Unsupervised and semi-supervised morphological analysis for Information Retrieval in the biomedical domain

Vincent Claveau

Unsupervised and semi-supervised morphological analysis for Information Retrieval in the biomedical domain

Vincent CLAVEAU
IRISA-CNRS, Campus de Beaulieu
F35042 Rennes, France
Vincent.Claveau@irisa.fr

Abstract
In the biomedical field, the key to access information is the use of specialized terms. However, in most of Indo-European languages, these terms are complex morphological structures. The aim of the presented work is to identify the various meaningful components of these terms and use this analysis to improve biomedical Information Retrieval. We present an approach combining an automatic alignment using a pivot language, and an analogical learning that allows an accurate morphological analysis of terms. These morphological analysis are used to improve the indexing of medical documents. The experiments reported in this paper show the validity of this approach with a 10% improvement in MAP over a standard IR system.

Keywords: Morphology, biomedical terminology, alignment, analogical learning, morpho-semantic indexing, biomedical information retrieval.

1 Introduction

In the biomedical domain, terminologies are the keystone of many applications. They are used for structuring the knowledge as well as retrieving and formalizing information contained in documents. For instance, the well-known MeSH®(Medical Subject Headings) `www.nlm.nih.gov/mesh` terminology is developed to index the very popular PubMED database (`www.pubmed.gov`). In most Indo-European languages, biomedical terms also have interesting inner characteristics in that they tend to be complex morphological constructions. Indeed, they are often resulting from the composition of several Greek or Latin roots, prefixes, and suffixes. This morphological complexity is an important point to take into account for basic operations like handling, understanding, translating or building semantic relationships between these terms, and furthermore for higher level applications like machine translation or, as we demonstrate in this paper, Information Retrieval (IR).

In this paper, we investigate the development of morphological resources and show how a biomedical Information Retrieval task can benefit from such resources. More precisely, we present several techniques aiming at breaking up a term into its morphological components, namely morphs[1], while labeling these morphs with some semantic information. To the contrary of existing studies (Deléger et al., 2008; Markó et al., 2005a, for example) which are chiefly based on human expertise, the techniques proposed here rely on unsupervised or semi-supervised approaches.

The original idea at the heart of our approach is to use the multilingualism of existing terminological databases. We exploit Japanese as a pivot language, or more precisely terms written in Kanjis, to help decompose the terms of other languages into morphs and associate them with the corresponding Kanjis, in a fully automatic way. Thus, Kanjis play the role of a semantic representation for morphs. The main advantage of Kanjis in this respect is that Japanese terms can be seen as a concatenation of elementary independent words that may even be found in general language dictionaries. For example, the term photochemotherapy can be translated in Japanese by 光化学法; splitting and aligning these two terms gives: photo ↔ 光 ('light'), chemo ↔ 化学 ('chemistry', 'medicine'), therapy ↔ 法 ('therapy'). As it is shown here, each morph is associated with Kanjis that may be used as descriptors more convenient to index a document than the term itself. In particular, we demonstrate here how such correspondences between morphs and Kanjis can be exploited, in different ways, to improve the results of an IR system.

The morphological analysis, and the document indexing that it allows, thus chiefly relies on an alignment step between morphs and Kanjis. This alignment is performed with an original technique, suited to the biomedical domain and based on a *Forward-Backward* algorithm and on analogy learning. In this paper, different versions of this alignment approach are proposed, either fully unsupervised, or semi-supervised.

The paper is structured as follows. After a review of the related studies in Section 2, we present the unsupervised alignment technique, its semi-supervised variants and their results respectively in Sections 3, 4 and 5. Then, in Section 6, the use of the obtained morphological decompositions in an IR framework is explained. Evaluations, conducted on a biomedical IR test collection, are detailed in Section 7.

[1] Following Mel'čuk (2006), we distinguish between morphs, elementary linguistic signs (segments), and morphemes, equivalence classes with identical signified and close signifiers.

