Lessons Learnt from Experiments on the Ad Hoc Multilingual Test Collections at CLEF

Jacques Savoy and Martin Braschler

Abstract This chapter describes the lessons learnt from the ad hoc track at CLEF in the years 2000 to 2009. This contribution focuses on Information Retrieval (IR) for languages other than English (monolingual IR), as well as bilingual IR (also termed "cross-lingual"; the request is written in one language and the searched collection in another), and multilingual IR (the information items are written in many different languages). During these years the ad hoc track has used mainly newspaper test collections, covering more than 15 languages. The authors themselves have designed, implemented and evaluated IR tools for all these languages during those CLEF campaigns. Based on our own experience and the lessons reported by other participants in these years, we are able to describe the most important challenges when designing a IR system for a new language. When dealing with bilingual IR, our experiments indicate that the critical point is the translation process. However, currently online translating systems tend to offer rather effective translation from one language to another, especially when one of these languages is English. In order to solve the multilingual IR question, different IR architectures are possible. For the simplest approach based on query translation of individual language pairs, the crucial component is the merging of the intermediate bilingual results. When considering both document and query translation, the complexity of the whole system represents clearly a main issue.

J. Savoy (✉)
Computer Science Department, University of Neuchâtel, Neuchâtel, Switzerland
e-mail: Jacques.Savoy@unine.ch

M. Braschler
Institut für Angewandte Informationstechnologie, Zürich University of Applied Sciences ZHAW, Winterthur, Switzerland
e-mail: bram@zhaw.ch

N. Ferro, C. Peters (eds.), *Information Retrieval Evaluation in a Changing World*, The Information Retrieval Series 41,
https://doi.org/10.1007/978-3-030-22948-1_7

1 Introduction

In the field of natural language research and applications, the English language is getting the most attention. With the growing presence of web sites not written in English, there is an increasing demand for effective tools to manipulate content in other natural languages. Such interest has also been supported by the process of globalization. Looking at the world around us, one can see many documents in digital libraries, newspapers, government archives and records, as well as legal and court decision documentation not written in English. For example, the European Union counts 24 official languages in 2017, and for each of them, effective IR tools must be designed, implemented, and evaluated. This objective corresponds to one of the main purposes of the mono-, bi-, and multilingual ad-hoc tracks in the CLEF evaluation campaigns.

At a first glance, one can think that a simple adaptation of approaches for handling English should be enough. After all, a cursory observer may assume that all European languages belong to the same Indo-European language family and stem from the same source. This assumption is not true. First, the Finnish, Hungarian and Estonian languages are members of the Uralic family, while the Maltese language is related to the Semitic group. All these languages serve as official EU languages. Second, morphology and word construction vary considerably between members of the Indo-European family, reducing the effectiveness of a simple adaptation from English. As a possible language-independent solution, one can design and implement search models based on the character *n*-grams approach (McNamee et al. 2009). Such a text representation approach was also shown effective for Chinese, Japanese or Korean languages (Savoy 2005). To reflect the language differences more closely, the current chapter describes an overview of approaches taking into account the morphology differences between the different languages. Most experiments in the CLEF ad hoc track have followed this approach.

Some of the use cases associated with accessing sources written in languages other than English and more generally in a multilingual context are as follows: in multilingual countries such as Switzerland, institutions such as the Federal Supreme Court may have to document legal cases, or parts of them, in one of the national languages (German, French, or Italian), depending on the involved parties, without providing translations into the other official languages. The information contained in these documents is still relevant for the whole country regardless of the language chosen. Also worth considering are the books and documents available in various languages in our libraries, in multinational companies or large international organizations (e.g., World Trade Organization, European Parliament or United Nations), where the typical user needs to overcome various language barriers. For example, users may write a request in one language and yet wish to retrieve documents written in one or more other languages. Frequently, users have some degree of proficiency in other languages that allows them to read documents, but not to formulate a query in that language or, at least, not to provide reliable search terms to retrieve the documents being searched. In other circumstances, monolingual users may want to retrieve documents in another language and then automatically or manually translate the texts retrieved into their own language. Finally, there are many documents in other languages containing information in non-textual formats such as images, graphics, and statistics that could be made accessible to monolingual users, based on requests written in a different language.

Based on more than a decade of experiments on developing CLIR systems, the rest of this paper is organized as follows. The next section describes the main problems when designing and implementing an IR system for a new language (monolingual IR). Section 3 discusses briefly the various solutions that can be applied to develop a bilingual system that does "cross-lingual" IR, returning information items in a language other than that used for the request. The description of different multilingual IR architectures is presented in Sect. 4 together with their advantages and drawbacks. Our main findings are summarized in the conclusion.

2 Monolingual (Non-English) Information Retrieval

The implementation of IR systems is conceptually subdivided into two major phases: the indexing and the matching phases. When moving from the English language to more (potentially all) languages, we have to re-think both phases. We start our discussion with the indexing phase, which is often implemented in the form of an indexing pipeline, where information items (and the requests) are methodically transformed into representations suitable for matching. Usually, the first step is to extract the words (tokenization). As white space characters and punctuation symbols are used to denote the word boundaries for all European languages, the tokenization to be applied for these languages does not differ fundamentally from that used for English (note, however, minor differences such as the handling of contractions, like "aujourd'hui" (today) in French). After being able to determine words, the morphology of the underlying language is of prime importance. Thus, knowing the part-of-speech (POS) of very frequent words is useful to define an appropriate stopword list as indicated in Sect. 2.1. Moreover, the word formation construction varies from one language to another. Thus, there is a real need to create an effective stemmer for each language as shown in Sect. 2.2. Finally, in Sect. 2.3 we explore the matching phase. Findings indicate that fundamental concepts used in IR weighting schemes such as term frequency (*tf*), inverse document frequency (*idf*), and length normalization are valid across all languages.

2.1 Stopword List

Information retrieval weighting schemes suffer from a drop in effectiveness if extremely frequent non-content bearing words are present. In such cases, the idf-weight that should account for global frequency of terms no longer balances the contribution to the overall score. Typically, function words (determiners, prepositions, conjunctions, pronouns, and auxiliary verbal forms) are affected. Assuming that these do not convey important meaning, they can be regrouped in a stopword list to be ignored during the indexing procedure. For all languages, the identification of determiners, prepositions (or, for some languages, postpositions), conjunctions, and pronouns does not present a real difficulty. Delimiting precisely whether an auxiliary verb form must appear or not in a stopword list is less clear. Forms such as those related to the verb "to be" and "to have" are good candidates for inclusion. For the modal verbs (e.g., can, would, should), the decision is debatable. For example, one can decide that "shall" must be included but not "can" or "must".

Reflecting the root cause of the problem (the very high occurrence count of some of these words), we can opt for a frequentist perspective instead of using POS information. In this case, a stopword list can be defined as the k most frequent words (with $k = 10$ to 500) in a given corpus or language (Fox 1990). With this strategy, some recurrent lexical words of the underlying corpus will appear in the top k most frequent words. For example with newspaper collections, very frequent words (e.g., government, president, world), names (e.g., France, Obama) or acronyms (e.g., PM, UK, GOP) will also appear in the top of the resulting ranked list. This would seem undesirable, but note that words that appear with very high frequency are in any case badly suited to discriminate documents even should they be content-bearing.

