Systems*. LNCS #3237, Springer-Verlag, Berlin, 286-300.

Xu, J., & Croft, B. (1998). Corpus-based stemming using cooccurrence of word variants. *ACM-Transactions on Information Systems*, 16(1), 61-81.

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2), 179-214.

How effective is Google's translation service in search?

Jacques Savoy, Ljiljana Dolamic

Computer Science Dept., University of Neuchatel,
Rue Emile Argand 11, 2009 Neuchâtel, Switzerland
{Jacques.Savoy, Ljiljana.Dolamic}@unine.ch

Introduction

In multilingual countries (Canada, Hong Kong, India, etc.) and large international organizations or companies (e.g., WTO, European Parliament), and among web users in general, accessing information written in other languages has become a real need (news, hotel or airline reservations, government information, statistics, etc.). While some users are bilingual, others can read documents written in another language but cannot formulate a query to search it, or at least cannot provide reliable search terms in a form comparable to those found in the documents being searched. There are also many monolingual users who may want to retrieve documents in another language and then have them translated into their own language, either manually or automatically. Translation services may however be too expensive, not readily accessible or not available within a short timeframe. On the other hand, many documents contain non-textual information such as images, videos and statistics that do not need translation and can be understood regardless of the language involved. In response to these needs and in order to make the web universally available regardless of any language barriers, in May 2007 Google launched a translation service that now provides two-way online translation services mainly between English and 11 other languages, namely Arabic, simplified and traditional Chinese, French, German, Italian, Japanese, Korean, Portuguese, Russian, and Spanish (http://translate.google.com/). Over the last few years other free internet translation services have been made available as for example by BabelFish (http://babel.altavista.com/) or Yahoo! (http://babelfish.yahoo.com/). These two systems are similar to that used by Google, given they are based on technology developed by Systran, one of the earliest companies to develop machine translation. Also worth mentioning here is the Promt system (also known as Reverso, http://translation2.paralink.com/), which was developed in Russia to provide mainly translation between Russian and other languages.

The question we would like to address here is to what extent a translation service such as Google can produce adequate results in the language other than that being used to write the query. Although we will not evaluate translations *per se* we will test and analyze various systems in terms of their ability to retrieve items automatically based on a translated query. To be adequate, these tests must be done on a collection of documents written in one given language plus a series of topics (expressing user information needs) written in other languages, plus a series of relevance assessments (relevant documents for each topic).

Evaluation Campaigns

In an effort to promote information retrieval (IR) in languages other than English and also to evaluate bilingual searches (queries expressed in one language, documents retrieved in another), there have been various evaluation campaigns conducted over the last few years. The first was the Text REtrieval Conference or TREC [1] in 1992, another took place in 1999 specifically for Far-East languages (the NTCIR series) [2], and beginning in 2000, CLEF [3] evaluation campaigns have been held for various European languages. The outcome of all these various international efforts was several test collections, created in various languages.

For our own tests and in an attempt to objectively evaluate Google's translation service, we used collections written in French and made up of articles published in the French newspaper *Le Monde* (1994 and 1995), plus others from the Swiss news agency (*ATS, Agence Télégraphique Suisse*) published during the same period. These collections were put together during six CLEF evaluation campaigns and contain a total of 177,452 documents (or about 487 MB of data). On average each article contained about 178 content-bearing terms (median: 126); not counting commonly occurring words such as "la," "de" or "et"). Typically, documents in this collection were represented by a short title plus one to four paragraphs of text.

These collections also contain 310 topics, each subdivided into a brief title (denoted as T), a full statement of their information need (called description or D), plus any background information that might help assess the topic (narrative or N). The topic titles consist of 2 or 3 words reflecting typical web requests, and are represented by a set of capitalized keywords rather than a complete grammatical phrase. These topics cover various subjects (e.g., "U.N./US Invasion of Haiti," "Consumer Boycotts," "Lottery Winnings", "Tour de France Winner" or "James Bond Films"), along with both regional ("Swiss Referendums," "Corruption in French Politics") and international coverage ("Crime in New York," "Euthanasia").

	2001	2002	2003	2004	2005	2006
Source	*Le Monde* 94 ATS 94	*Le Monde* 94 ATS 94	*Le Monde* 94 ATS 94-95	*Le Monde* 95 ATS 95	*Le Monde* 94-95 ATS 94-95	*Le Monde* 94-95 ATS 94-95
Size	243 MB	243 MB	331 MB	244 MB	487 MB	487 MB
# docs	87,191	87,191	129,806	90,261	177,452	177,452
# topics	49	50	52	49	50	49
Topics	#41 - #90	#91 - #140	#141 - #200	#201 - #250	#251 - #300	#301 - #350

Table 1. General statistics on our test-collection for each year

Relevance judgments (correct answers) were supplied by human assessors throughout the various CLEF evaluation campaigns. For example, Topics #201 to #250 were created in 2004 and responses were to result from searches in the *Le Monde* (1995) and *ATS* (1995) collections, a subset representing 90,261 documents. Of the 50 queries originally available in 2004, we found that only 49 having at least one correct answer.

In all, 11 queries were removed because they did not have any relevant information, meaning only 299 (310 minus 11) topics were used in our evaluation. Upon an inspection of these relevance assessments, the average number of correct responses for each topic was 30.57 (median: 16), with Topic #316 ("Strikes") obtaining the greatest number of correct responses (521).

Information Retrieval Models

To search for pertinent items within this corpus, we used a vector-space model based on the classical *tf idf* scheme [4]. In this case the weight attached to each indexing term in the document (or in the query) was the product of the term occurrence frequency (or *tf*) and the inverse of the document frequency (or *idf*). Based on this formula, greater importance is attached to terms occurring frequently in the document (*tf* component), and in relatively few different documents (*idf* component).

We also applied the Okapi probabilistic model [5] in which a term's weight also depends on its discriminating power (the fact that this term occurs mainly in the relevant or non-relevant items) and on document length (weights attached to longer items are reduced).

Finally, we also applied an approach based on a statistical language model (LM) [6], which tries to estimate the occurrence probability of words, or in more sophisticated models, sequences of two words. In our experiments, the underlying estimates were based on a linear combination of occurrence frequencies both within the document and within the entire corpus.

Evaluation Methodology

To measure the retrieval performance obtained with these three IR models, we adopted a method known as the mean reciprocal rank (MRR) [7]. For any given query, *r* is the rank of the first relevant document retrieved and the query performance is computed as 1/*r* or the reciprocal rank (RR). This value varies between 1 (the first retrieved item is relevant) and 0 (no correct response among the top 1,000 documents). It should be noted here that ranking the first relevant item in second place instead of first would seriously reduce the RR value, making it 0.5 instead of 1. Similarly, ranking the first relevant item in the 20th position (0.05) or lower would produce a very small RR. To measure the retrieval performance resulting from several queries, we simply computed the mean over all the queries. This value served as a measure of any given search engine's ability to extract one correct answer and list it among the top-ranked items. We thus believe that MRR value closely reflects the expectation of those internet surfers who are looking for a single good response to their requests.

In IR, not only do we want to measure a search system's ability to rank one relevant item, but also to extract all relevant information from the collection [7]. Users want both high precision (fraction of retrieved items that are relevant) and high recall (fraction of relevant items that have been retrieved). In other words they want "the

truth, only the truth (precision), and nothing but the truth (recall)". To meet this need we compute the average precision for each query by measuring the precision achieved at each relevant item extracted and then computing an overall average. Then for a given set of queries we calculate the mean average precision (MAP), which varies between 0.0 (no relevant items found) and 1.0 (all relevant items always appear in the top of the ranked list). Higher MAP values are thus more difficult to obtain than higher MRR values, due to the fact that the MAP accounts for the rank of all relevant items, and not just the first one.

Using the mean to measure a system's performance signifies that equal importance has been attached to all queries. Comparisons between two different IR strategies would therefore not be based on a single query but rather demonstrates that a single IR approach should not be rejected. Our approach is thus based on the importance of conducting experiments involving a large number of observations (in this study there were 299).

Finally, in an effort to statistically determine whether or not a given search strategy would be better than another, we applied the bootstrap methodology [8] in our statistical tests. With this method the null hypothesis H_0 stated that both retrieval schemes produced similar MRR (or MAP) performance, and the null hypothesis would be accepted if two retrieval schemes returned statistically similar retrieval performance, otherwise it would be rejected. In this study our experiments detected statistical significant differences by applying a two-sided non-parametric bootstrap test (significance level $\alpha = 5\%$).

Evaluation of Monolingual and English to French Searches

To define a baseline, we tested three IR models by submitting queries to search our corpus using the 299 topics written in the French language. The resulting MRR for topic titles only are depicted in the second column of Table 2 (labeled "Monolingual") and the corresponding MAP in the fourth column. We then took this value as a baseline and compared its retrieval effectiveness with other search models, while applying the same conditions. For both MRR and MAP, the Okapi model always provided the best retrieval results, and these results were significantly better than that of other search approaches.

In a second experiment, we took the English language topics and had them translated into French using Google's translation service, and then searched the French corpus with the translated topics. Through applying these three IR models, our MRR evaluations produced the results shown in the third column of Table 2 (labeled "From EN") or in the fifth column when using the MAP as retrieval effectiveness measures. In all cases, the Okapi approach performed significantly better than did the two other IR models.

When comparing original with translated topics, the performances decreased due to the automatic translation process. For the MRR, this difference was around 12% when using the Okapi search model (0.6631 vs. 0.5817) while with the MAP, this difference was slightly larger (0.4008 vs. 0.3408, or -15% in relative value). Taking the column labeled "Monolingual" as the baseline, retrieval performance differences for the

translated queries are always statistically significant for both the MRR or the MAP, and for all three retrieval models.

	MRR		MAP	
	Monolingual	From EN	Monolingual	From EN
Okapi	**0.6631**	**0.5817**	**0.4008**	**0.3408**
Language Model	0.5948	0.5093	0.3647	0.3085
$tf \cdot idf$	0.5072	0.3895	0.2591	0.2091

Table 2. Mean reciprocal rank (MRR) and mean average precision (MAP) for both monolingual and bilingual searches (299 title-only queries)

Although we know that the mean is a useful method for representing an entire distribution of observations, it may hide certain underlying irregularities. An inspection of the MRR performance obtained using the Okapi model for monolingual queries shows that out of 299 cases, 166 (55.5%) ranked the first relevant document highest, while for English queries this value was lower (142 queries or 47.5%). Second, a count of the number of queries ranking a good response among the top five shows that there were 241 monolingual vs. 213 English queries. A count of the number of hard queries (those having no relevant document ranked among the top twenty) shows that when comparing monolingual 30 vs. 60 with English queries, there was a relatively large difference. Clearly the automatic translation was not perfect and thus the retrieval quality had been decreased.