2 Related work

Many studies have used morphology for terminological analysis. This is more particularly the case in the biomedical domain where terminologies are central to many applications and where terms are constructed by operations like neo-classical composition (e.g. *chemotherapy*, built from the Greek pseudo-word *chemo*, and *therapy*), which are very regular, and very productive. Unfortunately, no comprehensive database of morphs with semantic information is available, and splitting a term into morphs is still an issue. One can distinguish two views of the use of morphology as a tool for term (or word) analysis. In the lexematic view, relations between terms rely on the word form, but without the need to split them into morphs (Grabar and Zweigenbaum, 2002; Claveau and L'Homme, 2005; Hathout, 2009). Beside this implicit use of morphology, the morphemic view chiefly relies on splitting the term into morphs as a first step. Many studies have been made in this framework. They either rely on partially manual approaches in which an expert gives morphs and combination rules (Deléger et al., 2008; Markó et al., 2005a) or heuristics (Baud et al., 1999), or on more automatic approaches. The latter usually try to find recurrent letter patterns in word lists as morph-candidates (Kurimo et al., 2010). But such techniques cannot associate a semantic meaning with these morphs. To our knowledge, our approach is the first to make the most of a pivot language to perform an automatic morphological analysis, as we propose in this study. It can be explained by three peculiarities of the biomedical domain: the morphology of its terms is known to be very regular, with few exceptions, the morphological composition (producing compounds) is very fertile, and there exists many multilingual terminologies.

From a more technical point of view, the use of a bilingual terminology also evokes studies in transliteration, particularly Katakana or Arabic (Tsuji et al., 2002; Knight and Graehl, 1998, for example), or in translation. In this framework, Morin and Daille (2010) propose to map complex terms written in Kanjis with French ones, by using morphological rules. Yet, here again, these rules are to be given by an expert, and this study only concerns a special case of derivation. Moreover such an approach cannot handle neo-classical compounds. In other studies, translation methods for biomedical terms which considers terms as simple sequences of letters have been proposed (Claveau, 2009, inter alia). Such approaches share some similarities with the one presented here: they require aligning the words at the letter level. In most cases, this is performed with 1-1 alignment algorithms (one character, possibly empty, of the source language word is aligned with one another character of the target language word), but in recent work about phonetization (Jiampojamarn et al., 2007), authors have shown that the interest of *many-to-many* alignment.

Concerning the use of morphological processing in Information Retrieval, the literature is more important (Moreau and Sébillot, 2005, for a panorama). Although the results depend on numerous factors (language, morphological tool, size of collection, domain...), there is a broad consensus about the benefit of simple processes like *stemming* (or, rarer in IR, lemmatization): such tools are available in many languages, conceptually simple, and they usually improve the results of IR systems. It is noteworthy that the only morphological phenomena addressed by these tools are inflection and derivation. As they mostly perform simple operations on the prefix and suffix of the words, morphological composition remains out of their scope. Yet, many authors have noted the importance of clever tokenisation based on morpheme for the biomedical indexing (Jiang and Zhai, 2007; Trieschnigg et al., 2007), but without proposing effective solutions. Recently, advanced morphological tools developed in the framework of MorphoChallenge have been applied to IR problems (Kurimo

et al., 2009). Here again, the authors have observed an improvement for some languages, such as Finnish which is highly compositional, but results on English were significantly lower than when using a simple stemmer. In that respect, the good results presented in the next sections confirm the interest of our approach.

3 Analogy for Alignment

As it was previously explained, our morphological decomposition technique relies on the alignment of terms with their translation in a pivot language (Japanese, Kanjis). Thus, this approach makes a strong parallelism assumption: the term in Kanjis must be built in the same way than the one in the studied language. This hypothesis may appear as unrealistic, but the results presented hereafter show that it is reasonable. It is noteworthy that the choice of Kanjis as pivot is not fortuitous. Kanjis do not have morphology, their form is therefore invariable whatever their position in the term. One only needs to test a few combinations when trying to segment a term made of Kanjis, compared with a term written with the latin alphabet. Kanjis are independent of the Greek and Latin roots used in most European languages. This prevents learning irrelevant regularities based on common etymology. Last, a segment of a Kanji term is most of the time a valid Kanji term by itself (to the contrary of morphs). It is thus possible to use dictionaries to access their meaning. These different reasons make Kanji-based Japanese a very good pivot when compared to other alternatives.

