Automatic morphological query expansion using analogy-based machine learning

Fabienne Moreau, Vincent Claveau, and Pascale Sébillot

IRISA, Campus universitaire de Beaulieu, 35042 Rennes cedex, France
{Fabienne.Moreau, Vincent.Claveau, Pascale.Sebillot}@irisa.fr

Abstract. Information retrieval systems (IRSs) usually suffer from a low ability to recognize a same idea that is expressed in different forms. A way of improving these systems is to take into account morphological variants. We propose here a simple yet effective method to recognize these variants that are further used so as to enrich queries. In comparison with already published methods, our system does not need any external resources or *a priori* knowledge and thus supports many languages. This new approach is evaluated against several collections, 6 different languages and is compared to existing tools such as a stemmer and a lemmatizer. Reported results show a significant and systematic improvement of the whole IRS efficiency both in terms of precision and recall for every language.

Key words: Morphological variation, query expansion, analogy-based machine learning, unsupervised machine learning

1 Introduction

Information retrieval systems (IRSs) aim at establishing a relation between users' information needs (generally expressed by natural language queries) and the information contained in documents. To this end, a commonly used method consists of making a simple match between the query terms and the document words. A document is said to be relevant if it shares terms with the query. IRSs face two problems with such a mechanism, mainly bound to the inherent complexity of natural language. The first problem is related to polysemy: a single term may have different meanings and represent various concepts (*e.g.* `bug`: `insect` or `computer problem`); because of term ambiguity, IRSs may retrieve non relevant documents. The second and dual issue reflects the fact that a single idea may be expressed in different forms (*e.g.* `bicycle-bike`). Therefore, a relevant document can contain terms semantically close but graphically different. To overcome those two limitations, a rather natural solution is to perform a linguistic analysis of documents and queries. Based on natural language processing (NLP) techniques, it enables to obtain richer and more robust descriptors than simple keywords. These descriptors are able to highlight the fact that a same word can have different meanings or undergo variations of form (`retrieve` ↔ `retrieval`), structure (`information retrieval` ↔ `information that is retrieved`) or

meaning (`seek` ↔ `search`). Among the various types of linguistic analysis that can be applied (*i.e.* morphological, syntactic or semantic), morphological analysis appears to be one of the most effective ones to improve IRS performances. It leads to recognize that words such as `produce`, `produced`, `producing`, and `producer`, although graphically different, are actually forms of the same word; in other terms they are morphological variants. Enabling the match of these graphically different but semantically close forms can be consequently relevant in information retrieval (IR).

Morphological variation is a well-known problem in IR and has been exhaustively investigated in the literature (for a state-of-the-art, see [1,2,3,4] for instance). Despite those studies, one main issue remains: the non-portability of the methods proposed for detecting morphological variants: a majority of them are developed for one given language and are based on external knowledge (list of endings, recoding rules, lexicon...); consequently, they cannot be re-used out of their framework of creation. Considering the potential impact of morphological information on the performances of IRSs, it is essential to conceive tools that exceed the limits of existing methods and are adapted to IR data specificities.

Therefore, a simple but effective approach using an unsupervised machine learning technique is proposed in order to detect morphological variants. This method has to fulfill the following requirements: it must not require any external knowledge or resources; it must be entirely automatic; it must be directly applicable to various languages. Our acquisition method is used in IR for query expansion. The goal of our approach is to detect, within a collection of texts, words that are in morphological relations with the terms of a query and to add them to it.

The rest of the paper has the following structure: Section 2 presents some of the approaches existing to take into account morphological variation in IR. Section 3 describes the method developed for the detection of morphological variants and its use in an IRS to extend queries. Section 4 details the experiment results obtained on various collections and discusses them. Finally, Section 5 concludes on the relevance of our method to improve IRS performances.

2 Background: morphological variation in IR

There are generally two ways for coping with morphological variation in IR: at indexing time (conflation approach) or at retrieval time (query expansion). In the conflation approach, the various forms of a same word (variants) are reduced to a common form (stem, root or lemma). Thus match between documents and query is done on the basis of this canonical form. In the expansion method, documents and queries are indexed with original word forms; and the terms of a user's query are expanded with their morphological variants at retrieval time (see [5] for instance). One usual technique to handle morphological variation is stemming, which is used to reduce variant word forms to common roots (stems) [1, for instance]. Other approaches choose more sophisticated tools based on

linguistic methods, like lemmatizers (inflectional morphology) or derivational analyzers [6,7,8].