The good news was that when using the Google's translation tool to search a French corpus based on English queries, the performance difference was not large (-12%) when compared to the original French queries. There are several possible explanations for this finding. First, the two languages are related with many words have similar meanings and some even the same spelling (e.g., "soldiers" and "soldats", "success" and "succès", "quota", "immigration" etc.). Proper names also have comparable spellings (e.g., "Clinton", "Israel", "Airbus", "Bosnia" vs. "Bosnie", "Iraq" vs. "Irak", "Alps" vs. "Alpes"). As an extreme example, Topic #280 appears the same in both languages ("Crime in New York" and "Crimes à New-York"). Secondly, acronyms tend to be well translated by Google (e.g., "UN" into "ONU", "EU" into "UE", "US" into "USA"). In certain cases English topics even improved the RR performance, such as with Topic #117 "European Parliament Elections" which is translated as "Élections du Parlement européen", while the original form is "Elections parlementaires européennes". This latter version is more readable in French but includes two adjectives and only one noun ("élections"). For this query the IR system did not choose the same stem for the noun "parlement" and the adjective "parlementaires" and thus the translated query provided better retrieval performance.

Generally speaking a translated topic does not perform as well as the corresponding original French topic, and based on our experiments with the Google's translation service, there are three main reasons for this. First a word's semantic coverage may differ from one language to the other. For example, in Topic # 113 "European Cup", the word "cup" was translated into the French "tasse" (in the sense of "coffee cup") instead of "coupe" (the winner's trophy). As another example, the word "court" in

Topic #75 "Euskirchen Court Massacre" could be translated into "tribunal" or "cour" in French. For this search the most efficient term was "tribunal", which in French is used more frequently than "cour". These examples demonstrate that Google tends to provide the same translation, regardless of the context. As another example, if we ask Google to translate "the ink is in the pen" or "the pig is in the pen", the term "pen" would always be translated into French as "stylo", an instrument for writing.

Second, Google is case sensitive and thus it distinguishes between uppercase and lowercase. For example a request for "made in turkey" and "Made in Turkey" would not return the same results when translated into French. In the first case Google selects the animal and in the second the country name. In some topics however Google may incorrectly tag certain terms beginning with an uppercase letter. With Topic #192 "Russian TV Director Murder" for example, the system assumes "Murder" is a personal name and thus does not translate it into French ("Directeur russe Murder de TV" vs. "Assassinat d'un directeur de la télévision russe"). The fact that words appearing in topic titles beginning with an uppercase letter may thus induce error into the translation system, causing it to wrongly assume that a proper name is present. A similar case occurs with Topic #244 "Footballer of the Year 1994" in which the term "Footballer is tagged as a proper name, or as a word not appearing in the dictionary. In this case therefore the translation into French contains a spelling error.

Third, when idioms or other compound terms are written with a hyphen, Google and other automatic translation tools tend to produce a word-by-word translation. With Topic #261 "Fortune-telling" for example the proposed translation "Fortune-dire" (with to tell = "dire") is far from being the correct translation ("Diseurs de bonne aventure"). Again, in the case of certain idiomatic expressions (e.g., "from the horse's mouth"), incorrect translations could occur when using Google or other automatic translation tools.

Using Other Translation Resources

The evaluations and explanations mentioned above are limited to the Google translation service and also to very short query formulations pertaining to a limited number of topic titles. In fact, during the last few years other freely available machine-based translation services have become available. We thus decided to compare performances achieved by the Google translation service (limited to the Okapi model), with the alternative translation systems Babelfish and Promt, when automatically translating English topics into French. The resulting MRR values are listed in Table 3 and display a larger query construction. This combination includes the title and descriptive (TD) sections of the topic formulation, mandatory during the CLEF evaluation campaigns [3]. Although the title is sometimes ambiguous, the descriptive part may help the translation system by providing a complete sentence and context, both being useful in the automatic translation process. For example, Topic #91 is titled "AI in Latin America" and its descriptive section consists of the following "Amnesty International reports on human rights in Latin America". This description indicates that the acronym AI does not mean "Artificial Intelligence". Adding the descriptive part increases the

mean query length to 10.78 content-bearing terms, when with the title section is limited to 2.86 content-bearing terms.

	T Query	TD Query
Monolingual	**0.6631**	**0.7360**
Google	0.5817	0.6551
Babelfish	0.5653	0.6426
Promt	0.5704	0.6457

Table 3. Mean reciprocal rank (MRR) for title (T) and title & descriptive (TD) topics using monolingual and bilingual searches (Okapi, 299 queries)

The data in Table 3 shows that the performance difference between the three translation tools are small, around 1% to 3%. For example, using the title-only topics the Google translation system produces an MRR of 0.5817 vs. 0.5704, or -1.9% in relative value for the Prompt system. Using the performance obtained by Google as baseline, we did not find any statistically significant difference when compared to other translation resources. Note however that the performance difference between the monolingual (second row in Table 3) and the three query translation approaches are always statistically significant and in favor of the monolingual search. As mentioned previously, we knew that both the Babelfish and Google systems are based on the same translation technology. When inspecting the MRR achieved by the title-only query formulation, we found that performances were different for only 27 queries out of 299 when comparing the Google and Babelfish translation services. When comparing the Promt and Google translated queries, the retrieval performance was different for 117 queries.

Evaluation of German to French Searches

We decided that the previous findings should be compared to another language, and thus we selected German for the query source language. Using the Google translation tool we automatically translated the queries into French. As shown in Table 4 under the column labeled "From DE", when compared to monolingual searches retrieval performances were shown to decrease significantly. In mean, the relative difference was around 30%, and there was a statistically significant performance difference between queries written in German and those written in French.

	Monolingual	From EN	From DE	From DE-EN
Okapi	0.6631	0.5817	0.4631	0.5273
Difference %		-12.3%	-30.2%	-20.5%

Table 4. Mean reciprocal rank (MRR) for both monolingual and bilingual searches (Title-only queries)

An inspection of the Google translation results for German shows that poor retrieval performances are for the most part caused by the factors cited above, and also by the inadequate processing of German compound words. Such linguistic constructions also occur in English (e.g., viewpoint, handgun) but in German they are more frequent, and also occur in various forms (e.g., "Friedensnobelpreis" = "Frieden" (peace) + "Nobel" + "Preis" (prize) or "Nobelpreis für den Frieden"). The fact that many German compound words were not translated had a very real impact on retrieval performance. For the topics written in French, we found that only 16 queries without having a correct answer ranked among the top 50 retrieved items while for German this value increased to 61.

As a final experiment, we used the queries written in German and then automatically translated them into English, and from this pivot language we translated them into French. This evaluation thus reflects commonly occurring situations in which one language is defined as a pivot language (*interlingua*) and serves as an intermediary between all possible language pairs. There are several advantages to using this translation strategy. For direct translations, n languages would require $n \cdot (n\text{-}1)$ possible translation services. In the European Union with its 23 official languages, this means that $23 \cdot 22 = 506$ possibilities would have to be covered. Thus, instead of a direct translation for all possible language pairs we can limit the resources to $2 \cdot (n\text{-}1)$ translation pairs (or 44 in our European example), namely $(n\text{-}1)$ from all languages to the pivot language, and $(n\text{-}1)$ from the pivot language to all the others.

As shown in Table 4, with the Okapi model the retrieval performance obtained was 0.5273, resulting in a mean performance significantly lower than that of the monolingual search (0.6631) but higher than the direct translation from German (0.4631). In an effort to explain this better performance when English was selected as the pivot language, we found that translation from German to English was better than from German to French. For example, Topic #235 "Seal-hunting" is written as a compound in German ("Robbenjagd" = "Robben"(seals) + "Jagd" (hunting)) which is correctly translated into English ("Seal hunting") but not into French ("Robbenjagd"). These experiments therefore demonstrate that query translation may be effective for some language pairs yet with other language pairs certain problems may be encountered, even when using the same translation system. Moreover, compared to direct translation, the pivot language approach does not always imply less effective translation performance.

Conclusion

Writing a topic in another language and then asking Google to automatically translate it before launching a search degrades retrieval effectiveness, compared to a monolingual search in which requests and documents are written in the same language. As revealed in our evaluations based on short topic formulations, retrieval performance reductions are not always impressive (see Table 4). Applying the Google translation tool to automatically translate an English topic into French may achieve retrieval effectiveness of around 88% compared to a corresponding monolingual search. From another perspective, a monolingual search provides at least one relevant item among the first five retrieved items for 241 queries out of 299 (or 80.6%). Using the English topics and

using Google to translate them into French will place a relevant item in the top five for 213 queries (or 71.2%). Clearly, in mean, a translated query may retrieve the needed information.

Using another translation service should allow us to obtain similar retrieval performance. For example, adopting the Babelfish that Yahoo! uses, 206 queries (or 68.9%) would find at least one good answer ranked among the top five, while for the Promt translation tool this number would be 212 (or 70.9%). Changing the language pairs may however degrade retrieval effectiveness. For example, using topics written in German instead of English clearly hinders retrieval performance by around 30% compared to a monolingual search (see Table 4). An inspection of the first five retrieved items among the German topics automatically translated into French shows that at least one pertinent item would be retrieved from only 174 queries out of 299 (or 58.2%). For some language pairs, the mean result obtained is around 10% lower than that of a monolingual search while for other pairs, the retrieval performance is clearly lower. In this study, we have investigated three important languages from an economic point of view, but automatic translation resources are not available for all language pairs, particularly for languages used by small numbers of users and having only modest economic importance.

For all search systems there are difficult queries for which the search engine encounters difficulties to find at least one relevant answer. These queries typically contain concepts expressed in an ambiguous way or use vocabulary that leads to incorrect identification of relevant and non-relevant items, and when adding a translation stage this phenomenon seems to increase. In our experiments for example we found 30 title-only queries for which a monolingual search was not able to extract any relevant items in the first 20 responses. With English topics and the Google translation system however this number increased to 60. Through making use of other freely available translation services, we obtained similar results (56 queries with Promt or 64 with Babelfish).

References

[1] Harman, D.K. (2005). The TREC ad hoc experiments. In TREC, Experiment and Evaluation in Information Retrieval (pp. 79-97). Cambridge: The MIT Press. See the web site http://trec.nist.gov

[2] Noriko, K., (Ed) (2007). NTCIR Workshop 6 Meeting. Tokyo: National Institute of Informatics. See the web site http://research.nii.ac.jp/ntcir/

[3] Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., & Stempfhuber, M. (Eds) (2007). Evaluation of multilingual and multi-modal information retrieval. Lecture Notes in Computer Science #4730. Berlin: Spinger-Verlag. See the web site http://www.clef-campaign.org

[4] Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern Information Retrieval. New York: The ACM Press.

[5] Robertson, S.E., Walker, S., & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. Information Processing & Management, 36(1), 95-108.

[6] Hiemstra, D. (2002). Term-specific smoothing for the language modeling approach to information retrieval. In Proceedings of the ACM-SIGIR (pp. 35-41). New York: The ACM Press.

[7] Buckley, C., & Voorhees, E. (2005). Retrieval system evaluation. In TREC, Experiment and Evaluation in Information Retrieval (pp. 53-75). Cambridge: The MIT Press
[8] Savoy, J. (1997). Statistical inference in retrieval effectiveness evaluation. Information Processing & Management, 33(4), 495-512.