Our alignment technique is mainly based on an *Expectation-Maximization* (EM) algorithm that we briefly present in the next sub-section (Jiampojamarn et al., 2007, for more details and examples of its use). The second sub-section explains the modification made to this standard algorithm so that it can naturally and automatically handle morphological variation, which is a phenomenon inherent to our morph splitting problem.

3.1 EM Alignment

The alignment algorithm at the heart of our approach is standard: it is a *Baum-Welch* algorithm, extended to map symbol sub-sequences and not only 1-1 alignments. In our case, it takes as input French terms with their Kanji translations, taken from a multilingual terminology for instance. The maximum length of the sub-sequences of letters and Kanjis considered for alignment are parametrized by $maxX$ and $maxY$.

For each term pair (x^T, y^V) to be aligned (T and V being the lengths of the terms in letters or Kanjis), the EM algorithm (see Algorithm 1) proceeds as follows. It first computes the partial counts of every possible mapping between sub-sequences of Kanjis and letters (*Expectation* step). These counts are stored in table γ, and are then used to estimate the alignment probabilities in table δ (*Maximization* step).

The *Expectation* step relies on a *forward-backward* approach (Algorithm 4): it computes the *forward* probabilities α and *backward* probabilities β. For each position t, v in the terms, $\alpha_{t,v}$ is the sum of the probabilities of all the possible alignments of (x_1^t, y_1^v), that is, from the beginning of the terms to the current position, according to the current alignment probabilities in δ (cf. Algorithm 2). $\beta_{t,v}$ is computed in a similar way by considering (x_t^T, y_v^V). These probabilities are then used to re-estimate the counts in γ. In this version of the EM algorithm, the *Maximization* (Algorithm 3) simply consists in computing the δ alignment probabilities by normalizing the counts in γ.

Algorithm 1 *EM Algorithm*

Input: list of pairs (x^T, y^V), $maxX$, $maxY$
while changes in δ **do**
 initialization of γ to 0
 for all pair (x^T, y^V) **do**
  γ = Expectation(x^T, y^V, $maxX$, $maxY$, γ)
 δ = Maximization(γ)
return δ

Algorithm 2 *Forward-many2many*

Input: (x^T, y^V), $maxX$, $maxY$
$\alpha_{0,0} := 1$
for $t = 0...T$ **do**
 for $v = 0...V$ **do**
  if $(t > 0 \vee v > 0)$ **then**
   $\alpha_{t,v} = 0$
  if $(v > 0 \wedge t > 0)$ **then**
   for $i = 1...maxX$ s.t. $t - i \geq 0$ **do**
    for $j = 1...maxY$ s.t. $v - j \geq 0$ **do**
     $\alpha_{t,v} \mathrel{+}= \delta(x^t_{t-i+1}, y^v_{v-j+1})\alpha_{t-i,v-j}$
return α

Algorithm 3 *Maximization*

Input: γ
for all sub-sequence a s.t. $\gamma(a, \cdot) > 0$ **do**
 for all sub-sequence b s.t. $\gamma(a, b) > 0$ **do**
  $\delta(a, b) = \frac{\gamma(a,b)}{\Sigma_x \gamma(a,x)}$
return δ

Algorithm 4 *Expectation*

Input: (x^T, y^V), $maxX$, $maxY$, γ
α := Forward-many2many(x^T, y^V, $maxX$, $maxY$)
β := Backward-many2many(x^T, y^V, $maxX$, $maxY$)
if $\alpha_{T,V} > 0$ **then**
 for $t = 1...T$ **do**
  for $v = 1...V$ **do**
   for $i = 1...maxX$ s.t. $t - i \geq 0$ **do**
    for $j = 1...maxY$ s.t. $v - j \geq 0$ **do**
     $\gamma(x^t_{t-i+1}, y^v_{v-j+1}) \mathrel{+}= \frac{\alpha_{t-i,v-j}\delta(x^t_{t-i+1}, y^v_{v-j+1})\beta_{t,v}}{\alpha^{T,V}}$
return γ

The EM process is repeated until the probabilities δ are stable. When the convergence is reached, the alignment simply consists in finding the mapping that maximizes $\alpha(T, V)$. In addition to this resulting alignment, we also store the final alignment probabilities δ, which are used to split unseen terms (cf. Section 6.2).