The principal limit of the existing tools is that they are, in most cases, based on external resources such as affix lists, morphological rules or dictionaries. Consequently, they can only be applied to one very particular language and present a restricted coverage. Many studies yet suggest to use them in IR [1,3,9, *inter alia*]; the experiments tend to show the added-value of taking into account morphological variants to improve both recall and precision of systems. However, obtained results depend on numerous factors, like collection language, query length or document type (general or from specialized fields for example). More generally, among those studies, very few are compatible with the three requirements given in introduction as a framework of our work. Some approaches that meet entirely our constraints rely on statistical techniques, which have the advantage of being independent of the language and may be unsupervised. Thus, several word segmentation tools were developed while being mainly based on frequency criteria [10, for instance] or on a N-grams technique [11]. Generally, those statistical methods, although they answer our requirements, show low reliability for the detection of morphological variants [12] and their contributions to IR has not been really proved.

3 New automatic acquisition of morphological variants used to extend query in IR

We describe here our method to extract morphological variants from documents. To fulfill the three requirements enumerated in introduction, our approach is based on a rather simple but flexible technique better suited to IR specificities. The principles are the followings: an original technique (*cf.* Section 3.1) is used to detect every morphologically related word pairs (joined up by a link of morphological variation); since we are looking for query extensions, we use it to locate within the document database all the words that are morphologically related to one of the terms of the query. All the detected words are then added to this query for its expansion. The proposed acquisition method is first explained; then its use within IRSs for query expansion is described in details.

3.1 Learning by analogy

Our approach for morphological variant acquisition of query terms is based on a technique initially developed to be used in the field of terminology [13]. Its principle is simple and based on analogy. Analogy can be formally represented as A : B $\doteq$ C : D, which means "A is to B what C is to D"; *i.e.* the couple A-B is in analogy with the couple C-D. The use of analogy in morphology, which is rather obvious, has already been studied [14]. For example, if we have analogies like `connector` : `connect` $\doteq$ `editor` : `edit`, and knowing that `connector` and `connect` share a morpho-semantic link, we can guess a same link between `editor` and `edit`.

The most important feature in learning by analogy is the notion of similarity that is used to determine if two pairs of propositions —in our case, two pairs of words— are analogous. The similarity notion we use, hereafter *Sim*, is quite simple but well fit to many languages in which inflection and derivation are mainly obtained by prefixation and suffixation. Intuitively, *Sim* checks that to go from a word w_3 to a word w_4, the same "path" of deprefixation, prefixation, desuffixation and suffixation is needed as to go from w_1 to w_2. More formally, let us name lcss(X, Y) the longest common substring shared by two strings X and Y (*e.g.* lcss(`republishing`, `unpublished`) = `publish`), X $+_{suf}$ Y (respectively $+_{pre}$) being the concatenation of the suffix (resp. prefix) Y to X, and X $-_{suf}$ Y (respectively $-_{pre}$) being the removal of the suffix (resp. prefix) Y from X. The similarity measure *Sim* can then be defined as follows:

Sim(w_1-w_2, w_3-w_4) = 1 if the four following conditions are simultaneously met:

$$\begin{cases} w_1 = \text{lcss}(w_1, w_2) +_{pre} \text{Pre}_1 +_{suf} \text{Suf}_1, \text{ and} \\ w_2 = \text{lcss}(w_1, w_2) +_{pre} \text{Pre}_2 +_{suf} \text{Suf}_2, \text{ and} \\ w_3 = \text{lcss}(w_3, w_4) +_{pre} \text{Pre}_1 +_{suf} \text{Suf}_1, \text{ and} \\ w_4 = \text{lcss}(w_3, w_4) +_{pre} \text{Pre}_2 +_{suf} \text{Suf}_2 \end{cases}$$

Sim(w_1-w_2, w_3-w_4) = 0 otherwise

Pre_i and Suf_i are any character strings. If *Sim*(w_1-w_2, w_3-w_4) = 1, the analogy $w_1 : w_2 \doteq w_3 : w_4$ stands, then we can suppose that the morphological relation between w_1 and w_2 is identical to the one between w_3 et w_4.