JACQUES SAVOY (Jacques.Savoy@unine.ch) is a professor in the Computer Science Department at the University of Neuchatel (Switzerland).

LJILJANA DOLAMIC (Ljiljana.Dolamic@unine.ch) is Ph.D. student in the Computer Science Department at the University of Neuchatel (Switzerland).

Retrieval Effectiveness of Machine Translated Queries

Ljiljana Dolamic, Jacques Savoy

Computer Science Dept., University of Neuchatel,
Rue Emile Argand 11, 2009 Neuchâtel, Switzerland
{Ljiljana.Dolamic, Jacques.Savoy}@unine.ch

Abstract. This paper describes and evaluates different IR models used to search document collections written in English with queries written in various other languages, either members of the Indo-European family (English, French, German, and Spanish) or radically different language groups such as Chinese. The evaluation is based on a rather large number of topics (around 300) and uses two commercial machine translation systems to cross the language barriers. The mean average precision (MAP) is applied in this study to measure differences in retrieval effectiveness when the query language differs from the document language. Although the performance difference is rather large for certain languages pairs, it does not mean that offering bilingual searches is not commercially viable. The reasons for difficulties incurred when searching or during translation are analyzed and some concrete examples are given.

Keywords. Cross-language information retrieval (CLIR); evaluation, automatic translation, morphology.

1 Introduction

In crossing language barriers, the English language plays a central role in facilitating communication for people speaking different languages. For example, in Europe as well as in large international organizations or companies (e.g., WTO, IBM, Novartis), the quantity of information written in English tends to be growing rapidly. Additionally, accessing information on the Web (Chung, 2008) in this language has become a real concern (news, hotel or airline reservations, government information, statistics, etc.). While some users are perfectly bilingual, others can read documents written in English but cannot formulate a query, or at least cannot provide reliable search terms in a form comparable to those found in the documents being searched. On the other hand, many documents contain non-textual information such as images, videos and statistics that do not need translation and can be understood regardless of the language involved.

If English is not the language spoken by the greatest number of persons around the world, it clearly plays a central role as an *interlingua* medium for transmitting knowledge or expressing opinions; CNN success story is an example of such increasing importance of this language. In Europe, India or Far-East, the first foreign language learned is often English. It is therefore important to provide good translation resources from other languages to English or vice-versa and also to analyze their quality.

This demand for translation resources from/to English has not been ignored by the most important commercial search engines. To permit improved searching of web pages available in English, regardless of language in which the query is written, Google launched a translation service in May 2007 that now provides two-way online translation services, mainly between English and 40 other languages (http://translate.google.com/). Over the last few years other free internet translation services have been made available. Yahoo! (http://babelfish.yahoo.com) for example offers also a freely available translation system. Thus, after more than ten years of research in this field, commercial products became freely available to Internet users.

The questions we would like to address is: "How effective is a bilingual search?" What is the "information retrieval cost" to web users who formulate requests in their own language and then search the web and find information written in English? If we compare two translation services, does their relative effectiveness depend only on the underlying IR model? Does the translation quality depend on the relationship between the source and target language, with a better quality being obtained by those languages having a close relationship with English (e.g., French, German) as opposed to Spanish or more distant languages such as Chinese?

The objective of this paper is to answer these questions. Although we will not evaluate translations *per se* we will test and analyze various IR and translation systems in terms of their ability to retrieve items written in English, based on an automatically translated query (experiments conducted in December 2008). The rest of this paper is divided as follows. Section 2 presents related works, while Section 3 depicts the main characteristics of the test collection. Section 4 briefly describes the IR models used during our experiments, while Section 5 evaluates them under different conditions and points out some of the main problems found in these automatic translation tools. A query-by-query analysis will complete this evaluation. The main findings of this paper are summarized in Section 6.

2 Translation Approaches

For a bilingual search to be effective (query expressed in one language, document retrieved in one or more other languages), we need to cross the language barrier. One way of achieving this is to assume that one language is merely a misspelled form of the other, as for example "English is French, misspelled" (Buckley *et al.*, 1998). When based on cognate matches, such an approach may work with closely related languages (and when an effective "spelling corrector" is included). However, an evaluation done by McNamee & Mayfield (2004) has shown that mean average precision varied from 9% to 27% compared to 45% achieved by a monolingual search. When compared with the performance of monolingual searches, this represents a relatively high decrease.

As a first real translation tool, various researchers suggested using machine-readable dictionaries (MRDs) (Ballesteros & Croft, 1997; Hull & Grefenstette, 1996; Hedlund *et al.*, 2004). However when employing MRDs we need to handle out-of-vocabulary (OOV) problems resulting from a dictionary's limited coverage. In a related issue, it could prove helpful to recognize proper nouns and acronyms and trans-

late them using a special dictionary (e.g., for the English-French languages we can find Putin-Poutine, UNO-ONU, SIDA-AIDS). Moreover, certain input words could be ambiguous and MRDs may suggest more than one translation (e.g., the word "bank" can take on a different meaning when used in the context of a river or a financial institution). Sometimes we need to automatically transform input words into base forms (lemma) listed in the dictionary, yet both errors and semantic shifts may appear during this process (e.g., the word "saw" in "I saw a man with a saw").

As a second translation tool, we might utilize machine translation (MT) systems (Chen & Gey, 2004) that are normally easy to use. However, such devices tend to perform poorly when translating entire documents, in part because translation is a semantic-based operation. For example, following the syntactical structure of the source language does not always produce the best translation (Mel'èuk & Wanner, 2006), even for closely related languages. For example, on the road we can see the sign "slow men at work" that is translated into French as "ralentir, travaux" (slow, works). This example illustrates the need for processing the idiosyncratic transformations between the source and the target syntactic structures.

As a third possibility, we might base the translation process on a corpus-based method used in conjunction with a statistical translation model in order to identify the proper translation candidates (Nie *et al*., 1999). In this case, we would need to access the corpora in order to automatically build the data structures from which direct translations or related term generation could be obtained (Sheridan & Ballerini, 1996), using the most probable match or the best *k* matches (Braschler & Schäuble, 2001). This presumes of course that for some domain-specific language pairs parallels and/or comparable corpora would be available. For certain language pairs such corpora would clearly be more difficult to find. The performance of these statistical translation approaches would however depend on very important factors such as source quality (e.g., extracted Web sites), and size (Nie & Simard, 2002). Moreover, cultural, thematic and time differences may also play a role in the effectiveness of such approaches (Kwok *et al*., 2001). Finally, assessing translation probabilities could be problematic and may result in disappointing performance levels, particularly when a lot of query terms and their correct translations cannot be found in the aligned corpus (Hiemstra *et al*., 2001).

3 Test-Collection

In an effort to promote information retrieval (IR) in languages other than English and to evaluate bilingual searches (queries expressed in one language, documents in others), various evaluation campaigns have been conducted over the last years such as TREC (Harman, 2005), NTCIR or CLEF (Peters *et al*., 2007).

To evaluate the retrieval effectiveness of bilingual searches, involving topic description written in various languages in order to retrieve document written in English, we used a corpus created during various the CLEF campaigns. This collection is made up of articles published in 1994 in the newspaper *Los Angeles Times*, as well as documents extracted from the *Glasgow Herald* and published in 1995. This collection contains a total of 169,477 documents (or about 579 MB of data). On average each

article contains about 250 (median: 191) content-bearing terms (not counting commonly occurring words such as "the," "of" or "in"). Typically, documents in this collection are represented by a short title plus one to four paragraphs of text.

	2001	2002	2003	2004	2005	2006
Source	*LA Times*	*LA Times*	*LA Times* *Glasgow H.*	*Glasgow H.*	*LA Times* *Glasgow H.*	*LA Times* *Glasgow H.*
Size	425 MB	425 MB	579 MB	154 MB	579 MB	579 MB
# docs	113,005	113,005	169,477	56,472	169,477	169,477
# topics	47	42	54	42	50	49
Topics	#41 - #90	#91 - #140	#141 - #200	#201 - #250	#251 - #300	#301 - #350

Table 1. General statistics on our test-collection for each year

This collection contains 310 topics, each subdivided into a brief title (denoted as T), a full statement of the information need (called description or D), plus any background information that might help assess the topic (narrative or N). The topic titles consist of 2 or 3 words (before stopword removal) reflecting typical web requests, and are represented by a set of keywords beginning with capitals rather than a complete grammatical phrase. These topics cover various subjects (e.g., "El Niño and the Weather," "Chinese Currency Devaluation," "Eurofighter," "Victories of Alberto Tomba," "Marriage Jackson-Presley" or "Computer Animation"), along with both regional ("Films Set in Scotland," "Area of Kaliningrad") and international coverage ("Oil Prices," "Sex in Advertisements").

Relevance judgments (correct answers) were supplied by human assessors throughout the various CLEF evaluation campaigns. As shown in Table 1, the entire corpus was not used during all the evaluation campaigns and thus pertinent articles must be searched in different parts of the corpus. For example, Topics #201 to #250 were created in 2004 and responses resulting from searches in the *Glasgow Herald* (1995) collection, a subset representing 56,472 documents. Of the 50 queries originally available in 2004, we found that only 42 had at least one correct answer.

In all, 26 queries were removed because they do not have any relevant documents in the corpus, meaning only 284 (310 minus 26) topics were used in our evaluation. Upon an inspection of these relevance assessments, the average number of correct responses for each topic was 22.46 (standard deviation: 28.9, median: 11.5), with Topic #254 ("Earthquake Damage") obtaining the greatest number of relevant documents (229).

The topics were manually translated in different languages and in this study we will use the German, French, Spanish, and simplified Chinese topic descriptions. These topics were encoded in ISO-8859-1 for the European languages, and in GB2312 for the Chinese language.

4 IR Models

In order to obtain a broader view of the relative merit of the various retrieval models, we used one vector-space scheme and three probabilistic models. First we adopted the classical *tf idf* model, wherein the weight attached to each indexing term was the product of its term occurrence frequency (or tf_{ij} for indexing term t_j in document d_i) and its inverse document frequency (or idf_j). To measure similarities between documents and requests, we computed the inner product after normalizing (cosine) the indexing weights (Manning *et al.*, 2008).

In addition to this classical vector-space scheme, we also considered probabilistic models such as that of Okapi (or BM25) (Robertson *et al.*, 2000) with parameters K and b being set to 1.2 and 0.55 respectively, which offers a high retrieval effectiveness. As a second probabilistic approach we implemented the $I(n_e)C2$ model, taken from the *Divergence from Randomness* (DFR) framework (Amati & van Rijsbergen, 2002) wherein the two information measures formulated below are combined:

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = -\log_2[Prob^1_{ij}] \cdot (1 - Prob^2_{ij}) \quad (1)$$

where $Prob^1_{ij}$ is the pure chance probability of finding tf_{ij} occurrences of the term t_j in a document. On the other hand, $Prob^2_{ij}$ is the probability of encountering a new occurrence of term t_j in the document, given tf_{ij} occurrences of this term had already been found. The $I(n_e)C2$ model was based on the following formulae:

$$Inf^1_{ij} = tfn_{ij} \cdot \log_2[(n+1) / (n_e+0.5)] \quad \text{with } n_e = n \cdot [1 - [(n-1)/n]^{tc_j}] \quad (2)$$

$$\text{with } n_e = n \cdot [1 - [(n-1)/n]^{tc_j}] \text{ and } tfn_{ij} = tf_{ij} \cdot \ln[1 + ((c \cdot mean\ dl)/l_i)]$$

$$Prob^2_{ij} = 1 - [(tc_j + 1) / (df_j \cdot (tfn_{ij}+1))] \quad (3)$$

where tc_j is the number of occurrences of term t_j in the collection, df_j indicates the number of documents in with the term t_j occurs, n the number of documents in the corpus, l_i the length of document d_i, *mean dl* (= 271), the average document length, and c a constant (fixed at 1.5).