After applying one of the two previous solutions, an inspection phase must verify whether the presence of a word in a stopword list could be problematic such as, for example, with homographs (e.g., "US" can be a country or a pronoun). For example, in French the word "or" can be translated into "thus/now," or "gold" while the French word "est" can correspond to "is" or "East". This verification must not be limited to the vocabulary but must take into account some acronyms (e.g., the pronoun "who" must be separated from the acronym "WHO" (World Health Organization) due, in this case, to the fact that uppercase letters are replaced by the lowercase equivalents.

Applying a stopword list generally improves the overall mean average precision (MAP). The precise value of such improvement depends on the language and the IR model, but an relative average enhancement may vary from 11.7% (English) to 17.4% (French). However, with either a long or a rather short stopword list, the retrieval effectiveness tends to be similar (MAP difference around 1.6% for the English language, 1.2% for French (Dolamic and Savoy 2010c)).

Some commercial IR systems consider that functional words may be entered by the user (e.g., search engines on the Web) or that they can be useful to specify the meaning more closely (e.g., specialized IR systems with "vitamin A"). Therefore, the size of the stopword list can be limited to a few very frequent words. As an extreme case and for the English language, the stopword list could be limited to a single entry (the article "the") (Moulinier 2004). Since stopword elimination always implies an information loss, however small, one is advised to use robust weighting schemes that allow the use of short stopword lists (Dolamic and Savoy 2010c).

2.2 Morphological Variations

A first visual difference between an English text and a document written in another European language could be the presence of a non-Latin script such as, for example, when the Cyrillic alphabet is employed for the Russian and Bulgarian languages. Another visual distinction is often the presence of diacritics (e.g., "élite", "Äpfel" (apples), "leão" (lion)). Different linguistic functions are attached to those additional glyphs such as discriminating between singular ("Apfel") and plural form ("Äpfel"), between two possible meanings (e.g., "tâche" (task) or "tache" (mark, spot)), or specifying the pronunciation. Keeping those diacritics or replacing them with the corresponding single letter modifies marginally the mean average precision (MAP), usually not in a significant way, and not always in the same direction. Note also that in some languages, it may be permissible to skip the writing of diacritics in certain circumstances, which may lead to an uneven use throughout a textual corpus. In such cases, elimination of diacritics may be advisable (e.g., in French, diacritics are usually not written in upper-case text). In German, umlauts are replaced if the corresponding keys are not available on a keyboard (e.g., "Zürich" can be written "Zuerich").

To achieve an effective semantic matching between words appearing in the user's request and the document surrogates, the indexing procedure must ignore small variations between a word stem (e.g., friend) and the various surface forms (e.g., friends). Such morphological variations may for example reflect the word's function in a sentence (grammatical cases), the gender (masculine, feminine, neutral), and the number (singular, dual, plural). For verbs, the tense, the person, and the mode may generate additional variations. These morphological variations are marked by inflectional suffixes that must be removed to discover the word stem. Of course, one can always find some exceptions such as, for example, having a plural form not always related to the singular one (e.g., "aids", the syndrome, and "aid" for help) while some words usually appear in only one form (e.g., scissors).

The English language has a comparatively simple inflectional morphology. For example, the noun plural form is usually indicated by the "-s" suffix. To denote the plural form in Italian, the last vowel (usually "-o", "-e", or "-a" for masculine nouns, "-a" or "-e" in feminine) must be changed into "-i" or "-e". In German, the plural can be indicated by a number of suffixes or transformations (e.g., "Apfel" into "Äpfel" (apple), "Auge" into "Augen" (eye), "Bett" into "Betten" (bed)). Variations in grammatical cases (nominative, accusative, dative, etc.) may imply the presence of a suffix (as, for example, the "'s" in "Paul's book"). In German, the four grammatical cases and three genders may modify the ending of adjectives or nouns. The same is valid for other languages such as Russian (6 cases), Czech (7 cases), Finnish (15 cases) or Hungarian (17 cases). As a simple indicator to define the morphological complexity of a language, one can multiply the number of possible genders, numbers, and grammatical cases. With this measure, the Italian or French language has a complexity of 2 (genders) $\times$ 2 (numbers) $=$ 4 (no grammatical case denoted by a suffix) while the German complexity is $3 \times 2 \times 4 = 24$.

New words can also be generated by adding derivational affixes. In IR, we assume that adding a prefix will change the meaning (e.g., bicycle, disbelief) and thus only suffix removal is usually considered (e.g, friendly, friendship).

Based on our experiments, it is not always clear whether a light stemmer (removing only inflectional suffixes or part of them) or an aggressive stemmer removing both inflectional and derivational suffixes proposes the best solution. For the English language, the conservative S-stemmer (Harman 1991) removes only the plural suffix while Porter's stemmer (Porter 1980) is a more aggressive approach. Such algorithmic or rule-based stemmers ignore word meanings and tend 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). In both cases, we suggest concentrating mostly on nouns and adjectives, and ignoring most of the verbal suffixes. Usually the meaning of a sentence can be determined more precisely when focusing more on the noun phrases than on the verbs.

While stemming approaches are normally designed to work with general texts, a stemmer may also be specifically designed for a given domain (e.g., medicine) or a given document collection, such as that developed by Paik and Parai (2011) or Paik et al. (2013) which used a corpus-based approach. This stemming approach reflects the language usage more closely (including word frequencies and other co-occurrence statistics), instead of a set of morphological rules.

A study demonstrates however that using a morphological analysis both light or more aggressive stemmers tend to produce statistically similar performance for the English language (Fautsch and Savoy 2009). When the stemmed words are shown to the user, we suggest applying a light stemmer approach for which the relationship between the surface form and the transformed one is usually simple and more understandable.

Using the CLEF test collections and the Okapi IR model (Robertson et al. 2000), one can find the following retrieval improvement (MAP) with a light stemmer over a non-stemming approach: +7% with the English language (Fautsch and Savoy 2009), +11% for German (Savoy 2006), +28% for Portuguese (Savoy 2006), +34% for French (Savoy 2006), +38% for Bulgarian (Savoy 2008a), +44% for Czech (Dolamic and Savoy 2009b), +55% for Hungarian (Savoy 2008b), and +96% with the Russian language (Dolamic and Savoy 2009a). Working with a morphologically rich language presenting numerous inflectional suffixes (e.g., Hungarian (Savoy 2008b)), even for names (e.g., Czech (Dolamic and Savoy 2009b); Russian (Dolamic and Savoy 2009a)), the presence of a stemming procedure is mandatory to achieve good retrieval effectiveness. Such IR tools are freely available for many languages.[1]

The choice between a light or a more aggressive suffix-stripping procedure for many languages remains not completely obvious. When looking only at the mean performance difference between a light and an aggressive stemmer, the variation

[1] Freely available at www.unine.ch/info/clef/ or at tartarus.org/martin/PorterStemmer/.

depends on the language, IR model, and test collection. For the English language, the average performance differences between a light (S-stemmer) and Porter's stemmer is 1% over five IR models and in favor of Porter's solution. This difference is however not statistically significant. With the Russian language, the difference is also 1% in average, but in favor of a light approach. For French, the aggressive stemmer performs, in mean, 6% better, but only 3% for Czech. Thus no clear and definitive conclusion can be reached when comparing the effectiveness of a light vs. more aggressive stemmer.