Our morphological acquisition process checks if an unknown pair is in analogy with one or several given examples. For instance, we can determine that the couple `rediscovering-undiscovered` is in analogy with one example-pair `republishing-unpublished`, since the similarity measure defined as follows:

$$\begin{cases} w_1 = \texttt{publish} +_{pre} \texttt{re} +_{suf} \texttt{ing}, \text{ and} \\ w_2 = \texttt{publish} +_{pre} \texttt{un} +_{suf} \texttt{ed}, \text{ and} \\ w_3 = \texttt{discover} +_{pre} \texttt{re} +_{suf} \texttt{ing}, \text{ and} \\ w_4 = \texttt{discover} +_{pre} \texttt{un} +_{suf} \texttt{ed} \end{cases}$$

worths 1.

For efficiency reasons during analogy search, rather than the word-pair examples, the prefixation and suffixation operations used in the similarity measure are stored. Thus, the example-couple `republishing-unpublished` is not stored as such, but retained according to the following rule:

$$w_2 = w_1 -_{pre} \texttt{re} +_{pre} \texttt{un} -_{suf} \texttt{ing} +_{suf} \texttt{ed}$$

To show the analogy `republishing` : `unpublished` $\doteq$ `rediscovering` : `undiscovered` consists in testing that `rediscovering-undiscovered` verifies the preceding rule.

As already emphasized in [6], prefixation and suffixation operations considered in our approach enable to take into account partly the light variations of roots as long as they are common enough to be present in one of our examples. More complex variations such the one existing in `go-went` are of course not supported. Yet it has been already proved that this simple analogy-based tech-

nique is able to detect morphological variants using examples of semantically and morphologically related words with a very good coverage and a high degree of accuracy in a context of computational terminology (*cf.* [13]). It is worth noting that it is moreover possible to identify the semantic link between these variants with excellent rates of success by annotating each rule with a label of semantic relation. Those are not used here: although it was shown that some semantic links are more relevant than others [15], we made the choice to take into account all the kinds of semantic links (synonymy, hyperonymy...) for query expansion.

3.2 Use for query expansion

In order to be operational, the previously presented detection method needs examples (*i.e.* morphologically related word couples). Such a supervised property is not well suited to a use within IR and does not correspond to the fully automatic aspect of the system in our requirements. To solve this problem, we substitute this supervision phase by a rustic technique that allows to constitute a set of word pairs that can be used as examples. This example-pair research proceeds in the following way:

1. randomly choose one document in the IRS collection;
2. form all the possible word pairs resulting from this document;
3. add to the example set couples w_1-w_2 such as $lcss(w_1,w_2) > l$;
4. return to step 1.

These steps are repeated until the resulting set of example-couples is large enough; in the experiments described in Section 4, 500 documents were analyzed. Notice that this operation also supposes that derivation and inflection are mainly done by prefixation and suffixation operations.

During this phase, it is necessary to avoid building word pairs that are not valid examples. The correct behavior of our analogy technique relies on it. That is why we have added two constraints. On the one hand, a minimal length of common substring l is fixed at a large enough value (in our experiments, $l = 7$ letters). Thus, the probability to aggregate two words that do not share any link is reduced. On the other hand, like what was already shown [5], variant search within a same document maximizes the probability that the obtained two words belong to the same domain.