Finally, we also considered an approach based on a language model (LM) (Hiemstra, 2000), known as a non-parametric probabilistic model (the Okapi and DFR are viewed as parametric models). Probability estimates would thus not be based on any known distribution, but rather estimated directly and based on occurrence frequencies in document d_i or the entire C corpus. Within this language model paradigm, various implementations and smoothing methods might also be considered, and in this study we adopted a model proposed by Hiemstra (2000) as described in Equation 4, which combines an estimate based on document ($P[t_j | d_i]$) and corpus ($P[t_j | C]$).

$$P[d_i | q] = P[d_i] \cdot \prod_{t_j \in Q} [\lambda_j \cdot P[t_j | d_i] + (1-\lambda_j) \cdot P[t_j | C]]$$

$$\text{with } P[t_j | d_i] = tf_{ij}/l_i \text{ and } P[t_j | C] = df_j/lc \quad \text{with } lc = \sum_k df_k \quad (4)$$

where λ_j is a smoothing factor (fixed at 0.35 for all indexing terms t_j), df_j indicates the number of documents indexed with the term t_j, and lc is a constant related to the underlying corpus C.

5 Evaluation

Given the relatively large number of queries (284), we were in a position to evaluate the retrieval effectiveness of various query languages when searching a corpus written in English. In order to achieve this, we wanted "the truth, the whole truth (recall), and nothing but the truth (precision)". To complete such an overall evaluation based on both the precision and recall, and to obtain a better understanding of the effect of a given search strategy, we analyzed the retrieval performance of some queries. Such a query-by-query inspection would provide detailed information on the drawbacks of the underlying IR scheme.

To evaluate the retrieval effectiveness of different IR schemes, we adopted the mean average precision (MAP) computed by the `trec_eval` software (based on a maximum of 1,000 retrieved records) (Buckley & Voorhees, 2005). This performance measure is the mean of all the average precision achieved by each query on the test-collection. Finally, we need to statistically determine whether or not a given search strategy would be better than another. To achieve this objective, we applied the non-parametric bootstrap test (Savoy, 1997). In our statistical tests, the null hypothesis H_0 stated that the two retrieval schemes used in the comparison produce similar retrieval performance. Thus, in the experiments presented in this paper, statistically significant differences were detected by a two-sided test (significance level 5%), and the corresponding computations were done using the R language.

The rest of this section is divided as follows. Section 5.1 presents the retrieval effectiveness of various IR models in a monolingual context that will be used as baseline in our further experiments. Section 5.3 evaluates these IR models using different query languages. Section 5.4 describes a query-by-query analysis revealing the main translation difficulties.

5.1 Monolingual Evaluation

To define a baseline, we tested four IR models using the topics written in the English language (monolingual search). In our experiments, we considered only the topic titles (T, mean number of search terms = 2.8, median: 3. min: 1, max: 7, standard deviation: 0.86). Such short topic formulations more closely reflect the current practice of web users sending their requests to commercial search engines. In this case, the mean number of search terms per request was estimated as 2.4 (Jansen & Spink, 2006). The resulting MAP is shown in Table 2. These performance values will be our baseline when evaluating bilingual searches.

As shown in Table 2, the I(n_e)C2 model provided the best retrieval results, and these results were statistically significant and better than those achieved by either the LM or *tf idf* approaches (as denoted by an "*"). On the other hand, the performance difference between the Okapi and I(n_e)C2 cannot be viewed as significant.

IR Model	MAP
I(n_e)C2	**0.4053**
Okapi	0.4044
Language Model	0.3708*
tf · idf	0.2392*

Table 2. Mean average precision (MAP) for monolingual searches (284 Title-only queries)

5.2 Bilingual Evaluation

In bilingual searches, the documents are written in one language (in English in our case) while the topics are written in another language. In order obtain a broader view point, we selected four different topic languages, two from the Latin family (French (FR) and Spanish (SP)), and the German (DE), in order to have another language related to English. To include a language with a very different morphology and writing system, we chose the Chinese (ZH) language.

From the title descriptions available in these languages, we used the Google translation service to automatically translate them into English, and these translations were then used to search our newspapers' corpus (translations done in December 2008). The resulting MAP are listed in Table 3, while Table 4 shows the same retrieval measures obtained by Yahoo's translation service.

	Mono	From ZH	From DE	From FR	From SP
I(n_e)C2	**0.4053**	**0.3340***	0.3618*	**0.3719***	0.3741*
Okapi	0.4044	0.3327*	**0.3625***	0.3692*	**0.3752***
LM	0.3708	0.3019*	0.3305*	0.3400*	0.3426*
tf · idf	0.2392	0.1920*	0.2266*	0.2294	0.2256*
Mean difference %		-18.2%	-9.3%	-7.3%	-7.1%

Table 3. Mean average precision (MAP) for both monolingual and bilingual searches using the Google translation service (284 Title-only queries)

In the last row of Tables 3 and 4 we listed the mean percentage differences with the corresponding monolingual search. The values listed show that bilingual searches always resulted in lower retrieval performance, and the differences were usually statistically significant (values denoted by an "*"). The only exception is the difference obtained using the *tf idf* model and the Google translation service with queries written in French. In this case performance difference between the monolingual and bilingual searches (0.2392 vs. 0.2294) is not statistically significant.

From comparing the different languages we saw that the translation from the French or Spanish language was easier for the Google system than was the Chinese language. Using the I(n_e)C2 IR model, the precision obtained for the bilingual French or Spanish language searches was 92% of the value obtained for monolingual search,

90% for German queries, and 82% for the simplified Chinese language. For Yahoo, the situation was somewhat comparable. The French language search obtained the best precision (82% for the monolingual search) and Chinese was the most difficult (only 55% for the monolingual search).

	Mono	From ZH	From DE	From FR	From SP
$I(n_e)C2$	**0.4053**	**0.2286***†	**0.2951***†	**0.3322***†	**0.2897***†
Okapi	0.4044	0.2245*†	0.2917*†	0.3268*†	0.2867*†
LM	0.3708	0.2000*†	0.2636*†	0.3006*†	0.2600*†
$tf \cdot idf$	0.2392	0.1289*†	0.1846*†	0.2065*†	0.1812*†
Mean difference %		-45.1%	-26.7%	-17.5%	-27.9%

Table 4. Mean average precision (MAP) for both monolingual and bilingual searches using the Yahoo translation service (284 Title-only queries)

Moreover, the translations produced by Yahoo seemed on average to encounter more problems compared to those obtained by the Google system. To verify this, we used the Google's translation service retrieval performance as baseline (Table 3) for comparing those obtained by the Yahoo system (Table 4). The performance differences between these two translation devices were always analyzed to be statistically significant (denoted by an "†" in Table 4).

5.3 Query Translation Difficulties

In order to discover the reasons why translation failed for some searches, we analyzed the retrieval performance of all individual queries. Our objective was also to potentially identify systematically occurring types of translation error. In order to limit our investigation somewhat, we mainly consider as problematic a decrease of more than 10% in average precision.

The first source of translation difficulties was the presence of proper names in the request. In some cases, the name did not change from one language to English (e.g., "France," or "Haiti") but usually a modification had to be made (e.g, "London" is written "Londres" in French). We encountered various topics depicting similar problems, such as Topic #94 ("Return of Solzhenitsyn") which was written as "le retour de Soljénitsyne" in French, "Retorno de Solzhenitsin" in Spanish, or "Rückkehr Solschenizyns" in German. When French or German was the query language, Yahoo's translation system was not able to return the correct English spelling for this name. It is interesting to note that when Spanish was the query language, both MT systems failed to translate this personal name correctly.

The correct translation of a name could be rendered more difficult due to the fact that a proper name may also have a specific meaning in the source language. For example, in Topic #89 ("Schneider Bankruptcy"), the name "Schneider" means also "cutter" in German and this meaning was selected by Yahoo's translation system producing the phrase "Cutter bankruptcy". Topic #43 ("El Niño and the Weather") dem-

onstrates another but related difficulty. In this case, the weather phenomenon was designated as a Spanish noun that also means "the boy". From the Spanish expression, Yahoo's translation service will return "the boy and the time", ignoring the fact that the topic contains a particular name. When selecting Chinese as query language and as shown in Table 5, both MT systems often cannot translate such proper name, leaving the Chinese word untouched or returning a weird expression (e.g., for Topic #89 "Schneider Bankruptcy" we obtained "史特加 bankruptcy" from Google and "Shi Tejia goes bankrupt" from Yahoo). Moreover, knowing that the Chinese language does not employ the same set of phonemes, the pronunciation and its resulting spelling forms did not have a bijective relationship with the English phonology. With Topic #121 for example "Edouard Balladur" Google returned "Edward Baladu" while Yahoo returned "Edward Baladoo". As another example, for Topic #208 "Sophie's World" Yahoo gave us "Su Fei world".

	Google				Yahoo			
	ZH	DE	FR	ES	ZH	DE	FR	ES
Name	21	2	1	2	37	11	3	13
Polysemy / Synonymy	16	4	11	11	27	21	23	14
Morphology	2	2	1	2	7	8	3	7
Compound		4	0	1	0	15	0	0
Other			2		6		2	19

Table 5. Translation error distribution according to source language and translation systems (284 Title-only queries)

The second main source of translation errors was the polysemy attached to a given word in the source language. More precisely, in order to find the appropriate word (or expression) in the target language (English in our case), the translation system had to take the context into account. In fact, a given word in one language can be translated by various words involving different semantics. As shown previously, in the Spanish Topic #43 "El Niño y el tiempo", the word "tiempo" could be translated as "weather" or "time", and the latter was selected by Yahoo's system. With Topic #341 ("Theft of 'The Scream'") written in French as "Vol" du 'Cri'", the French word "vol" could be translated by "flight" or "theft". In this case, the translation produced by Google was "The Flight of the 'Scream'" and that by Yahoo was "Flight of the 'Cry'."

This latter translation demonstrates another problem related to the synonymy of a given set of words. In this case, the translation system was faced with different translations but with related meanings. In our example, the French word "cri" could be translated using either "scream" or "cry". This synonymy aspect was also found in various topics with the related terms "car" and "automobile". In the original English version of the topics, the term "car" is used more frequently (five times to be precise), in the topic titles (e.g., Topics #106 "European car industry", Topics #288 "US Cars Import") and 18 times in all topic formulations. On the other hand, the term "automobile" was never used in the titles and only twice in the description part of two topics

(recall that our evaluations were only done on the topic titles). Of course the semantic relationship between two (or more) alternatives is not always that close, as illustrated by Topic 67 "Ship Collisions". In this case, Yahoo returned "Naval collisions" as the translation from Spanish.