Finally, compounding, i.e. a word formation process where new words are formed based on multiple simpler "components" (e.g., ghostwriter, dishwasher), is another linguistic construction that can affect the IR quality. This form is active in many languages (e.g., "capogiro" (dizziness) in Italian, "rakkauskirje" (love (rakkaus) and letter (kirje) in Finnish) but especially in German compounding is frequent and raises a specific challenge (Sanders 2010). First, this language allows long constructions (e.g., "Lebensversicherungsgesellschaftsangestellter" = "Leben" + s + "Versicherung" + s + "Gesellschaft" + s + "Angestellter" for life + insurance + company + employee)). Second, the same concept can equally be expressed using a compound term (e.g., "Computersicherheit") or a noun-phrase ("die Sicherheit für Computer"). As one form can appear in a relevant text and the second in the query, this aspect renders semantic matching more difficult. Thus, for the German language, a decompounding procedure (Chen 2004; Savoy 2003b) must be applied to achieve higher effectiveness. Such an automatic word decomposition can improve the MAP by 23% for short queries (title-only) or +11% for longer request formulation (Braschler and Ripplinger 2004). Similar mean performance differences have been found by Hedlund et al. (2004).

2.3 IR Models

In designing, implementing, and testing IR tools for European languages, different IR models have been used, such as variants of the vector-space models (Buckley et al. 1995; Manning et al. 2008), Okapi (Robertson et al. 2000), language models (Hiemstra 2000), and probabilistic approaches derived from "Deviation From Randomness" (DFR) (Amati and van Rijsbergen 2002). The formulations underlying these approaches are based on three main components, namely (1) the term frequency (*tf*) of the corresponding term in the document or the user's request, (2) the inverse document frequency (or *idf*), and (3) a length normalization procedure.

Essentially, these three factors encode the notion that a term should contribute most to the calculation of the item's score (or RSV, retrieval status value), if that term is found frequently in a document ("locally" frequent) and rarely in the overall collections ("globally" rare). The three factors have proven to be useful to discriminate between the major and minor semantic aspects of a document (or a request). Moreover, this formulation does not depend on the underlying natural language which can be an Indo-European one (Savoy 2003a), an Indian language (Dolamic and Savoy 2010a), or even Chinese, Japanese, Korean (Savoy 2005), the last three requiring however a more complex tokenization procedure.

Overall, our experiments indicate that both Okapi and variants of DFR tend to produce the highest retrieval effectiveness over numerous languages using the CLEF test collections (composed mainly of newspapers), and are thus most "robust" towards the different characteristics of the languages we have studied. The IR schemes derived from a language model tend to produce high mean average precision, marginally lower that those achieved by the Okapi or some DFR approaches. In all these implementations however, the best values for the underlying parameters are not known in advance and may have an impact of the overall effectiveness.

3 Bilingual Information Retrieval

Bilingual Information Retrieval (BIR) corresponds to the simplest form of information retrieval in which the requests are written in one language and the information items in another. Often, the term "cross-language" (or "cross-lingual") information retrieval (CLIR) is used as an alternative. The latter term is, however, less precise and can also be applied to scenarios with more than two languages involved. In nearly all cases, a direct matching between the query and the document surrogates does not work effectively in a bilingual scenario, and a translation stage must thus be incorporated during the IR process.

To achieve this, the simplest strategy is to translate the requests into the target language, knowing that queries are usually shorter than documents (query translation). The second approach consists of translating the whole text collection into the query language(s) (document translation). In this case, the translation process can be done off-line, and thus the translation process does not increase the response delay.

In some particular circumstances, the translation step can be ignored. Belonging to the same language family, some words may appear in different languages with the same or similar spelling (e.g., cognates such as, for example music, "Musik" (German), "musica" (Italian), "musique" (French), "música" (Spanish)). For some closely related languages, a rather large part of the vocabulary has similar spellings in the two languages, as for example, English and French, or German and Dutch. This aspect can be also explained by the presence of numerous loanwords (e.g., joy and "joie" (French)). Therefore, retrieval is possible when assuming that "English is simply misspelled French" (Buckley et al. 1997).

In this perspective for retrieval purposes, the translation stage is then replaced by a soft matching based on a spell corrector. This ingenious strategy is only possible for a limited number of closely related languages.

Moreover, this approach does not usually perform as well as an IR system with an explicit translation procedure (the solution achieves approximatively 60% of the effectiveness of a monolingual retrieval). In addition, sometimes the meaning differs even if the spelling looks similar (e.g., "demandes du Québec" must be translated into "requests of Quebec" and not as "demands of Quebec").

Therefore, to achieve a good overall IR performance, a form of explicit translation must be included during the IR process. This can be achieved using various techniques as shown in Sect. 3.1. The next section presents an architecture based on a query-translation approach and indicates some effectiveness measures.

3.1 Translation Strategies

A good translation requires knowing the meaning of the source text, and therefore could be hard to perform perfectly automatically. Note, however, that in a retrieval scenario, it may not be necessary to render a translation in the classical sense. The role of the "translated" query is merely the retrieval of relevant items in the other language; for this, any representation of the query *intent* in the target language, whether directly recognizable as translation or not, is suitable.

During the translation process, different forms of ambiguity must be resolved. For example, the correct translation of a word or expression depends on the context (word sense disambiguation) as, for example, the translation of the word "bank" differs if one considers a river or a financial context. Similarly, the French word "temps" could be translated into "time," "weather," or even "tense". Thus, for a given term, the translation process could be hard in one direction, but not in the other.

Moreover, not every word in one language does necessarily have a direct corresponding one in the target language (e.g., the occurrence of "have" in "have to" or "have" must usually be translated differently). Therefore, a word-by-word translation does not provide the best solution.

Multi-word expressions raise another set of ambiguities. Idiomatic expressions (e.g., "to see stars") cannot be translated as is into the target language. In other cases, the culture generates expressions that do not have a direct equivalent in the target language (e.g., "a lame duck Congressman").

As translation strategies (Zhou et al. 2012), the BIR experiments performed during the CLEF campaigns have tried different tools and IR models. As a first solution, one can use machine-readable bilingual dictionaries (MRD). In this case, each surface word in one language is searched in an MRD and the set of possible translations is returned. Even if some MRDs return, on average, only one or two possible translations, for some words the number of translations can be far larger (we have observed up to 15). It is not clear whether the IR system must take account only of the first one, the first k (with $k = 3$ to 5), or simply all translations. Usually, the IR system assumes that the translations are provided in a rank reflecting their decreasing frequency or usefulness. Thus, a weight assigned to each translation can depend on its position in the returned list.

The issue of how many candidate translations for a term should be included in the translated query representation has been handled in different ways. Assuming that the MRD returns the candidate in descending order of frequency of occurrence, the output can be pruned by accepting at most k translation candidates. This approach is problematic if $k > 1$, since unambiguous source language terms will then be under-represented in the translated rendering. Hedlund et al. (2004) present a remedy to this with their "structuring of queries" approach, where the k translation candidates are weighted as a "synonym set", instead of individually. They give results from experiments with three source languages (Swedish, Finnish, and German) and find consistent benefits of using structured queries. Greatest benefits are reported for Finnish, where they obtained an increase in retrieval effectiveness of up to 40%.

A more linguistically motivated alternative is the attempt to select "the" optimal translation candidate, e.g., through word sense disambiguation. Approaches using automatically generated dictionaries from corpora can be helpful here, as they can reflect specific domains in the context of which translation is less ambiguous. We will discuss relevant approaches to produce such statistical translation resources below.

MRDs as a translation tool must be integrated with caution. An MRD is not the same as a paper-based bilingual dictionary. In this latter case, each dictionary entry corresponds to a lemma (e.g, "to see"), but the surface word may include inflectional suffixes (e.g., "sees", "saw", "seen"). Thus, the link between the surface word and the lemma could be problematic.