At the end of this step, a set of morphologically related word-pair examples is available; analogy rule learning can be conducted (*cf.* Section 3.1). It is then possible to check if unknown word pairs are in derivation or inflection relation. In our case, we precisely want to retrieve query term variants. Each query term is thus confronted with each word of the document collection. If a formed pair is in analogy with one of the example-pairs, then the document word is used to enrich the query. In order to speed up treatments, analogy rules are in fact used in a generative way. Words are produced from the query terms according to prefixation and suffixation operations indicated in the morphological rules and are kept only if they appear in the index of the collection's terms. Rule

learning being made off-line, only the morphological variant search for query terms within the index is made on-line. Search complexity is $O(n)$ where n is the number of distinct terms in the collection. In our experiments, it takes some tenths of a second using a Pentium 1.5 GHz (512 MB). For instance, for the original query: `Ineffectiveness of U.S. Embargoes or Sanctions`, the result of the expansion will be: `ineffectiveness ineffective effectiveness effective ineffectively embargoes embargo embargoed embargoing sanctioning sanction sanctioned sanctions sanctionable`.

During expansion, only words directly related to query terms are added; the words themselves related to the extensions are not taken into account. This voluntary absence of transitivity aims at avoiding propagating errors, such as `reduce` → `produce` → `product` → `productions` → `production` → `reproduction`... In our experiments, an average of three variants is added to each query term. No manual filtering is performed; thus, some extensions are not relevant. The quality of the extensions is evaluated by measuring their impact on the IRS performances. An intrinsic evaluation, out of the context of use, turns out to be non relevant to estimate their impact.

4 Experimental results

This section details the evaluation of our query expansion method. We first present the various document collections that have been used (Section 4.1), and then successively describe different experiments: results obtained from French and English collections (Sections 4.2 and 4.3) are first reviewed; then the impact of the query length (Section 4.4) is analyzed; and finally the portability of our approach on other languages (Section 4.5) is evaluated.

4.1 Document collections

Three different document collections are used for our experiments. The evaluation of our method is carried out for English on a subset of the TIPSTER collection used in TREC. More precisely the Wall Street Journal subcollection made up of 175,000 documents and a set of 50 queries (from TREC-3) has been chosen. In order to emulate the usual short-query behavior, only the *title* field containing few words has been employed.

The evaluation on French is based on the INIST collection, made up of 30 queries and 163,000 documents, which are paper abstracts from various scientific disciplines. The portability of our method is controlled on the ELRA collection, made up of 30 queries and 3,511 documents that are questions/answers of the European Commission, available in French, English, German, Portuguese, Spanish, and Italian. Short queries (*title* field) are also chosen for these two collections, except in Section 4.4 where the impact of the query length is studied. The IRS used is LEMUR (`http://www.lemurproject.org`), implemented with the well-known Okapi-like (BM-25) weighting scheme.

4.2 French experiments

The first experiment is performed on the French INIST collection. In order to evaluate the added-value of query expansion with morphological variants detected with our method, results are computed with and without extensions. Standard IR measures are used for evaluation: precision and recall (computed for several threshold values), interpolated average precision (calculated at 11 recall points (IAP)), R-precision and non-interpolated average precision (MAP). For comparison, we also present the results obtained by applying on the same collection three traditional morphological tools: 2 French stemmers based on a set of fixed rules of desuffixation — one developed by Savoy [16], the other is an CPAN Perl adaptation of the Porter algorithm for French — and a French lemmatizer — part-of-speech tagger TREETAGGER (`http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger`).

In contrast with our method, these tools perform by conflation. Results are given in Table 1. Those considered as being not statistically significant (using paired t-test with the condition p-value < 0.05) are indicated in italic. The average length (number of words, stop-words included) of the queries ($|Q|$) is also indicated.