This technique is not very different from the one used in statistical translation. Yet, some particularities are worth noting: this approach allows us to handle *fertility*, that is the capacity to align from or to empty substrings (for lack of space, it does not appear in the above simplified version); conversely, *distortion*, that is reordering of morphs, cannot be handled easily without major changes in this algorithm.

3.2 Automatic morphological normalisation

The maximization step simply compute the translation probabilities of a Kanji sequence into a letter sequence. For example, for the Kanji 菌 ('*bacteria*'), there may exist one entry in δ associating it with bactérie, one with bactério (as in bactério/lyse) and another one with bactéri (in myco/bactéri/ose), each with a certain probability. This dispersion of probabilities, which is of course harmful for the algorithm, is caused by morphemic variation: bactério, bactérie, and bactéri are 3 morphs of the same morpheme, and we would like their probabilities to reinforce each other. The adaptation we propose aims at making the maximization phase able to automatically group the different morphs belonging to a same morpheme. To achieve this goal, we use a simple but well suited technique relying on formal analogical calculus.

3.2.1 Analogy

An analogy is a relation between 4 elements that we note: $a : b :: c : d$ which can be read *a is for b what c is for d* (Lepage, 2000, for more details about analogies). Analogies have been

used in many NLP studies, especially for translation of sentences (Lepage, 2000) or terms (Langlais and Patry, 2007; Langlais et al., 2008). Analogies are also a key component in the previously mentioned work on terminology structuring (Claveau and L'Homme, 2005). We rely on this latter work to formalize our normalization problem. In our framework, one possible analogy may be: dermato : dermo :: hémato : hémo. Knowing that dermato and dermo belong to a same morpheme, one can infer that this is the case for hémato and hémo. Such an analogy, build on the graphemic representation of words, is said a formal analogy.

After Stroppa and Yvon (2005), formal analogies can be defined in terms of factorizations. Let us note $\overleftarrow{\oplus}$ the (non-commutative) concatenation operator at the right ($abc\overleftarrow{\oplus}d = abcd$), and $\overleftarrow{\ominus}$ its associated string subtraction operator ($abc\overleftarrow{\oplus}d\overleftarrow{\ominus}d = abc\overleftarrow{\ominus}c\overleftarrow{\oplus}c = abc$), and similarly for $\overrightarrow{\oplus}$ and $\overrightarrow{\ominus}$ operating at the left of the first argument. Let a be a string (a term in our case) over an alphabet Σ, a factorization of a, noted f_a, is a sequence of n factors $f_a = (f_a^1, ..., f_a^n)$, such that $a = f_a^1 \overleftarrow{\oplus} f_a^2 \overleftarrow{\oplus} ... \overleftarrow{\oplus} f_a^n$. A formal analogy can be defined by as:

Definition 1 $\forall (a,b,c,d) \in \Sigma, [a : b :: c : d]$ *iff there exist factorizations* $(f_a, f_b, f_c, f_d) \in (\Sigma^{*n})^4$ *of* (a,b,c,d) *such that,* $\forall i \in [1,n], (f_b^i, f_c^i) \in \left\{(f_a^i, f_d^i), (f_d^i, f_a^i)\right\}$. *The smallest* n *for which this definition holds is called the degree of the analogy.*

For most European languages, as French and English, morphology is mostly concerned with prefixation and suffixation. Thus, we are looking for formal analogies of degree at most 3 (ie, 3 factors: prefix $\oplus$ base $\oplus$ suffix). In our approach, such analogies are searched by trying to build a rule rewriting the prefixes and the suffixes to move from dermato to dermo and to check that this rule also applies to hémato-hémo. The base is considered as the longest common sub-string (lcss) between the 2 words. In the previous example, the rewriting rule r would be:
$r = \text{lcss}(\text{morph}_1,\text{morph}_2) \overleftarrow{\ominus}$ ato $\overleftarrow{\oplus}$ o.
This rule makes it possible to rewrite dermato into dermo and hémato into hémo; thus, hémato,hémo is in analogy with dermato,dermo.