Moreover, names may raise additional difficulties when they do not appear in the dictionary and sometimes the spelling varies from one language to the other (e.g., Putin, Poutine, Poetin). Of course, names are not limited to well-known politicians but can denote a product or an artwork (e.g., "Mona Lisa" (Italian), "La Joconde" (French) or "La Gioconda" (Spanish)). When names are relatively frequent, their translations can be obtained by consulting specialized thesauri (e.g., *JRC-Names*, *Arts and Architectures Thesaurus*, *The Getty Thesaurus of Geographic Names*). Similar data structures can also be built from other sources such as the *CIA World Factbook*, various gazetteers, or by downloading Wikipedia/DBpedia pages written in different languages. A similar solution can be applied to translate acronyms (e.g., UN must appear as ONU (in Spanish, French, Italian), UNO (in German), ONZ (in Polish), or YK (in Finnish)), under the assumption that a short sequence of uppercase letters corresponds to an acronym.

When a translation is not returned for a given word (out-of-vocabulary problem) resulting from a dictionary's limited coverage, the usual reason is the presence of a name (e.g., London, Renault) and the corresponding word can be kept as it is or translated by the previously mentioned tools. In other cases, a word (corresponding to a name) should not be translated (e.g., Bush).

Finally, the most appropriate translation can depend on the national origin of the target collection. Each language is strongly related to a culture. Therefore, one word or expression can appear in a given region, not in another one (or with a different meaning). For example, the translation of "mobile phone" into French can be “téléphone portable” (France), “téléphone mobile” (Belgium), “cellulaire” (Canada) or “natel” (Switzerland).

As a second translation strategy, one can adopt a machine translation (MT) system that will automatically provide a complete translation of a given request (or document) into the target language. As well-known examples, one can mention Google or Yahoo! online translation services. Various other systems have been developed such as Systran, Promt/Reverso, Babel Fish or WorldLingo. A classic example of the use of such automatic translation system is the Canadian weather forecast (started in 1971), while the latest version translates also weather warnings (Gotti et al. 2013).

As a third possibility of identifying proper translation candidates, we can apply a statistical translation model (Kraaij et al. 2003).[2] Advances in the effectiveness of machine translation systems reduce the role of statistical translation models for bilingual and multilingual retrieval to something of a niche role; however, there is still considerable potential for cases where special vocabulary (e.g., many proper names) and/or less frequently spoken languages are involved. Ideally, the model is built on the basis of a parallel corpus (i.e., a corpus that contains high-quality translations for all the documents) written in the desired languages. By aligning the translated documents at sentence level, pairs of terms across the languages are identified as translation candidates. Building a data structure from these pairs, the most probable match or the best *k* matches (Braschler and Schäuble 2001) can serve as retrieval terms.

In principle, this approach is workable independently of the languages considered. The availability of a suitable parallel corpus covering both the languages and the desired target domain, however, remains a concern. In Braschler (2004), we show that the requirement for a parallel corpus is not a strict one; instead, a comparable corpus that works on a much coarser “document similarity” basis, may be sufficient and may be much easier to obtain. Nie et al. (1999) discusses how suitable candidate documents can be identified in publicly accessible Web resources. Starting from a comparable corpus (Braschler 2004), shows how documents are “aligned” if they describe the same news event, even if produced independently by different authors. By modifying the *tf idf*-weighting formula to retrieve terms that co-occur in a training set of documents, a very large translation resource can be built that covers a vocabulary that is potentially much larger than that of MRDs (“similarity thesaurus”). Of course, the overall performance of such statistical translation systems depends on important factors, such as quality and size of the sources (Kraaij et al. 2003), along with the role played by cultural, thematic and time differences between the training corpora and the target domain.

[2] For example, by using the freely available Moses system, see www.statmt.org/moses/.

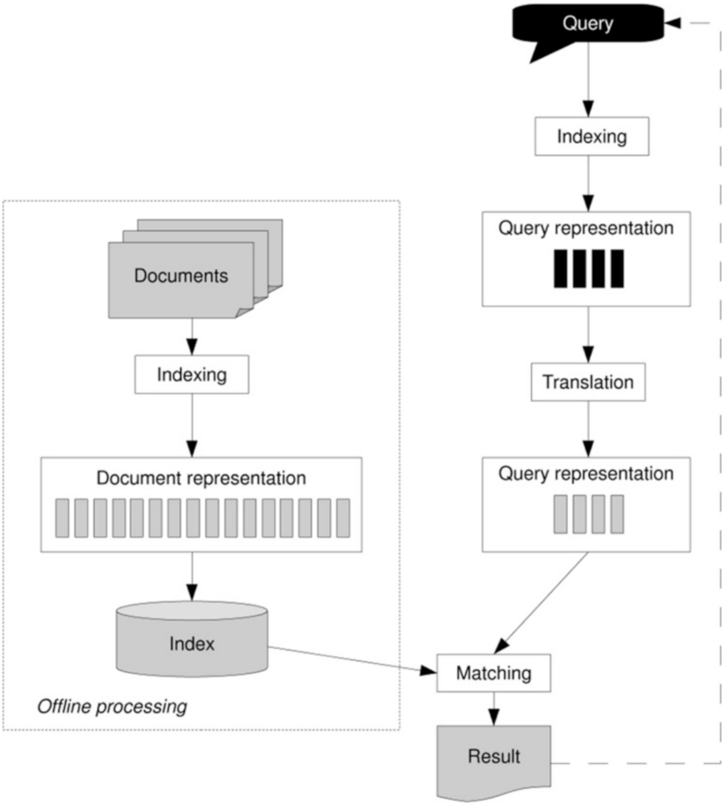

Fig. 1 Main architecture for a bilingual information retrieval system

3.2 Query Translation

To implement a query translation process, one can insert the automatic translation phase between the request acquisition and query indexing stage. As the request is usually rather short, the translation delay can be brief and done in real time at query time. As a translation strategy, one can implement an approach based on MRDs, an MT system or using a statistical translation model. Based on our experiments, the MT approach tends to produce the highest average performance level. As a variant depicted in Fig. 1,[3] the query representation can be generated in the query language and then translated into the target language in which the search is performed. If needed, the search result can be translated into the query language.

The overall quality of a translation system depends also on the language pair, and having English as one of the two languages tends to produce better results

[3] The figures appearing in this chapter are reproduced from Peters et al.'s book (Peters et al. 2012).

(the demand for automatic translation from/to English is clearly higher than for another language). In comparing different languages when using the Google system (Dolamic and Savoy 2010b), we observe that the translation from queries written in French or Spanish in order to search in an English collection was easier than it was from the Chinese. Based on a DFR model, the MAP obtained for the bilingual search using the French or Spanish language in the query language achieves 92% of the MAP obtained for the monolingual search. This value decreases to 90% with German topics, and 82% with simplified Chinese as language. With the Yahoo! translation service, the situation was somewhat comparable, with the French language achieving the best MAP (82% of the monolingual search), and using Chinese as the query language was the most difficult (only 56% for the monolingual search).

As the first source of translation errors, one can find the problem of polysemy and synonymy attached to a word. With the French request "Vol du cri" ("Theft of *The Scream*"), the word "vol" can be translated into "flight" or "theft", both with a high probability of being correct. In other cases, the choice in the target language seems irrelevant from a semantic point of view because two words are viewed as synonyms (e.g., the German word "Wagen" could be translated into "car" or "automobile"). From an IR perspective, one of these possible correct translations will provide more relevant items (e.g., car) than the other (e.g., automobile).