	Without extension	With extension (improvement %)	Stemming (Savoy) (improvement %)	Stemming (Porter) (improvement %)	Lemmatization (improvement %)
$\|Q\|$	5.46	16.03	5.2	5.2	5.17
MAP	14.85	18.45 (+24.29%)	*17.31 (+16.63%)*	15.89 (+7.00%)	17.82 (+20.07%)
IAP	16.89	19.93 (+17.97%)	*18.85 (+11.57%)*	17.69 (+5.92%)	19.72 (+16.73%)
R-Prec	17.99	21.63 (+20.24%)	*19.88 (+10.53%)*	18.77 (+4.34%)	*19.71 (+9.56%)*
P(10)	34.33	38.67 (+12.62%)	*36.67 (+6.80%)*	*34.33 (0%)*	39.67 (+15.53%)
P(20)	27.83	31.83 (+14.37%)	*29.00 (+4.19%)*	*26.50 (-4.78%)*	31.6 (+13.77%)
P(50)	18.33	21.27 (+16.00%)	*20.13 (+9.82%)*	*18.33 (0%)*	*20.87 (+13.82)*
P(100)	12.23	14.80 (+20.98%)	15.23 (+24.52%)	13.87 (+13.41%)	14.97 (+22.34%)
P(500)	3.88	4.80 (+23.71%)	4.55 (+17.18%)	4.56 (+17.53%)	4.47 (+15.29%)
P(1 000)	2.21	2.68 (+21.30%)	2.53 (+14.80%)	2.54 (+15.26%)	2.48 (+12.39%)
P(5 000)	0.56	0.67 (+20.38%)	0.63 (+13.47%)	0.62 (+11.81%)	0.64 (+15.14%)
R(10)	8.00	8.99 (+12.36%)	*8.45 (+5.64%)*	*8.19 (+2.38%)*	*9.04 (+13.02%)*
R(20)	12.33	14.50 (+17.59%)	*12.81 (+3.90%)*	*12.00 (-2.75%)*	*13.62 (+10.48%)*
R(50)	19.65	24.07 (+22.47%)	*20.78 (+5.74%)*	*19.71 (+0.31%)*	*21.56 (+9.71%)*
R(100)	26.85	32.87 (+22.41%)	31.32 (+16.64%)	29.28 (+9.05%)	31.58 (+17.59%)
R(500)	43.09	53.83 (+24.92%)	49.31 (+14.43%)	50.16 (+16.42%)	49.35 (+14.54%)
R(1 000)	48.43	59.45 (+22.74%)	55.27 (+14.12%)	56.94 (+17.57%)	55.03 (+13.62%)
R(5 000)	59.32	72.20 (+21.71%)	67.22 (+13.31%)	67.82 (+14.32%)	68.20 (+14.96%)

Table 1. Query expansion performances on the INIST collection

The reported figures show that, for each measure, our query expansion method obtains very good results that are all statistically significant. For most measures, query expansion appears not only more effective than stemming or

lemmatization, but also more stable since several results of the last two techniques have been found not statistically significant. It is also worth noting that improvements are distributed on every precision and recall thresholds (from 10 to 5000 documents). Thus, improvement does not only correspond to a re-ranking of relevant documents at the head of the result list but also to the obtaining of relevant documents that would not have been retrieved without query extensions.

4.3 English experiments

This experiment proposes to test if the good results obtained for French can also be observed on English. The preceding experiments are reiterated on the English TIPSTER collection. Table 2 shows the results obtained compared with those of a traditional research without extension, and of researches with lemmatization (using TREETAGGER) and stemming (based on Porter's stemmer [17]).

	Without extension	With extension (improvement %)	Stemming (improvement %)	Lemmatization (improvement %)
|Q|	6.5	16.6	6.48	6.48
MAP	23.15	27.18 (+17.40%)	28.09 (+21.33%)	*23.85 (+3.02%)*
IAP	25.02	29.44 (+17.65%)	29.71 (+18.73%)	*25.68 (+2.63%)*
R-Prec	27.52	31.96 (+16.15%)	32.66(+18.69%)	*27.68 (+0.59%)*
P(10)	39.60	47.00 (+18.68%)	45.00 (+13.63%)	*41.40 (+4.54%)*
P(20)	36.10	40.90 (+13.29%)	41.00 (+13.57%)	*36.60 (+1.38%)*
P(50)	28.28	32.80 (+15.98%)	31.72 (+12.16%)	*28.68 (+1.41)*
P(100)	21.44	25.50 (+18.93%)	23.76 (+10.82%)	*22.10 (+3.07%)*
P(500)	7.94	9.21 (+15.90%)	*8.72 (+9.81%)*	*8.35 (+5.13%)*
P(1 000)	4.66	5.37 (+15.08%)	*5.09 (+9.12%)*	*4.85 (+3.89%)*
P(5 000)	1.17	1.31 (+12.21%)	1.30 (+12.10%)	*1.22 (+4.09%)*
R(10)	10.20	11.18 (+9.52%)	*12.77 (+25.13%)*	*10.68 (+4.66%)*
R(20)	16.10	16.82 (+4.46%)	19.36 (+20.24%)	*17.79 (+10.45%)*
R(50)	29.68	32.41 (+9.18%)	32.84 (+10.65%)	*30.43 (+2.53%)*
R(100)	39.48	44.59 (+12.95%)	43.86 (+11.09%)	*41.61 (+5.40%)*
R(500)	61.11	67.68 (+10.74%)	67.82 (+10.98%)	*62.46 (+2.20%)*
R(1 000)	68.68	75.50 (+9.92%)	74.81 (+8.92%)	*69.28 (+0.87%)*
R(5 000)	80.59	87.66 (+8.77%)	87.22 (+8.22%)	81.20 (+0.75%)