As a third translation difference with the original English description, we found different morphologies and grammatical categories. Also, when expressing an idea and we can select different forms from the same root, (e.g., "merger," "merge" or "merging"). For example, from the original Topic #196 "Merger of Japanese Banks", the system was able to rank the first relevant item in the first position, while with the translation "Merging of Japanese Banks", the first relevant article appeared in the sixth position. The same problem occurred in Topic #165 "Golden Globes 1994" from which the retrieval system returned a relevant document and ranked it in first position. However with the translated query "Gold Globes 1994", the first relevant item only appeared in 6th position. This last example also demonstrated that using a more aggressive stemmer such as that of Porter (1980) conflating word variants into a common stem and this tended to be the most appropriate solution. In our case, with the form "golden" the IR system was able to rank a relevant item in the first position. On the other hand, Porter's stemmer was not able to conflate the forms "merger" and "merging" into the same root ("merging" is transformed into "merg" while "merger" is left untouched).

As fourth main source of translation problems, we found that compound constructions, occurring frequently in the German language were not always translated into English. For example with Topic #84 "Shark Attacks", from the German formulation we obtained "Haifischangriffe" (Google) or for Topic #105 "Bronchial asthma" we obtained "Bronchialasthma" from Yahoo. Of course, in both cases, the retrieval system was unable to find any relevant items with the German query formulation. Using the English original form, the IR system ranked a relevant item in the first position for both topics.

Other sources of translation problems can be found in the different languages. For example, the French Topic #200 ("Innondationeurs en Hollande et en Allemagne") contains a spelling error in the word "Innondationeurs" (instead of "Inondations"). The original French Topic #259 is written as "Lions d'or" (award name of the Venice Film Festival), which is incorrectly translated from "Golden Bear", the award name of the Berlin Film Festival. Even if the translation was correct, the translated query was not able to find any relevant item in the top 10.

Our classification errors are based on translated queries resulting in clear and significant retrieval performances with the original English topics. In many other cases, the translation was not perfect or was even incorrect but the MAP performance was similar or only produced a slight variation, and usually a slight degradation. For example, for the Topics #192 ("Russian TV Director Murder") the first relevant item was ranked third in a monolingual search. Yahoo returned "Assassination of a director of Russian television" from the French language, and with this formulation the IR system ranked the first relevant item in the 5th position. Using Google and French as the search language we obtained "The assassination of a head of Russian television" as the query translation. In this case the first relevant document appeared in the 30th position. With Chinese as the search language, we obtained "Russian television mur-

der charge" with Google and "The Russian television station managers murder" with Yahoo. With both MT systems, the translated search performed reasonably well compared to the original English search (the first retrieved item was relevant with Google translation, and the sixth with Yahoo instead of the third with the English language).

6 Conclusion

Compared to a monolingual search, writing a topic in another language and then asking Google or Yahoo to automatically translate it before launching a search significantly degrades retrieval effectiveness. When using the MAP as a performance measure, this finding is valid when German, French, Spanish, or Chinese is the query language and for all IR models analyzed in this study (see Tables 3 and 4).

Through an inspection of the hardest queries, we can determine whether the automatic translations really hurt the retrieval process. From the set of 284 queries, the best IR model was not able to find a correct answer in the top 10 of the 38 queries (representing 13.4% of the cases). Using the French queries and then automatically translated them with the Google system, this value increases to 57 (or 20% of the queries). With Yahoo, this number of difficult queries increases to 73 (25.7%). Looking at the results query-by-query, we see can clearly see a difference and degradation when using a query translation approach.

When querying with the Spanish language instead of the French, we can expect similar performance (Google) or a slight performance degradation (Yahoo). When using simplified Chinese as the query language however, retrieval performance decreases significantly, as does the number of queries for which no relevant items were listed among the top ten (69 with Google, 114 with Yahoo). With German as the query language, retrieval performances tend to lie between these two extremes.

Finally, we analyzed the query translations produced by the two MT systems to investigate their main difficulties, and we found four main reasons for this performance degradation (Section 5.3). Better translation of names (personal, geographical, product) and better processing of German compounds will clearly improve bilingual searches. In our opinion better matching between ambiguous terms would also further improve translation quality (e.g., the French word "temps" could be translated as "time" or "weather" depending on the context). For the moment however it is not clear how the context could be clearly taken into account when handling 2.6 terms per query, on average. The synonymy problem (e.g., film/movie, ship/boat, car/automobile) was also a source of performance variations between the original and translated queries. Finally, the choice of the most appropriate word form (even that taken from a common root) plays a role in the final retrieval performance (e.g., "merging" or "merger", "prehistorical" or "prehistoric", "golden" or "gold"). In this case however when designing a MT system delimiting the precise boundary between good and less effective query formulation can be difficult. On the other hand, it is worth noting that we were surprised to verify that frequently used acronyms were usually correctly translated (e.g., "ONU" and "UNO" or "UN") a feature that was absent a few years ago.

Acknowledgments
This research was supported in part by the Swiss NSF under Grant #200021-113273.

References

Amati, G., & van Rijsbergen, C.J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM-Transactions on Information Systems*, 20(4), 357-389.

Ballesteros, L., Croft, B.W. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings ACM SIGIR'97*, The ACM Press, 84-91.

Buckley, C., Singhal, A., Mitra, M., & Salton, G. (1996). New retrieval approaches using SMART. In *Proceedings of TREC-4*. NIST Publication #500-236, Gaithersburg (MA), 25-48.

Buckley, C., & Voorhees, E. (2005). Retrieval system evaluation. In E.M. Voorhees, D.K. Harman (Eds), *TREC, Experiment and Evaluation in Information Retrieval* (pp. 53-75). Cambridge: The MIT Press.

Braschler, M., Schäuble, P. (2001). Experiments with the Eurospider retrieval system for CLEF 2000. In *Cross-Language Information Retrieval and Evaluation*, Springer-Verlag, LNCS #2069, 140-148.

Chen, A., Gey, F.C. (2004). Multilingual information retrieval using machine translation, relevance feedback and decompounding. *IR Journal*, 7(1-2), 149-182.

Chung, W. (2008). Web searching in a multilingual world. *Communications of the ACM*, 51(5), 32-40.

Harman, D.K. (2005). The TREC ad hoc experiments. In E.M. Voorhees, D.K. Harman (Eds), *TREC, Experiment and Evaluation in Information Retrieval* (pp. 79-97). Cambridge: The MIT Press.

Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A., Järvelin, K. (2004). Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000-2002. *IR Journal*, 7(1-2), 99-119.

Hiemstra, D. (2000). *Using Language Models for Information Retrieval*. CTIT Ph.D. Thesis.

Hiemstra, D., Kraaij, W., Pohlmann, R., Westerveld, T. (2001). Translation resources, merging strategies, and relevance feedback for cross-language information retrieval. In *Cross-language information retrieval and evaluation*, Springer-Verlag, LNCS #2069, 102-115.

Hull, D., Grefenstette, G. (1996). Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings ACM SIGIR'96*, The ACM Press, 49-57.

Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine large search engine transaction logs. *Information Processing & Management*, 42(1), 248-263.

Kwok, K.L., Grunfeld L., Dinstl, N., Chan, M. (2001). TREC-9 Cross-language, Web and question-answering track experiments using PIRCS. In *Proceedings TREC-9*, NIST Publication #500-249, 417-426.

Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

McNamee, P., Mayfield, J. (2004). Character *N*-gram tokenization for European language text retrieval. *IR Journal*, 7(1-2), 73-97.

Mel'čuk, I., Wanner, L. (2006). Syntactic mismatches in machine translation. *Machine Translation*, 20, 81-138.

Nie, J.Y., Simard, M., Isabelle, P., Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proceedings ACM SIGIR '99*, The ACM Press, 74-81.

Nie, J.Y., Simard, M. (2002). Using statistical translation models for bilingual IR. In *Evaluation of cross-language information retrieval systems*, Springer-Verlag, LNCS #2406, 137-150.

Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., & Stempfhuber, M. (Eds). (2007). *Evaluation of multilingual and multi-modal information retrieval.* Lecture Notes in Computer Science #4730. Berlin: Springer-Verlag.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130-137.

Robertson, S.E., Walker, S., & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. *Information Processing & Management*, 36(1), 95-108.

Savoy, J. (1997). Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management*, 33(4), 495-512.

Sheridan, P., Ballerini, J.P. (1996). Experiments in multilingual information retrieval using SPIDER system. In *Proceedings ACM SIGIR'96*, The ACM Press, 58-65.

Appendix B

Documents Examples

B.1 Example of a document included in the Czech test-collection

<DOC>

<DOCID> LN-20020129005 </DOCID>

<DOCNO> LN-20020129005 </DOCNO>

<DATE> 01/29/02 </DATE>

<TITLE> Spisovatelka Astrid Lindgrenová (94 let) </TITLE>

<GEOGRAPHY> STOCKHOLM </GEOGRAPHY>

<TEXT> Autorka proslulé dětské knihy Pipi Dlouhá punčocha zemřela včera ve svém stockholmském bytě. Dožila se 94 let. Po několikadenním boji podlehla blíže nespecifikované nemoci. Podle spisovatelčiny přítelkyně Margarety Strömstedtové strávila Astrid Lindgrenová poslední dny obklopena svými blízkými. Nejslavnějším, dodnes populárním románem Lindgrenové je Pipi Dlouhá punčocha. Příběh o samorostlé dívce šokující vrstevníky i dospělé nekonvenčními názory a chováním vyšel poprvé v roce 1945. Lindgrenová napsala více než stovku literárních děl, mezi nimiž jsou kromě románů a povídek také divadelní hry, básně a písňové texty. Ve světě se dosud prodalo více než 130 milionů výtisků jejích knih. STOCKHOLM - Autorka proslulé dětské knihy Pipi Dlouhá punčocha zemřela včera ve svém stockholmském bytě. Dožila se 94 let. Po několikadenním boji podlehla blíže nespecifikované nemoci. Podle spisovatelčiny přítelkyně Margarety Strömstedtové strávila Astrid Lindgrenová poslední dny obklopena svými blízkými. Nejslavnějším, dodnes populárním románem Lindgrenové je Pipi Dlouhá punčocha. Příběh o samorostlé dívce šokující vrstevníky i dospělé nekonvenčními názory a chováním vyšel poprvé v roce 1945. Lindgrenová napsala více než stovku literárních děl, mezi nimiž jsou kromě románů a povídek také divadelní hry, básně a písňové texty. Ve světě se dosud prodalo více než 130 milionů výtisků jejích knih. </TEXT>