The second main source of translation errors comes from names. For example, in the request "Death of Kim Il Sung", the last word can be incorrectly analyzed as the past participle of the verb "to sing". Therefore, the returned translation is inappropriate to retrieve all pertinent information items. With another translation tool, the term "Il" was incorrectly recognized as the chemical acronym for Illinium (an discontinued chemical element). Finally, the Spanish word "El Niño" must not be translated into English (i.e. "the boy") but must be kept as is when the underlying domain concerns global warming. Of course, manual translation does not guarantee correct expressions.[4]

In order to limit translation ambiguity, one can automatically add terms to the submitted request before translating it into the target language (Ballesteros and Croft 1997). In this case, the query is first used to search within a comparable collection of documents written in the request language. Based on a pseudo-relevance feedback scheme, new and related terms can then be added to the query before translation. Such new terms may reflect morphological variations (e.g., from a query about "London", the extended query may include additional terms or related concepts such as "Britain, British, PM, England").

As a second strategy to improve the BIR system, the translation stage can take account of more that one translation approach or source. It was shown that combining multiple translation sources (Savoy 2004) tends to improve the overall retrieval effectiveness (Savoy and Berger 2005). For example, using queries written

[4] In a hotel cloakroom in Germany, the following faulty translation was found: "Please hang yourself here." (Crocker 2006).

in English to search a collection written in another language, we have combined two alternative translated representations of the query. In the best case, searching in a French collection, the MAP can be improved from 8% to 12% compared to a single translation. Similar average enhancements can be found using the Spanish or Russian language (Savoy 2004). With the Italian language, the improvement was even higher, from 18% to 30%. When compared to the corresponding monolingual search and combining two translation tools, the performance difference is similar when searching in the French corpus (with English requests), with a 8% decrease for a collection written in German and around 10% decrease for the Spanish or Italian language. Those performance levels can be achieved when having the English as one of the languages. Of course, such a translation strategy is clearly more complex to design and to maintain in a commercial environment.

4 Multilingual Information Retrieval

Designing effective Multilingual Information Retrieval (MIR) systems corresponds to a very challenging issue. In such a context, the request can be written in one language while the information items appear in many languages. As for BIR, the translation process must be included in the IR process generating an additional level of uncertainty. In such an IR system, we usually assume that one document collection corresponds to one language. Therefore, the search must be done across different separate collections or languages. However, an MIR system can be built with different architectures, and the simplest one is based on a query-translation approach as described in Sect. 4.1. More complex approaches, usually achieving better retrieval effectiveness, implement a document translation phase as discussed in Sect. 4.2 or both a document and query translation process as described in Sect. 4.3.

4.1 Multilingual IR by Query Translation

As a first MIR architecture, one can simply translate the submitted request into all target languages. Note, however, that this approach suffers from scaling issues: as the number of languages to be covered grows, so does the number of translated representations that need to be produced. The number of bilingual language pairs can thus quickly become prohibitively large. After producing the individual translations, the search is performed separately in each language (or collection), each returning a ranked list of retrieved items. MIR then presents an additional problem. How can one merge these results to form a single list for the user in an order reflecting the pertinence of the retrieved items, whatever the language used (“merging”)? Figure 2 depicts the overall MLIR process based on a query translation (QT) strategy.

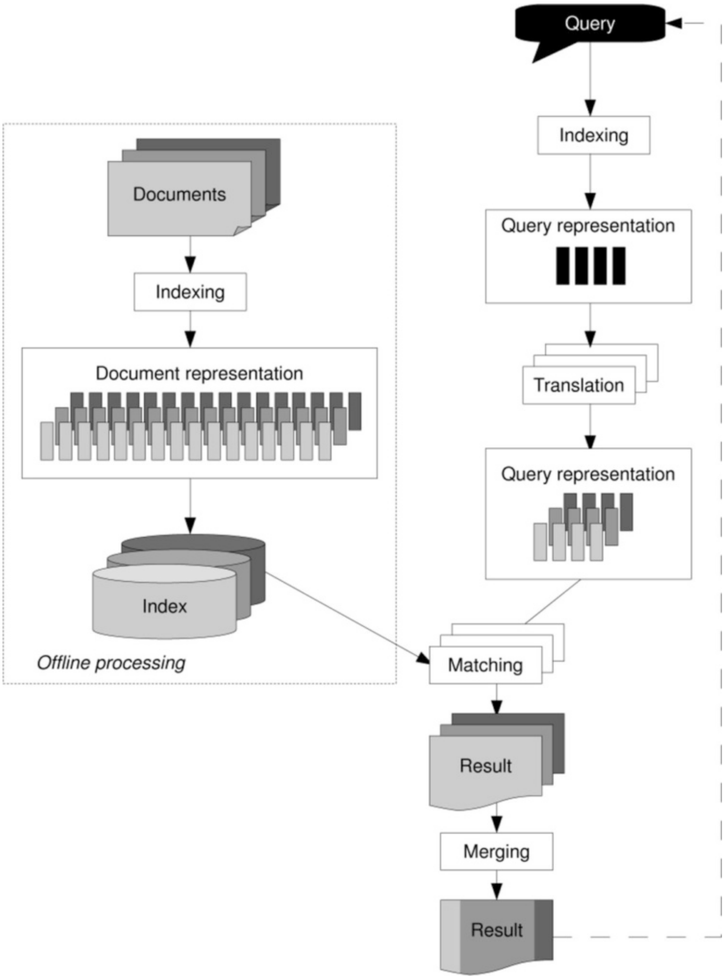

Fig. 2 Main architecture for a query translation model for a cross-language information retrieval system

As a first merging approach, one might assume that each language contains approximately the same number of pertinent items and that the distribution of relevant documents is similar across the result lists. Using the rank as the sole criteria, the simplest solution is then to interleave the retrieved records in a round-robin fashion. As an alternative, one can suggest a biased round-robin approach which extracts not one document per collection per round but one document for each smaller collections and more than one for larger ones (Braschler et al. 2003).

To account for the document score (or RSV) computed for each retrieved item (or the similarity value between the retrieved record and the query), one can formulate the hypothesis that each collection is searched by the same or a very similar search engine. In such cases, the similarity values are directly comparable across languages/collections. Such a strategy, called raw-score merging, produces a final list sorted by the document score computed separately by each collection.

However, as demonstrated by Dumais (1994), collection-dependent statistics in document or query weights may vary widely among collections, and therefore this phenomenon may invalidate the raw-score merging hypothesis. But different evaluations carried out using English only documents have demonstrated that the raw-score merging strategy sometimes leads to satisfactory performance (Rasolofo et al. 2003).

As a third merging strategy, one can normalize document scores within each collection by dividing them by the maximum score (i.e. the document score appearing in the first position (Fox and Shaw 1994), a strategy denoted "Norm Max"). This procedure could generate more comparable document scores across all languages/collections. As a variant of this normalized score merging scheme, Powell et al. (2000) suggest normalizing the document scores by taking the maximum and minimum document score (approach denoted "Norm RSV") and explained by Eq. (1).

$$Norm\ RSV(D_k^i) = \frac{RSV_k^i - Min(RSV^i)}{Max(RSV^i) - Min(RSV^i)} \tag{1}$$

where RSV_k^i indicates the retrieval score of document k in the ith retrieved list, and $Max(RSV^i)$ (respectively $Min(RSV^i)$) the maximum (minimum) RSV value appearing in the ith list.