Table 2. Query expansion performances on the TIPSTER collection

The results are positive. The contribution of our approach using query expansion on English is important since the observed gain on the IRS performances is ranging from 4 and 18% according to the measures. Although improvements are sometimes slightly lower than those observed for stemming, they are all statistically significant and constant for all measures. These observations highlight the robustness of our method, and its ability of self-adaptation to English. Other experiments are proposed in Section 4.5 in order to precisely evaluate its portability.

4.4 Impact of query length

In order to measure impact of the query length on our expansion method, the French experiment is repeated using the other fields of INIST queries so as to cope with increasingly long queries. In this collection, a query is associated with a set of concepts, each one being represented in a distinct field. The fields are added one by one to the original query (*i.e.* the *title* field). Figure 1 shows results according to the query length that is measured in number of words before expansion. The IRS performance is measured by non-interpolated average precision.

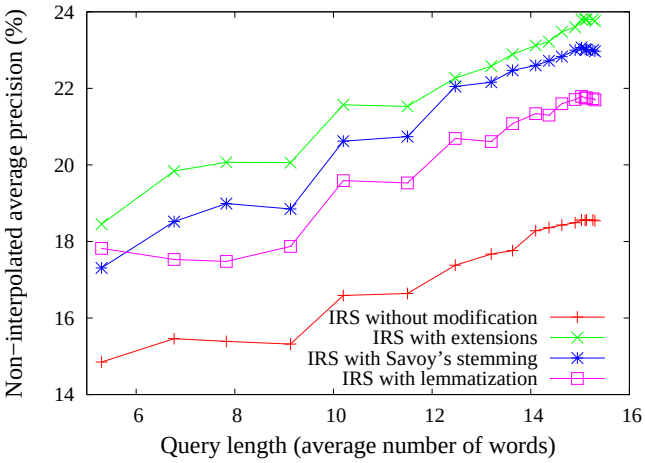

Fig. 1. Precision evolution according to the query length

Broadly speaking, these results prove the interest of taking into account the morphological variants whatever the query length and the morphological process. Among the three evaluated techniques, our approach for query expansion has yet shown better results than stemming and lemmatization.

4.5 Portability

The principal asset of our approach compared with other existing tools is its portability. It is supposed to be directly usable on any language whose morphology is done by prefixation and suffixation. In order to establish the truth of this assertion, Table 3 presents the results obtained on the ELRA collection for German, English, Spanish, French, Italian and Portuguese. For each language, variation (expressed as a percentage) compared to the same search without query extension is indicated.

Results are all very positive since improvements given by query extensions are ranging from 10 to 20% according to languages and measures. As for the other experiments, this gain concerns all precision and recall thresholds. However, for low