</DOC>

B.2 Example of a document written in the Russian language in the GIRT corpus

<DOC>

<DOCNO> ISISS-RAS-ECOSOC-20060324-45864 </DOCNO>

<DOCID> ISISS-RAS-ECOSOC-20060324-45864 </DOCID>

<TITLE-RU> Материалы выездного совещания Министерства образования и науки Российской Федерации в Уральском федеральном округе 18-19 марта 2004 года "О ходе реализации молодежной политики в Российской Федерации. Студенчество России: задачи, перспективы развития" Правительство Ханты-Манс. авт. округа - Югры. Ком. по молодеж. Политике </TITLE-RU>

<KEYWORDS-RU> конференции; молодежная политика; студенты; Россия </KEYWORDS-RU>

</DOC>

B.3 Example of a document written in the Hindi language (FIRE collection)

<Doc>

<DocNo> range20_save202_d00000_f00292 </DocNo>

<Text> मुख्य पृष्ठ राष्ट्रीय अंतरराष्ट्रीय खेल वाण्ज्याि संपादकीय दृष्टिकोण्ा फीचर राज्यवार खबरें पिछले अंक समाचार पंचांग 2006 धर्म मार्ग जूनियर जागरण्ा 7272 जरा हट के ग्रीटिंग्स साहित्य जागरण्ा रेडियो इर्-पेपर जागरण्ा सखी अंापकी बात जागरण्ा यात्रंाा जागरण्ा क्रिकेट सिने मजा मेरा जागरण्ा वर्गीकृत जागरण्ा जोशंा राश्फाि्ल चर्चा में खाना खजाना जागरण्ा विश्षंाेा जागरण्ा इमेज गुदगुदी मंडियां रिजल्ट जागरण्ा गेम्स जागरण्ा फैश्ना ' जागरण्ा ---> समाचार ---> अंतरराष्ट्रीय प्रिंट करें स्थगित हो सकती है नेपाल श्ंाांति वार्ता काठमांडो। नेपाल के प्रधानमंत्रंाी गिरिजा प्रसाद कोइराला अंौर विद्रोही माअेावादी नेताअंाे के बीच तय कार्यक्रम के मुताबिक इस सप्ताह होने वाली श्ंाांति वार्ता के दूसरे दौर की बातचीत तैयारियो के अभाव के चलते स्थगित हो सकती है। नेपाल के दो उप प्रधानमंत्र्याेंि मे से एक अमिक श्रोचान ने कहा कि उच्च स्तरीय वार्ता को तयश्दाुंा कार्यक्रम के मुताबिक श्क्राुवार को ही कराने के लिए हम पूर्व तैयारियो मे जुटे हैं, लेकिन तैयारियो के अभाव के चलते यह कुछ दिन के लिए स्थगित हो सकती है। एक अंौर मंत्रंाी ने पहचान जाहिर न करने की श्र्ता पर बताया कि श्क्राुवार को होने वाली वार्ता की स्थगित होने की संभावना है, क्योकि सरकारी दल को तैयारियो के लिए अंौर वक्त चाहिए। "इस समाचार पर अपनी प्रतिक्रिया देने के लिए यहां "क्लिक " करें या पर मेल करे। " [' कुल समाचार : 38 'भारत से सुधरते रिश्तों के बीच नहीं अंाएगा सीमा मुद्दा' स्थगित हो सकती है नेपाल श्ंाांति वार्ता दिल्ली से कोइर् गिरफ्तारी नहीं: एटीएस टीअंारएस ने दी संप्रग सरकार छोड़ने की धमकी सरकार को संसद मे अविश्वास प्रस्ताव पेश्ंा होने की अंाश्काुं अपहरण्कार्ताअंाेें से समझ्ताौेे के खिलाफ थे जसवंत ...अब मुसलिम सिविल जजों की गण्नाुं होगी केंद्र ने की महत्वपूर्ण्ा स्थानों की सुरक्ष्ंाा समीक्ष्ंाा राजस्थान में छह हिरण्ंाांेें का श्कंाि एम्स ने रचा रोबोट सर्जरी का नया इतिहास पृष्ठ </Text>

</Doc>

B.4 Example of a document written in the Marathi language (FIRE collection)

<doc>

<docno> 863332.cms.txt </docno>

<text> महामुंबई मंत्रिपद आणि सत्तेत वाटा देण्याच्या ग्वाहीमुळे रिपब्लिकन पक्ष अखेर काँग्रेस आघाडीत [Saturday, September 25, 2004 03:17:22 pm] म. टा. प्रतिनिधी मुंबई : विधानसभेच्या चार जागा, सत्तेत दहा टक्के वाटा व किमान एक मंत्रीपद देण्याचे आश्वासन मिळाल्यानंतर रिपब्लिकन पक्षाने काँग्रेस आघाडीत सहभागी होण्याचा निर्णय घेतला आहे. दरम्यान काँग्रेसप्रमाणेच राष्ट्रवादी काँग्रेसमध्येही किमान 20 ते 22 ठिकाणी बंडखोरी झाली असल्याची कबुली देतानाच बंडखोर शनिवारपर्यंत माघार घेतील, असा विश्वास प्रदेशाध्यक्ष आर. आर. पाटील यांनी व्यक्त केला. विधानसभा निवडणुकीत रामदास आठवले यांच्या नेतृत्वाखालील रिपब्लिकन पक्षही काँग्रेस व राष्ट्रवादी काँग्रेसच्या आघाडीत सहभागी होत असल्याची घोषणा शुक्रवारी तिन्ही पक्षांच्या नेत्यांच्या उपस्थितीत करण्यात आली. राष्ट्रवादी काँग्रेसच्या कार्यालयात झालेल्या वार्ताहर परिषदेला पक्षाचे प्रदेशाध्यक्ष आर. आर. पाटील यांच्याबरोबरच काँग्रेसच्या प्रदेशाध्यक्षा प्रभा राव, स्वत: रामदास आठवले तसेच उपमुख्यमंत्री विजयसिंह मोहिते पाटील, माजी उपमुख्यमंत्री छगन भुजबळ आदी उपस्थित होते. वरळी, उल्हासनगर सोडले रिपब्लिकन पक्षाबरोबर झालेल्या चर्चेत त्या पक्षाला वरळी, उल्हासनगर, इंदापूर व गंगाखेड या चार जागा सोडण्याचा निर्णय झाला आहे. याशिवाय विधान परिषदेवर तीन जागा, सत्ता आल्यानंतर किमान एक मंत्रीपद, राष्ट्रवादी काँग्रेसच्या वाट्याला येणारी महामंडळांवरील 10 टक्के जागा रिपब्लिकन पक्षाला देण्याचा निर्णय घेण्यात आला असल्याचे या युतीची घोषणा करताना पाटील यांनी सांगितले. रिपब्लिकन पक्षाने राष्ट्रवादी काँग्रेसला नेहमीच साथ केली मात्र गेल्या चार वर्षांत या पक्षाला सत्तेत वाटा देताना अन्याय झाल्याची कबुलीही पाटील यांनी यावेळी दिली. महात्मा फुले, शाहू महाराज, डॉ. बाबासाहेब आंबेडकर आणि शिवाजी महाराजांची परंपरा आम्ही मानतो, त्यांची विचारधारा अधिक बळकट व्हावी, या व्यापक हेतूने अवघ्या चार जागा वाट्याला येत असूनही काँग्रेसच्या आघाडीत सहभागी होण्याचा निर्णय रिपब्लिकन पक्षाने घेतला असल्याचे यावेळी बोलताना रामदास आठवले म्हणाले. बंडखोरांची समजूत काढणार राष्ट्रवादी काँग्रेसच्या वाट्याला आलेल्या जागांपैकी 20 ते 22 ठिकाणी एकाहून अधिक उमेदवारांनी उमेदवारी अर्ज भरले असल्याचे यावेळी आर. आर. पाटील यांनी सांगितले. या सर्वांशी संवाद साधण्यात येत आहे. सांगलीत मदन पाटील यांनी बंडखोरी केली असली तरी त्यांनीही माघार घ्यावी म्हणून प्रयत्न सुरू असल्याचे पाटील म्हणाले. </text>

</doc>

B.5 Example of a document written in the Bengali language (FIRE collection)

<DOC>

<DOCNO> 1050701_1banga5.pc.utf8 </DOCNO>

<TEXT> **16** আষাঢ় **1412** শুক্রবার **1** জুলাই **2005** বঙ্গ সম্মেলনের টুকিটাকি চার দেশের জাতীয় সঙ্গীত উদ্বোধনী অনুষ্ঠানে চারটি দেশের জাতীয় সঙ্গীত হবে ম্যাডিসন স্কোয়ার গার্ডেনে। আমেরিকা, কানাডা, ভারত ও বাংলাদেশ। মূলত এই চারটি দেশের বাঙালিরাই আসছেন বঙ্গ সম্মেলনে। পাশাপাশি থাকবে শান্তীর জন্য সংস্কৃত স্তোত্রও। উদ্যোক্তাদের আরও প্রয়াস, দেশ ও বিদেশের শিল্পীরা মিলে এত বড় মাপে এই প্রথম একই সঙ্গে মঞ্চে অনুষ্ঠান করবেন। অতিথি শিল্পীদের সঙ্গে মিলেমিশে যাবেন স্থানীয় প্রতিভারাও। ফিল্মোৎসবে বঙ্গদর্শন দু'দিনে মোট **15**টি ছবি। বঙ্গ সম্মেলনের ফিল্মোৎসব। উদ্যোক্তারা জানিয়েছেন, তার মধ্যে যেমন রয়েছে হেমেন গুপ্তের 'বিয়াল্লিশ', তেমনই আছে সত্যজিত্ রায়ের 'দেবী' **।** এক দিকে গৌতম হালদারের 'ভাল থেকো' তো অন্য দিকে রবি ওঝার 'আবার আসব ফিরে' **।** তার মধ্যেই জায়গা করে নিয়েছে হরনাথ চক্রবর্তীর 'সংগ্রাম', অনুপ সেনগুপ্তের 'মায়ের আঁচল', পিনাকী চৌধুরীর 'ক্যান্সার', গৌতম ঘোষের 'দেখা' এবং শেখর দাসের 'মহুলবনির সেরেং' **।।** হালফিলের তারকাদের মধ্যে রচনা বন্দ্যোপাধ্যায়ের থাকার কথা। থাকবেন সৌমিত্র চট্টোপাধ্যায় ও শর্মিলা ঠাকূরও। উদ্যোক্তাদের তরফে বিলি-করা অনুষ্ঠানসূচি অনুযায়ী (এই সাবধানবাণী সমেত যে, 'বিনা নোটিসে যখন তখন বদলাতে পারে), উৎসবের অনেকটা জায়গাই নিয়ে নেবে ফিল্মোৎসব। নিরাপত্তার জন্য বিশ্ববঙ্গ সম্মেলনের উদ্বোধনী অনুষ্ঠান শুরু শুক্রবার সন্ধে সাতটায়, ম্যাডিসন স্কোয়ার গার্ডেনের 'থিয়েটার'-এ। তার অন্তত দু-ঘন্টা আগেই সেখানে পৌঁছোতে হবে শিল্পীদের। নিরাপত্তার খাতীরে। কলকাতা থেকে গাইতে এসেছেন যে শিল্পীরা, তঁারা আমেরিকার প্রবাসী শিল্পীদের সঙ্গে টানা দু-দিন ধরে মহড়া দিয়েছেন কূইন্সে। গানের সঙ্গে গ্রন্থনায় আবৃত্তিশিল্পী ব্রততী বন্দ্যোপাধ্যায়। উদ্বোধনী অনুষ্ঠানে মাল্টিমিডিয়া প্রেজেন্টেশনের সঙ্গেই মঞ্চে থাকবেন দেশ-বিদেশ মিলিয়ে প্রায় **70** জন নৃত্যশিল্পী। পরের কয়েক দিন গান গাইবেন মান্না দে, শান এবং কবিতা কৃষ্ণমূর্তি। আধুনিক গানের জলসায় গাওয়ার কথা মিতালি সিংহ, শ্রীরাধা বন্দ্যোপাধ্যায়, মণিময় ভট্টাচার্য এবং রাঘব চট্টোপাধ্যায়ের। বঙ্গ বিপণন বঙ্গ সম্মেলনে বাঙালির সঙ্গে বাঙালির দেখা হওয়ার আবহের মধ্যে বাণিজ্যের হাওয়াটুকুও থাক, চাইছেন উদ্যোক্তারা। তাই সব মিলিয়ে **60**টি বুথ থাকছে এ বারের সম্মেলনে, পেনসিলভেনিয়া হোটেলের পেন প্যাভিলিয়ন এবং ম্যাডিসন স্কোয়ার গার্ডেনের টেরেসে। সেই সব বুথের মধ্যে যেমন বই এবং সাময়িকীর জন্য রয়েছে আনন্দবাজার পত্রিকা, তেমনই রয়েছে শাড়ি, এথনিক গয়না, কাটলারি, বাদ্যযন্ত্র, রিয়েল এস্টেটের বুথও। অবশ্য তার সঙ্গেই মিলেমিশে থাকছে বামফ্রন্টের চেয়ারম্যান তথা সিপিএমের পলিটব্যুরো সদস্য বিমান বসুর 'বিদ্যাসাগর মেলা'-র বুথও। ফ্লাইট মিস করায় বিমানবাবুর আসতে এক দিন দেরি হয়ে গিয়েছে। এসে গিয়েছেন দলে তাঁর দুই সহকর্মী-- সাংসদ নীলোৎপল বসু ও সুজন চক্রবর্তী। </TEXT>