As a fifth merging strategy, the "Zscore" approach (Savoy 2004) has been suggested in which the normalization of the RSV values depends on the RSV distribution, using its mean ($Mean(RSV^i)$) and estimated standard deviation ($Std(RSV^i)$). The precise definition is provided by Eq. (2).

$$Z\ score(D_k^i) = \frac{RSV_k^i - Mean(RSV^i)}{Std(RSV^i)} + \delta^i \quad \delta^i = \frac{Mean(RSV^i) - Min(RSV^i)}{Std(RSV^i)} \tag{2}$$

Finally, machine learning methods can be applied to improve the merging operation. In this perspective, a logistic regression approach can be used to estimate the probability of relevance for a given document, based on its retrieval status value and the natural logarithm of its rank. The final list is sorted according to these estimates. The evaluation is performed based on the leaving-one-out evaluation strategy producing an unbiased estimator of the real performance.

To analyze the quality of these merging operators, the CLEF 2004 test collection has been selected (Savoy and Berger 2005). This corpus contains newspapers articles written in English, French, Finnish, and Russian. Table 1 indicates the number of queries with relevant items in each language, as well as the MAP achieved when applied to the original queries (column denoted "Manual" or monolingual run).

Table 1 MAP of each single run

Query (TD)		Mean average precision (MAP)		
Language	Number of queries	Manual	Condition A	Condition B
English	42	0.5580	0.5580	0.5633
French	49	0.4685	0.4098	0.4055
Finnish	45	0.4773	0.2956	0.2909
Russian	34	0.3800	0.2914	0.2914

Table 2 MAP of various multilingual merging strategies

Query (TD)	Mean average precision (MAP)		
Merging operator	Condition A	Condition B	Difference
Round-robin	0.2386	0.2358	−1.2%
Biased round-robin	0.2639	0.2613	−1.0%
Raw-score	0.0642	0.3067	377.7%
Norm max	0.2552	0.2484	−2.7%
Norm RSV	0.2899	0.2646	−8.7%
Z-score	0.2669	0.2867	7.4%
Logistic regression	0.3090	0.3393	9.8%
Optimal selection	0.3234	0.3558	

Under Condition A (bilingual runs with English queries), we have tried to obtain a high MAP per language, applying different IR models with distinctive parameter values for each language. Under Condition B, the same IR model (a variant of the DFR family) is used for each language (with similar parameter values). This last choice reflects the case where a single IR model is used to search across different collections/languages.

Table 2 reports the MAP achieved when applying different merging operators. The round-robin method must be viewed more as a baseline than a really effective approach. When distinct IR models are merged (Condition A), the raw-score merging strategy resulted in poor retrieval effectiveness. On the other hand, when applying the same IR model (with similar parameter values), the raw-score approach offers higher MAP. The normalization procedures (either by the Norm Max or the Norm RSV) or the Z score technique tend to produce better retrieval results than the round-robin technique under both conditions.

In some circumstances, an effective ranking can be learnt from past results. As an example, a logistic regression model can use both the rank and the document score as explanatory variables to predict the probability of document relevance. When such training sets are available and the similarity between trained and test topics is high, the merging achieved can be significantly better than the round-robin merging as well as better than the simple normalization approaches (see Table 2). Finally, the last row of Table 2 reports the optimal merging result that can be achieved based on the returned lists per language. Compared to the round-robin strategy, this optimal merging offers a 36% improvement under Condition A (0.3234 vs. 0.2386) and +50% under Condition B (0.3558 vs. 0.2358).

4.2 Document Translation

Document translation (DT) provides an attractive alternative approach avoiding the merging problem. By translating all documents into a single, unified target language, the multilingual retrieval problem is essentially reduced to a monolingual one. Interestingly, the merging problem is thus avoided altogether. For reasons of its superior language resources, a pertinent choice for the pivot language is English. To justify this choice, we describe the following experiment.

In the experiment (Savoy and Dolamic 2010), we needed to translate 299 queries written in German to search in a French collection. Compared to a monolingual run (MAP: 0.6631), the achieved MAP was 0.4631 resulting in a decrease of around 30%. Using English as the query language, the MAP was 0.5817, for a performance difference of 12% compared to the monolingual run. Clearly the translation quality was higher from English than from the German language. Moreover, we need to limit the number of translation pairs. In our case, we are using English as pivot language. In a second stage, we first translate the German queries into English and then into French. After this two-stage translation, it is reasonable to expect a poor retrieval performance. Using English as pivot language, the resulting MAP was 0.5273, with an average decrease of only 20% (compared to the 30% with a direct translation from German to French). Similar good retrieval performances with a pivot language were observed in Hedlund et al. (2004). An example of the resulting MLIR process is depicted in Fig. 3.

As a second model, we can translate all text collections into all query languages. Receiving the query in one of the available languages, the search is then performed as a monolingual one. In this case, no translation is performed during query processing.

All translation strategies outlined in Sect. 3.1 equally apply to document translation. Since a document (retrievable item) is typically much longer than a query, more context is available, and problems with out-of-vocabulary terms and synonymy tend to be less pressing. Moreover, there is justified hope that the information contained in some of the untranslatable terms is represented, at least partially, in the remainder of the document. Note that, analogously to the situation in query translation, a translation in the "classical sense" is not necessary; any rendering of the document into a representation in the target language that is suitable for retrieval will do (e.g., the syntax of the target language is not always perfectly respected ("pseudo translation")).

Translation of large document collections, even if automated, is a costly task. The document translation approach also does not scale well as the number of query languages grows—in essence, the collection has to be replicated (and re-translated) for each target language. On the positive side, it is possible to do this translation offline, with no performance impact of translation during query time.

Examples of document translation-based experiments in the CLEF ad hoc tracks are reported in Braschler (2004) and McNamee and Mayfield (2002, 2004). In our experiments, we have gained the most insight in document translation behavior

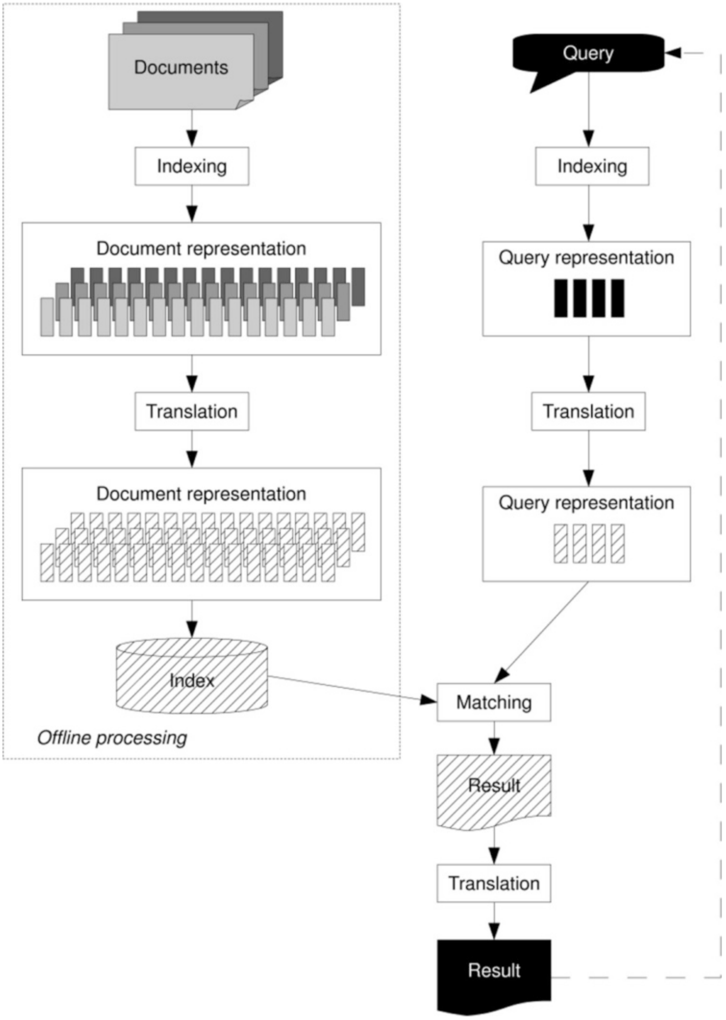

Fig. 3 Main architecture for a document translation model for a cross-language information retrieval system

from using the CLEF 2002 test collection containing documents written in English, German, French, Italian, and Spanish. We have used the German query set.