	Languages					
	German	English	Spanish	French	Italian	Portuguese
MAP	+16.25%	+17.52%	+10.03%	+11.89%	+10.45%	+9.69%
IAP	+15.93%	+16.66%	+8.70%	+10.99%	+9.79%	+9.25%
R-Prec	*+3.03%*	+10.23%	+7.97%	+9.43%	+10.23%	*+6.20%*
P(10)	*+10.68%*	+7.03%	*0%*	*+3.53%*	*+2.54%*	*0%*
P(20)	*+8.33%*	*+3.62%*	+7.41%	+6.85%	+11.15%	*+4.38%*
P(50)	*+6.69%*	+8.23%	+13.40%	+13.85%	+13.48%	+8.31%
P(100)	+9.54%	+14.31%	+16.76%	+16.24%	+18.98%	+14.24%
P(500)	+13.18%	+20.49%	+18.13%	+17.19%	+18.94%	+23.35%
P(1 000)	+12.97%	+21.60%	+20.32%	+18.26%	+22.13%	+24.64%
R(10)	*+6.82%*	+2.90%	*+1.88%*	*+5.43%*	*-0.67%*	*-0.47%*
R(20)	*+5.95%*	*+3.27%*	+7.40%	+7.36%	+7.82%	*+7.55%*
R(50)	*+11.12%*	+8.48%	+7.72%	+10.82%	+7.37%	*+6.21%*
R(100)	+11.87%	+13.23%	+10.14%	+10.11%	+8.93%	+9.39%
R(500)	+16.45%	+21.68%	+14.49%	+12.69%	+14.31%	+17.71%
R(1 000)	+18.15%	+20.93%	+17.38%	+13.20%	+18.35%	+19.23%

Table 3. Query extension performances on different languages

thresholds (10 to 50 documents), some not statistically significant figures seem to indicate results varying from one query to another. In contradiction with what is usually claimed in some studies, we would like to emphasize here some original remarks. First, query extension with morphological variants has more impact for English, which is generally seen as a morphologically poor language, than for so-called richer languages like Spanish, Italian... It also appears that it is German that benefits the most from the extension technique; this is most probably related to the fact that frequent word agglutinations are better taken into account by our approach (the pair `Menschenrechte-Menschenrechtsorganisation` for instance).

4.6 Discussion

Reported experimental results have shown that our approach for morphological variant detection and its use in query expansion significantly improves IRS performances. Its portability has been demonstrated: good results are observed even for languages that are traditionally found to be morphologically poor. These conclusions are distinct from those in several studies of the same field [18, for instance]. Moreover, contrary to what is sometimes observed in other studies, query length appears to have almost no impact on the results: improvement is constant and comparable for query lengths between 5 and 15 words.

Our method for enriching query terms with their morphological variants is nevertheless not perfect. Some terms actually related to query terms are not detected. For instance, for the English collection, our method did not allow to find the variant `hazard` related to the original term `hazardous` nor the term `paid`

related to the conjugated verb `pays` of the initial query. These errors are avoided by the methods based on resources, thus explaining why in some cases results obtained by Porter's stemmer are better. What is on the other hand more prejudicial for query extensions is that non relevant terms can be sometimes added. Concerning this last point, several cases can be distinguished. First, some detected words are not semantically related to the original term; the morphological link is fortuitous or no longer used, like `composition-exposition` for instance. Then, some polysemous terms cause errors that are difficult to avoid. For example, `production` and `reproduction`, detected as morphologically related, are indeed linked in `result production` and `result reproduction` but not in `fish reproduction`. To limit the impact of these errors, words that are themselves related to extensions are not used to enrich queries. This voluntary absence of transitivity aims at avoiding propagating errors. For this reason, the approach by expansion seems more flexible than the conflation method in which `production` and `reproduction` together with their variants would all be transformed to one single form.

5 Conclusion

In this paper, we have proposed a simple and original technique, relying on an analogy-based learning process, able to automatically detect morphological variants within documents and use them to expand query terms. This morphological expansion approach yields very good results. It rivals and even almost always outperforms results obtained with existing tools such as rule-based stemmer or lemmatizer, and also provides more stable performances. Moreover, contrary to most existing techniques, our method is fully unsupervised and thus can be used for many languages; in this paper, we successfully used it on English, French, German, Italian, Portuguese and Spanish test collections.

From a broader point of view, the conclusions of our experiments confirm those generally claimed in state-of-the-art studies since taking into account morphological variation always improves IRS performances, whatever the language. However, our results go against what is sometimes concluded. Indeed, we have shown that morphology can improve IRS performances whatever the query length or the morphological complexity of language providing that a flexible enough method is used.