</DOC>

Appendix C

Light Stemming Procedures

C.1 Light Stemming Procedure for the Russian Language

```
RussianStemmer(word) {
   RemoveCase(word);
   Normalize(word);
   return;
}
RemoveCase(word) {
   if(word ends with "-иями") then remove "-иями" return;
   if(word ends with "-оями") then remove "-оями" return;
   if(word ends with "-оиев") then remove "-оиев" return;
   if(word ends with "-иях") then remove "-иях" return;
   if(word ends with "-иям") then remove "-иям" return;
   if(word ends with "-ями") then remove "-ями" return;
   if(word ends with "-оям") then remove "-оям" return;
   if(word ends with "-оях") then remove "-оях" return;
   if(word ends with "-ами") then remove "-ами" return;
   if(word ends with "-его") then remove "-его" return;
   if(word ends with "-ему") then remove "-ему" return;
   if(word ends with "-ери") then remove "-ери" return;
   if(word ends with "-ими") then remove "-ими" return;
   if(word ends with "-иев") then remove "-иев" return;
   if(word ends with "-ого") then remove "-ого" return;
   if(word ends with "-ому") then remove "-ому" return;
   if(word ends with "-ыми") then remove "-ыми" return;
   if(word ends with "-оев") then remove "-оев" return;
   if(word ends with "-яя") then remove "-яя" return;
   if(word ends with "-ях") then remove "-ях" return;
   if(word ends with "-ею") then remove "-ею" return;
   if(word ends with "-ая") then remove "-ая" return;
   if(word ends with "-ах") then remove "-ах" return;
   if(word ends with "-ею") then remove "-ею" return;
   if(word ends with "-их") then remove "-их" return;
   if(word ends with "-ия") then remove "-ия" return;
   if(word ends with "-ию") then remove "-ию" return;
   if(word ends with "-ие") then remove "-ие" return;
   if(word ends with "-ий") then remove "-ий" return;
   if(word ends with "-им") then remove "-им" return;
   if(word ends with "-ое") then remove "-ое" return;
   if(word ends with "-ом") then remove "-ом" return;
   if(word ends with "-ой") then remove "-ой" return;
   if(word ends with "-ов") then remove "-ов" return;
   if(word ends with "-ые") then remove "-ые" return;
   if(word ends with "-ый") then remove "-ый" return;
   if(word ends with "-ым") then remove "-ым" return;
   if(word ends with "-ми") then remove "-ми" return;
   if(word ends with "-ою") then remove "-ою" return;
   if(word ends with "-ую") then remove "-ую" return;
   if(word ends with "-ям") then remove "-ям" return;
   if(word ends with "-ых") then remove "-ых" return;
```

```
   if(word ends with "-ых") then remove "-ых" return;
   if(word ends with "-ея") then remove "-ея" return;
   if(word ends with "-ам") then remove "-ам" return;
   if(word ends with "-ее") then remove "-ее" return;
   if(word ends with "-ей") then remove "-ей" return;
   if(word ends with "-ем") then remove "-ем" return;
   if(word ends with "-ев") then remove "-ев" return;
   if(word ends with "-[яюйы]") then remove "-[яюйы]"return;
   if(word ends with "-[аеиоу]") then remove "-[аеиоу]" return;
   return;
}
Normalize(word) {
   if(word ends with "-ь") then remove "-ь" return;
   if(word ends with "-и") then remove "-и"return;
   if(word ends with "-нн") then replace by "-н" return;
   return;
}
```

C.2 Light Stemming Procedure for the Hindi Language

```
HindiStemmerLight(word) {
   RemoveSuffix(word);
   return;
}
RemoveSuffix(word) {
   if (word ends with "-िया") then remove "-िया" return;
   if (word ends with "-ियो") then remove "-ियो" return;
   if (word ends with "-ाए") then remove "-ाए" return;
   if (word ends with "-ाओ") then remove "-ाओ" return;
   if (word ends with "-ुआ") then remove "-ुआ" return;
   if (word ends with "-ुओ") then remove "-ुओ" return;
   if (word ends with "-ये") then remove "-ये" return;
   if (word ends with "-ेन") then remove "-ेन" return;
   if (word ends with "-ेण") then remove "-ेण" return;
   if (word ends with "-ीय") then remove "-ीय" return;
   if (word ends with "-टी") then remove "-टी" return;
   if (word ends with "-ार") then remove "-ार" return;
   if (word ends with "-ाई") then remove "-ाई" return;
   if (word ends with "-ा") then remove "-ा" return;
   if (word ends with "-े") then remove "-े" return;
   if (word ends with "-ी") then remove "-ी" return;
   if (word ends with "-ो") then remove "-ो" return;
   if (word ends with "-ि")then remove "-ि" return ;
   if (word ends with "-अ") then remove "-अ" return;

   return;
}
```

C.3 Light Stemming Procedure for the Marathi Language

```
MarathiStemmerLight(word) {
   RemoveCase(word);
   RemoveNoGender(word);
   return;
}
RemoveCase(word) {
   if (word ends with "-शया") then remove "-शया" return;
   if (word ends with "-शे") then remove "-शे" return;
   if (word ends with "-शी") then remove "-शी" return;
   if (word ends with "-चा") then remove "-चा" return;
   if (word ends with "-ची") then remove "-ची" return;
   if (word ends with "-चे") then remove "-चे" return;
   if (word ends with "-हून") then remove "-हून" return;
   if (word ends with "-नो") then remove "-नो" return;
   if (word ends with "-तो") then remove "-तो" return;
   if (word ends with "-ने") then remove "-ने" return;
   if (word ends with "-नी") then remove "-नी" return;
   if (word ends with "-ही") then remove "-ही" return;
   if (word ends with "-ते") then remove "-ते" return;
   if (word ends with "-या") then remove "-या" return;
   if (word ends with "-ना") then remove "-ना" return;
   if (word ends with "-ून") then remove "-ून" return;
   if (word ends with "-[ेीा स ल त म]")
     then remove "-[ेीा स ल त म]" return ;
return;
   }
RemoveNoGender(word) {
   if (word ends with "-उरडा") then remove "-उरडा" return;
   if (word ends with "-ढा") then remove "-ढा" return;
   if (word ends with "-रु") then remove "-रु" return;
   if (word ends with "-डे") then remove  "-डे" return;
   if (word ends with "-ती") then remove "-ती" return;
   if (word ends with "-ान") then remove "-ान" return;
   if (word ends with "-ीण") then remove "-ीण" return;
   if (word ends with "-डा") then remove "-डा" return;
   if (word ends with "-डी") then remove "-डी" return;
   if (word ends with "-गा") then remove "-गा" return;
   if (word ends with "-ला") then remove "-ला" return;
   if (word ends with "-या") then remove "-या" return;
   if (word ends with "-वा") then remove "-वा" return;
   if (word ends with "-ये") then remove "-ये" return;
   if (word ends with "-वे") then remove "-वे" return;
   if (word ends with "-ती") then remove "-ती" return;
   if (word ends with "-[अ े ि ै ै उ ]")
     then remove "-[अ े ि ै ै उ]"return;
   if (word ends with "-[ा ी ू त]")
     then remove "-[ा ी ू त]" return;
   return;
}
```