In order to have an idea about the performance differences between a QT and DT approaches, we have conducted the following experiment. First, we considered two query translation (QT) approaches, namely round-robin, and biased round-robin. As shown in Table 3, these two QT approaches tend to produce similar overall mean average precision. In the last column, we have indicated the performance difference with the round-robin solution.

Table 3 Eurospider experiments on the CLEF 2002 multilingual corpus, German queries

Strategy	Mean average precision	Difference
Query translation, round-robin	0.3249	
Query translation, biased round-robin	0.3369	+3.7%
Document translation	0.3539	+8.9%
Optimal selection	0.4876	+50.1%

Table 4 Hybrid approach vs. document translation only or query translation only

	Mean average precision (MAP)		
Strategy	CLEF 2000	CLEF 2001	CLEF 2002
Query translation (QT)	0.2500	0.2773	0.2876
Document translation (DT)	0.2816	0.3099	0.3539
DT + QT	0.3107	0.3416	0.3554

Second, in the fourth row, our document translation (DT) is evaluated. One can see that this DT approach outperformed the three QT strategies. However, when comparing to the benchmark of "Optimal Selection" (see Sect. 4.1), i.e. under the condition that the merging problem is "solved", a different conclusion must be drawn. Note, however, that compared to simple merging strategies, we have found consistently better results for document translation across all years where we have made such comparisons (CLEF 2000–2002) as reported in Table 4.

4.3 Hybrid Approaches

Using the mean as a measure, we obtain a synthetic value reflecting the overall performance of an IR system. The differences between the average precision achieved by each query are however hidden. Looking at individual queries, it becomes evident that performance differences between query translation and document translation approaches vary greatly. To take advantage of both translation models, a hybrid approach can combine their outputs. In this scenario, a more robust solution can be proposed with respect to outliers. Indeed, our experiments on CLEF 2000–2002 test collections have shown an increase in mean average precision for all three years as reported in Table 4. As indicated previously, the document translation strategy performs better that the query translation approach over the three years. When comparing the document translation (second row) with the hybrid model (last row), the performance differences are always in favor of the hybrid model, although the difference for 2002 is negligible.

Analyzing query-by-query these results, we can see that the hybrid strategy proposes a better average precision for the majority of the queries. In Table 5, we have depicted the number of queries performing better in terms of average precision (over the set of 50 queries available each year). For example, for the CLEF 2001 collection, 41 out of 50 queries benefit from the hybrid approach when compared to

Table 5 Impact on individual queries

Strategy	CLEF 2000	CLEF 2001	CLEF 2002
DT + QT vs. DT only	32:8	41:9	28:22
DT + QT vs. QT only	31:9	36:14	41:9

document translation only, while this value reaches 36 when comparing to the query translation approach.

5 Conclusion

During our ten years of participation in the mono-, bi-, and multilingual tracks at CLEF, we have designed, implemented, and evaluated various IR tools for a dozen of European natural languages. Those experiments tend to indicate that the IR models validated on various English collections (e.g., TREC, NTCIR, CLEF, INEX) perform also very well with other European (Savoy 2003a), Indian (Dolamic and Savoy 2010a), or Far-East (Savoy 2005) languages. No special adaptation is really required when considering the *tf*, *idf*, and length normalization components. On the other hand, some IR procedures must take into account the specifies of each language.

Each natural language presents its own difficulties when building effective IR systems. To generate a stopword list, we suggest considering all closed part-of-speech categories (determiners, prepositions, conjunctions, pronouns, and auxiliary verb forms). In this list, an inspection is needed to verify, according to the target application or domain, whether some forms must be removed or not from the stopword list (e.g., the article “a” can appear in the context of “vitamin A”).

To develop a stemmer for a new language, we suggest focusing mainly on morphological variations related to nouns and adjectives, and to ignore the usually too numerous suffixes related to verbs. Moreover, removing only the inflectional suffixes seems to be good practice for many languages. Adopting this approach, the edit distance between the search term introduced by the user and its internal representation is rather small. With a light stemmer, one can improve the MAP in the range of 5% to 10% (e.g., French or German language) up to 96% (Russian).

If needed, and according to the target application, an advanced stemmer can be proposed to remove both inflectional and derivational suffixes. The enhancement over a light stemmer is between −1% (Russian) to +6% (French). Trying to remove verbal suffixes tends to be more problematic by generating too many incorrect conflations for nouns and adjectives. For the German language only, we recommend implementing an automatic decompounding procedure, leaving both the compound and its separate components in the document or query surrogate. This strategy can increase the mean performance by 23% (Braschler and Ripplinger 2004).

Recent research has been conducted to analyze in a more systematic way the effect of different stopword lists and stemmers, as well as their combined effect (Ferro and Silvello 2016a,b).

When implementing a bilingual IR system, the crucial component is clearly the translation procedure. When the pair of languages includes English and one of the most widely spoken languages (such as Spanish, German, or French), currently available machine-translation systems offer high effectiveness from an IR point of view (Dolamic and Savoy 2010b). Even if the translation is not fully correct from a linguistic standpoint, the search engine is able, on average, to find the appropriate related search terms and to retrieve the pertinent items. In such circumstances, the decrease of the mean performance compared to the monolingual setting is rather limited (−5% to −12%), and in the best case, no degradation occurs. For other languages (e.g., Finnish, Polish), the number of translation tools is rather limited and their quality is clearly inferior to those available for the most frequently spoken languages. The retrieval performance can however be improved by combining multiple translations of the same texts on the one hand, and on the other, by applying some query expansion before the translation. However, such IR strategies render the final system more complex and difficult to maintain.

When the translation resources available are limited or absent, the usual solution is to generate a statistical translation system based on parallel corpora (Kraaij et al. 2003). In this case, the mean retrieval precision typically decreases substantially (from 10% to 40%). Finally, more specific IR models have been proposed to take account of the additional uncertainty generated by the translation process.

Multilingual IR corresponds to our most complex situation in which the overall performance depends on many factors and where the quality of the translation plays an important role. From an architecture point of view, two main approaches have been tested. The simplest one is based on query translation (QT) in which the submitted query is translated automatically into all the target languages. The search is then done separately in all languages, and the results are then merged to generate a single ranked list of retrieved items to be presented to the user. The main difficulty in this model is the merging process that can substantially degrade the overall performance. Our experiments indicate that selecting a form of normalization of the document score (e.g., Norm RSV or the Z score) can offer a reasonable overall IR performance.

In a document translated (DT) model, all documents are translated into a single pivot language (usually in English). The submitted request is also automatically translated into this pivot language. The search process is then done in a single language and the resulting ranked list can be directly returned to the user. Such solutions tend to produce a better overall retrieval performance compared to query translation approaches.

Acknowledgement The authors would like to thank the CLEF organizers for their efforts in developing the CLEF test collections.