This paper opens many future prospects that need further consideration. As further studies, there might be some added-values not to include all variants related to a query term but only retain the most relevant ones. Expansion decision could be thus based on the level of confidence of detected analogy (according to its productivity for instance) and on the importance of the query term directly or indirectly related. It would also be interesting to work on the weighting of the variants that are added to the original query and to integrate it in the ranking function. Reported results on studied languages require to checked and consolidated on other collections and to be extended to other languages. Finally, within a framework of translinguistic IR, a similar approach based on analogy used for translation of specialized terms is being studied.

References

1. Harman, D.: How Effective is Suffixing? Journal of the American Society for Information Science **42**(1) (1991) 7–15
2. Kraaij, W., Pohlmann, R.: Viewing Stemming as Recall Enhancement. In: Proceedings of the 19th ACM International Conference on Research and Development in Information Retrieval (SIGIR), Zürich, Switzerland (1996)
3. Hull, D.: Stemming Algorithms - A Case Study for Detailed Evaluation. Journal of the American Society of Information Science **47**(1) (1996) 70–84
4. Moreau, F., Sébillot, P.: Contributions des techniques du traitement automatique des langues à la recherche d'information. Research report, IRISA, Rennes, France (2005)
5. Xu, J., Croft, W.B.: Corpus-Based Stemming Using Cooccurrence of Word Variants. ACM Transactions on Information Systems **16**(1) (1998) 61–81
6. Gaussier, E.: Unsupervised Learning of Derivational Morphology from Inflectional Corpora. In: Proceedings of Workshop on Unsupervised Methods in Natural Language Learning, 37th Annual Meeting of the Association for Computational Linguistics (ACL), College Park, United-States (1999)
7. Vilares-Ferro, J., Cabrero, D., Alonso, M.A.: Applying Productive Derivational Morphology to Term Indexing of Spanish Texts. In Gelbukh, A., ed.: Computational Linguistics and Intelligent Text Processing. Springer-Verlag (2001) 336–348
8. Moulinier, I., McCulloh, J.A., Lund, E.: West Group at CLEF 2000: Non-English Monolingual Retrieval. In: Proceedings of the Workshop of Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal (2000)
9. Fuller, M., Zobel, J.: Conflation-Based Comparison of Stemming Algorithms. In: Proceedings of the 3rd Australian Document Computing Symposium, Sydney, Australia (1998)
10. Goldsmith, J.A., Higgins, D., Soglasnova, S.: Automatic Language-Specific Stemming in Information Retrieval. In: Proceedings of Workshop of Cross-Language Evaluation Forum (CLEF), Lisbon, Portugal (2001)
11. Frakes, W.B.: Stemming Algorithms. In Frakes, W.B., Baeza-Yates, R., eds.: Information Retrieval: Data Structures and Algorithms. Prentice Hall (1992) 131–160
12. Savoy, J.: Morphologie et recherche d'information. Technical report, Neuchâtel University, Neuchâtel, Switzerland (2002)
13. Claveau, V., L'Homme, M.C.: Structuring Terminology by Analogy Machine Learning. In: Proceedings of the International Conference on Terminology and Knowledge Engineering (TKE), Copenhagen, Denmark (2005)
14. Hathout, N.: Analogies morpho-synonymiques. Une méthode d'acquisition automatique de liens morphologiques à partir d'un dictionnaire de synonymes. In: Proceedings of 8ème conférence annuelle sur le traitement automatique des langues naturelles (TALN), Tours, France (2001)
15. Voorhees, E.M.: Query Expansion Using Lexical-Semantic Relations. In: Proceedings of the 17th ACM International Conference on Research and Development in Information Retrieval (SIGIR), Dublin, Ireland (1994)
16. Savoy, J.: A Stemming Procedure and Stopword List for General French Corpora. Journal of the American Society for Information Science **50**(10) (1999) 944–952
17. Porter, M.F.: An Algorithm for Suffix Stripping. Program **14**(1) (1980) 130–137
18. Arampatzis, A., Weide, T.P.V.D., Koster, C.H.A., Van Bommel, P. In: Linguistically Motivated Information Retrieval. Volume 69. M. Dekker, New York, United-States (2000) 201–222