C.4 Light Stemming Procedure for the Bengali Language

```
BengaliStemmerLight(word) {
   RemoveIO(word);
   RemoveCase(word);
   RemoveArticle(word);
   Normalize(word);
   return;
}
RemoveOI(word){
if (word ends with "-[ি া ]")
      then remove "-[ি া]"return;
   return;
}
RemoveCase(word) {
   if (word ends with "-য়েদেরকে") then remove "-য়েদেরকে" return;
   if (word ends with "-াদেরকে") then remove "-াদেরকে" return;
   if (word ends with "-দেরকে") then remove "-দেরকে" return;
   if (word ends with "-য়েদের") then remove "-য়েদের" return;
   if (word ends with "-াদের") then remove  "-াদের" return;
   if (word ends with "-য়েরা") then remove  "-য়েরা" return;
   if (word ends with "-দের") then remove "-দের" return;
   if (word ends with "-ারা") then remove "-ারা" return;
   if (word ends with "-য়ের") then remove "-য়ের" return;
   if (word ends with "-কায়") then remove "-কায়" return;
   if (word ends with "-কার") then remove "-কার" return;
   if (word ends with "-িলা") then remove "-িলা" return;
   if (word ends with "-ের") then remove "-ের" return;
   if (word ends with "-ার") then remove "-ার" return;
   if (word ends with "-ান") then remove "-ান" return;
   if (word ends with "-েন") then remove "-েন" return;
   if (word ends with "-কু") then remove "-কু" return;
   if (word ends with "-কা") then remove "-কা" return;
   if (word ends with "-বি") then remove "-বি" return;
   if (word ends with "-তে") then remove "-তে" return;
   if (word ends with "-রা") then remove "-রা" return;
   if (word ends with "-াঁয়") then remove "-াঁয়" return;
   if (word ends with "-বে") then remove "-বে" return;
   if (word ends with "-মি") then remove "-মি" return;
   if (word ends with "-সি") then remove "-সি" return;
   if (word ends with "-ড়ে") then remove "-ড়ে" return;
   if (word ends with "-কে") then replace with "ড" return;
   if (word ends with "-[ে র য় ব ম স াঁ]")
      then remove "-[ে র য় ব ম স াঁ]"return;
   return;
}
```

```
RemoveArticle(word) {
   if (word ends with "-খানা") then remove "-খানা" return;
   if (word ends with "-খানি") then remove "-খানি" return;
   if (word ends with "-গুলো") then remove "-গুলো" return;
   if (word ends with "-গুলি") then remove "-গুলি" return;
   if (word ends with "-য়োন") then remove "-য়োন" return;
   if (word ends with "-খান") then remove "-খান" return;
   if (word ends with "-গুল") then remove "-গুল" return;
   if (word ends with "-টা") then remove "-টা" return;
   if (word ends with "-টি") then remove "-টি" return;
   if (word ends with "-টু") then remove "-টু" return;
   if (word ends with "-কা") then remove "-কা" return;
   if (word ends with "-তা") then remove "-তা" return;
   if (word ends with "-িল") then remove "-িল" return;
   if (word ends with "-িক") then remove "-িক" return;
   if (word ends with "-েক") then remove "-েক" return;
   if (word ends with "-েত") then remove "-েত" return;
   if (word ends with "-লি") then remove "-লি" return;
   if (word ends with "-য়া") then remove "-য়া" return;
   if (word ends with "-ায়") then remove "-ায়" return;
   if (word ends with "-ড়ি") then remove "-ড়ি" return;
   if (word ends with "-িস") then remove "-িস" return;
   if (word ends with "-ান") then remove "-ান" return;
   if (word ends with "-[ঠ ট া য়]")
      then remove "-[ঠ ট া য়]"return;
   return;
   }
Normalize(word) {
   if (word ends with "-িনি") then remove "-িনি" return;
   if (word ends with "-নি") then remove "-নি" return;
   if (word ends with "-িন") then remove "-ো" return;
   if (word ends with "-না") then remove "-না" return;
if (word ends with "-[ো ে ন র ি]")
      then remove "-[ো ে ন র ি]"return;
   return;
}
```

Appendix D

List of Publication

to appear

- Ljiljana Dolamic, Jacques Savoy
  Comparative Study of Indexing and Search Strategies for the Hindi, Marathi and Bengali Language
  In *Special Issue of ACM TALIP on IR for Indian languages*

- Ljiljana Dolamic, Jacques Savoy
  Retrieval Effectiveness of Machine Translated Queries
  In *Journal of the American Society for Information Science*

2010

- Ljiljana Dolamic, Jacques Savoy
  Unine at FIRE 2010: Hindi, Bengali, and Marathi IR
  In *Working notes of the FIRE 2010 Workshop* Grandhinagar, India, February 19-21, 2010.

- Ljiljana Dolamic, Jacques Savoy
  When Stopword Lists Make the Difference
  In *Journal of the American Society for Information Science*, Volume 61(1), pages 200-203, 2010.

2009

- Jacques Savoy, Ljiljana Dolamic
  How Effective is Google's Translation Service in Search?
  In *Communications of the ACM*, Volume 52(10), pages 139-143, 2009.

- Ljiljana Dolamic, Claire Fautsch, Jacques Savoy
  UniNE at CLEF 2009: Persian ad hoc and CLEF-IP
  In *the Working notes for the CLEF 2009 Workshop* Corfu, Greece, September 30 - October 2, 2009.

- Ljiljana Dolamic, Jacques Savoy
  Indexing and Stemming Approaches for the Czech language
  In *Information Processing & Management*, Volume 45(6), pages 714-720, 2009.

- Ljiljana Dolamic, Jacques Savoy
  Persian Language, is Stemming Efficient?
  In *Proceedings of 6th International Workshop on Text-based Information Retrieval* August 31 - September 4 2009, Linz, Austria.

- Ljiljana Dolamic, Jacques Savoy
  Indexing and Searching Strategies for the Russian Language
  In *Journal of the American Society for Information Science*, Volume 60(12), pages 2540-2547, 2009.

- Ljiljana Dolamic, Claire Fautsch, Jacques Savoy
  UniNE at CLEF 2008: TEL and Persian IR
  In *Carol Peters et al. (Eds):Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, Volume 5706/2008, pages 178-185, Springer.

- Claire Fautsch, Ljiljana Dolamic, Jacques Savoy
  UniNE at Domain-Specific IR - CLEF 2008: Scientific Data Retrieval: Various Query Expansion Approaches
  In *Carol Peters et al. (Eds):Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 199-202, 2008, Revised Selected Papers*, Volume 5706/2008, pages 196-199, Springer.

2008

- Ljiljana Dolamic,Jacques Savoy
  Unine at FIRE 2008: Hindi, Bengali, and Marathi IR
  In *the Working notes of the FIRE 2008 Workshop*, Kolkata, India, December 12-14, 2008.

- Ljiljana Dolamic, Jacques Savoy
  Variations autour de tf idf et du moteur Lucene
  In *Proceedings of 9th International Conference of the Statistical Analysis of the Textual Data* March 12-14 2008, Lyon, France, pages 1047-1058.

- Claire Fautsch, Ljiljana Dolamic, Samir Abdou, Jacques Savoy
  Domain-Specific IR for German, English and Russian Languages
  In *Carol Peters et al. (Eds): Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, Volume 5152/2008, pages 196-199, Springer Berlin/Heidelberg.

- Ljiljana Dolamic, Jacques Savoy
  Stemming Approaches for East European Languages
  In *Carol Peters et al. (Eds): Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, Volume 5152/2008, pages 37-44, Springer Berlin/Heidelberg.

Bibliography

Amati, G. & van Rijsbergen, C. J. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM-Transactions on Information Systems*, 20(4):357–389, 2002.

Buckley, C. & Voorhees, E. M. Retrieval system evaluation. In Voorhees, E. & Harman, D., editors, TREC. *Experiment and Evaluation in Information Retrieval*, pages 53–75, Cambridge (MA), 2005. The MIT Press.

Buckley, C., Singhal, A., Mitra, M., & Salton, G. New retrieval approaches using SMART. In *Proceedings of* TREC-*4*, volume 500-236, pages 25–48, Gaithersburg (MA), 1996. NIST Publication.

Dolamic, L. & Savoy, J. Stemming Approaches for East European Languages. In *CLEF*, pages 37–44, 2007.

Dolamic, L. & Savoy, J. Persian language, is Stemming Efficient? In Tjoa, A. M. & Wagner, R., editors, *DEXA Workshops*, pages 388–392. IEEE Computer Society, 2009. ISBN 978-0-7695-3763-4.

Dolamic, L. & Savoy, J. When stopword lists make the difference. *Journal of the American Society for Information Science*, 61(1):200–203, 2010.

Fautsch, C. & Savoy, J. Stratégies de recherche dans la blogosphére. *Document Numrique*, 11:109–132, 2008.

Fautsch, C. & Savoy, J. Algorithmic stemmers or morphological analysis? An evaluation. *Journal of the American Society for Information Science and Technology*, 60(8):1616–1624, 2009.

Grefenstette, G. *Cross-Language Information Retrieval*. Kluwer Academic Publishers, Norwell, MA, 1998.

Harman, D. K. How effective is suffixing? *Journal of the American Society for Information Science*, 42(1):7–15, 1991.

Harman, D. K. Beyond English. In Voorhees, E. M. & Harman, D. K., editors, *TREC Experiment and Evaluation in Information Retrieval*, pages 153–182, Cambridge, MA, 2005. The MIT Press.

Harter, S. P. A probabilistic approach to automatic keyword indexing. *Journal of the American Society for Information Science*, 26:197–206, 1975.

Hiemstra, D. *Using Language Models for Information Retrieval.* PhD thesis, CTIT, 2000.

Hollink, V., Kamps, J., Monz, C., & de Rijke, M. Monolingual document retrieval for European languages. *Information Retrieval*, 7(1-2):33–52, 2004.

Malhebre, M. *Les langues de l'humanité.* Robert Laffont, Paris, 1995.

Manning, C., Raghavan, P., & Schtze, H. *Introduction to Information Retrieval.* Cambridge University Press, Cambridge,UK, 2008.

McNamee, P. & Mayfield, J. Character n-gram tokenization for European language text retrieval. *IR Journal*, 7(1-2):73–97, 2004.

McNamee, P., Nicholas, C., & Mayfield, J. Addressing morphological variation in alphabetic languages. In *Proceedings of the ACMSIGIR'09*, pages 75–82, New York, 2009. ACM.

Navalkar, G. R. *The Students Marathi Grammar.* Asian Education Services, New Delhi, 2001.

Oard, D. W. & Resnik, P. Support for interactive document selection in cross-language information retrieval. *Information Processing & Management*, 35(3):363–379, 1999.

Peters, C., Jijkoun, J., Mandl, T., Müller, H., Oard, D. W., Peñas, A., Petras, V., & Santos, D., editors. *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, volume 5152 of *Lecture Notes in Computer Science*, 2008. Springer.

Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G. J. F., Kurimo, M., Mandl, T., Peñas, A., & Petras, V., editors. *Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, volume 5706 of *Lecture Notes in Computer Science*, 2009. Springer.

Pirkola, A. Morphological typology of languages for IR. *Journal of Documentation*, 57 (3):330–348, 2001.

Robertson, S. E. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

Robertson, S. E., Walker, S., & Beaulieu, M. Experimentation as a way of life: Okapi at TREC. *Information Processing & Management*, 36:95–108, 2000.

Salton, G., Ed. *The* SMART *retrieval system. Experiements in automatic document processing.* Prentice-Hall Inc., Englewood Cliffs (NJ), 1971.

Salton, G. & McGill, M. *Introduction to Modern Information Retrieval.* McGraw-Hill Book Company, 1984.

Savoy, J. Combining multiple strategies for effective monolingual and cross-language retrieval. *Inf. Retr.*, 7(1-2):121–148, 2004.

Singhal, A., Choi, J., Lewis, D. D., & Pereira, F. AT&T at TREC-7. In *Proceedings* TREC -*7*, pages 239–251, Gaithersburg (MA), 1999. NIST.

Sproat, R. *Morphology and Computation.* The MIT Press, Cambrige, 1992.

Tomlinson, S. Lexical and algorithmic stemming compared for 9 European languages with Hummingbird SearchServer™ at CLEF 2003. In *Comparative Evaluation of Multilingual Information Access Systems*, volume 3237 of *Lecture Notes in Computer Science*, pages 286–300, Berlin, 2004. Springer-Verlag.

Zhai, C. & Lafferty, J. A study of smoothing methods for language models applied to information retrieval. *ACM-Transactions on Information Systems*, 22(2):179–214, 2004.