References

Amati G, van Rijsbergen CJ (2002) Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans Inf Syst 20:357–389

Ballesteros L, Croft BW (1997) Phrasal translation and query expansion techniques for cross-language information retrieval. In: Proceedings ACM SIGIR. ACM Press, New York, pp 84–91

Braschler M (2004) Combination approaches for multilingual text retrieval. Inform Retrieval J 7:183–204

Braschler M, Ripplinger B (2004) How effective is stemming and decompounding for German text retrieval? Inform Retrieval J 7:291–316

Braschler M, Schäuble P (2001) Experiments with the eurospider retrieval system for CLEF 2000. In: Peters C (ed) Cross-language information retrieval and evaluation. LNCS, vol 2069, Springer, Berlin pp 140–148

Braschler M, Göhring A, Schäuble P (2003) Europsider at CLEF 2002. In: Peters P, Braschler M, Gonzalo J, Kluck M (eds) Advances in cross-language information retrieval: third workshop of the cross–language evaluation forum (CLEF 2002) revised papers. LNCS, vol 2785. Springer, Berlin, pp 164–174

Buckley C, Singhal A, Mitra M, Salton G (1995) New retrieval approaches using SMART. In: Proceedings TREC-4, NIST, Gaithersburg, pp 25–48

Buckley C, Singhal A, Mitra M, Salton G (1997) Using clustering and superconcepts within SMART: TREC-6. In: Proceedings TREC-6, NIST, Gaithersburg, pp 107–124

Chen A (2004) Report on CLEF-2003 monolingual tracks: fusion of probabilistic models for effective monolingual retrieval. In: Peters C, Gonzalo J, Braschler M, Kluck M (eds) Comparative evaluation of multilingual information access systems, LNCS, vol 3237. Springer, Berlin, pp 322–336

Crocker C (2006) Løst in Tränšlatioπ. Misadventures in English abroad. Michael 0'Mara Books, London

Dolamic L, Savoy J (2009a) Indexing and searching strategies for the Russian language. J Am Soc Inf Sci Technol 60:2540–2547

Dolamic L, Savoy J (2009b) Indexing and stemming approaches for the Czech language. Inf Process Manag 45:714–720

Dolamic L, Savoy J (2010a) Comparative study of indexing and search strategies for the Hindi, Marathi and Bengali languages. ACM Trans Asian Lang Inf Process 9(3):11

Dolamic L, Savoy J (2010b) Retrieval effectiveness of machine translated queries. J Am Soc Inf Sci Technol 61:2266–2273

Dolamic L, Savoy J (2010c) When stopword lists make the difference. J Am Soc Inf Sci Technol 61:200–203

Dumais ST (1994) Latent semantic indexing (LSI) and TREC-2. In: Proceedings TREC-2, vol #500-215. NIST, Gaithersburg, pp 105–115

Fautsch C, Savoy J (2009) Algorithmic stemmers or morphological analysis: an evaluation. J Am Soc Inf Sci Technol 60:1616–1624

Ferro N, Silvello G (2016a) A general linear mixed models approach to study system component effects. In: Proceedings ACM SIGIR. ACM Press, New York, pp 25–34

Ferro N, Silvello G (2016b) The CLEF monolingual grid of points. In: Fuhr N, Quaresma P, Gonçalves T, Larsen B, Balog K, Macdonald C, Cappellato L, Ferro N (eds) Experimental IR meets multilinguality, multimodality, and interaction. Proceedings of the eighth international conference of the CLEF association (CLEF 2017). LNCS, vol 9822. Springer, Berlin, pp 13–24

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

Fox EA, Shaw JA (1994) Combination of multiple searches. In: Proceedings TREC-2, vol 500-215. NIST, Gaithersburg, pp 243–249

Gotti F, Langlais P, Lapalme G (2013) Designing a machine translation system for the Canadian weather warnings: a case study. Nat Lang Eng 20:399–433

Harman DK (1991) How effective is suffixing? J Am Soc Inf Sci 42:7–15

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. Inf Retrieval J 7:99–120

Hiemstra D (2000) Using language models for IR. PhD thesis, CTIT, Enschede

Kraaij W, Nie JY, Simard M (2003) Embedding web-based statistical translation models in cross-lingual information retrieval. Comput Linguist 29:381–419

Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press, Cambridge

McNamee P, Mayfield J (2002) Scalable Multilingual Information Access. In: Peters P, Braschler M, Gonzalo J, Kluck M (eds) Advances in cross-language information retrieval. LNCS, vol 2785. Springer, Berlin, pp 207–218

McNamee P, Mayfield J (2004) Character N-gram tokenization for European language text retrieval. Inf Retrieval J 7:73–98

McNamee P, Nicholas C, Mayfield J (2009) Addressing morphological variation in alphabetic languages. In: Proceedings ACM - SIGIR. ACM Press, New York, pp 75–82

Moulinier I (2004) Thomson legal and regulatory at NTCIR-4: monolingual and pivot-language retrieval experiments. In: Proceedings NTCIR-4, pp 158–165

Nie JY, 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. ACM Press, New York, pp 74–81

Paik JH, Parai SK (2011) A fast corpus-based stemmer. ACM Trans Asian Lang Inf Process 10(2):8

Paik JH, Parai SK, Dipasree P, Robertson SE (2013) Effective and robust query-based stemming. ACM Trans Inf Syst 31(4):18

Peters C, Braschler M, Clough P (2012) Multilingual information retrieval. From research to practice. Springer, Berlin

Porter MF (1980) An algorithm for suffix stripping. Program 14:130–137

Powell AL, French JC, Callan J, Connell M, Viles CL (2000) The impact of database selection on distributed searching. In: Proceedings ACM-SIGIR. ACM Press, New York, pp 232–239

Rasolofo Y, Hawking D, Savoy J (2003) Result merging strategies for a current news metasearcher. Inf Process Manage 39:581–609

Robertson SE, Walker S, Beaulieu M (2000) Experimentation as a way of life: Okapi at TREC. Inf Process Manage 36:95–108

Sanders RH (2010) German, biography of a language. Oxford University Press, Oxford

Savoy J (2003a) Cross-language information retrieval: experiments based on CLEF 2000 corpora. Inf Process Manage 39:75–115

Savoy J (2003b) Cross-language retrieval experiments at CLEF 2002. In: Peters P, Braschler M, Gonzalo J, Kluck M (eds) Advances in cross-language information retrieval. LNCS, vol 2785. Springer, Berlin, pp 28–48

Savoy J (2004) Combining multiple strategies for effective monolingual and cross-lingual retrieval. Inf Retrieval J 7:121–148

Savoy J (2005) Comparative study of monolingual and multilingual search models for use with Asian languages. ACM Trans Asian Lang Inf Process 4:163–189

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

Savoy J (2008a) Searching strategies for the Bulgarian language. Inf Retrieval J 10:509–529

Savoy J (2008b) Searching strategies for the Hungarian language. Inf Process Manage 44:310–324

Savoy J, Berger PY (2005) Selecting and merging strategies for multilingual information retrieval. In: Peters C, Clough P, Gonzalo J, Jones GJF, Kluck M, Magnini B (eds) Multilingual information access for text, speech and images. LNCS, vol 3491. Springer, Berlin, pp 27–37

Savoy J, Dolamic L (2010) How effective is Google's translation service in search? Commun ACM 52:139–143

Zhou D, Truran M, Brailsford T, Wade V, Ashman H (2012) Translation techniques in cross-language information retrieval. ACM Comput Surv 45(1):1