3.2.2 Using analogy for normalization

The main problem is that we do not have examples of morphs that are known a priori to be related (like dermato and dermo in the previous example). Thus, we use a simple bootstrapping technique: if two morphs are stored in γ as possible translations of the same Kanji sequence, and if these two morphs share a sub-string longer than a certain threshold, then we assume that they both belong to the same morpheme. From these bootstrap pairs, we build the prefixation and suffixation rewriting rules allowing us to detect analogies, and thus to group pairs of morphs (which can be very short, unlike the bootstrapping pairs). The more a rule is found, the more certain it will be. Therefore, we keep all the analogical rules generated at each iteration along with their number of occurrence, and we only apply the most frequently found ones. The whole process is thus completely automatic.

This new *Maximization* step is summarized in Algorithm 5. It ensures that all the morphs supposed to belong to the same morpheme have equal and reinforced alignment probabilities.

4 Semi-supervision and bootstrapping

The approach described above can be considered as unsupervised since no example of alignment or decomposition is provided. Yet, in some cases, expert knowledge is available

Algorithm 5 *Maximization* with analogical normalization

Input: γ
for all sub-sequence a s.t. $\gamma(a, \cdot) > 0$ **do**
 for all m_1, m_2 s.t. $\gamma(a, m_1) > 0 \wedge \gamma(a, m_2) > 0 \wedge$ lcss$(m_1, m_2) >$ threshold **do**
  build the prefixation and suffixation rule r for m_1, m_2
  increment the score of r
 for all sub-sequence b s.t. $\gamma(a, b) > 0$ **do**
  build the set $\mathscr{M}$ of all morphs associated to b with the help of the n most frequent analogical rules from the previous iteration

$$\delta(a, b) = \frac{\sum_{c \in \mathscr{M}} \gamma(a, c)}{\sum_{x} \gamma(a, x)}$$

return δ

and can be used to add some supervision to this morpho-semantic alignment task. It is important to note that this human intervention can be more or less costly and requires more or less expertise. While manually building a full morphological resource from scratch is a tedious task, providing light information during the alignment process is more accessible. In the following sub-sections, we propose different strategies to improve the unsupervised alignment process, implementing different trade-off between human cost and performance.

4.1 Active alignment

In analogy with active learning (Settles, 2009), the first semi-supervised strategy that we propose is active alignment. Its principle is the following: a human expert, the oracle, adds information about the pair during the expectation step. Different information can be used: decomposition of the Kanji term, of the English term, partial or full alignment. The interest of these pieces of information is twofold. First, it helps reduce the complexity, by avoiding to consider certain alignments in the pair. Secondly, it possibly improves the final results by reinforcing the probability for this pair on a few realistic alignments.

From an algorithmic point-of-view, the implementation is straightforward. The information provided by the oracle is interpreted as a set of constraints on the possible decomposition and/or alignments used to compute $\alpha_{t,v}$, $\beta_{t,v}$, and $\gamma(x^t_{t-i+1}, y^v_{v-j+1})$ in the Forward, Backward, and Expectation steps respectively.

As for active learning, one can think of different strategies to choose the pairs to be presented to the oracle (Settles, 2009). The goal is of course to find the strategy that best *helps* the alignment algorithm and thus results in a faster convergence and/or better performance. In this paper, two strategies are experimented and both will ask the oracle for providing full alignment of a pair. The first one is a random strategy and serves as baseline: at each iteration, randomly selected pairs are proposed to the oracle. The second strategy is a difficulty-driven one: at each iteration, pairs with many equi-probable alignments are proposed to the oracle. In practice, the equi-probability is measured with the probability gathered at the previous iteration.

4.2 Bootstrapping the oracle

The previous active alignment approach can also be used, to some extent, without human intervention. Indeed, when processing multiple languages, it is sometimes possible to make the most of the existing probabilities from a language L1 to help estimate the alignment

probabilities for the new language L2. Several ways to use these alignment probabilities of other languages can be imagined. In the experiments presented below, we used a simple approach. We adopt the difficulty-driven presented above, but instead of presenting the pairs (of L2) to a human oracle, they are processed as follows:
– if the Kanji term is known in the L1 alignment pairs, the closest term of L1 (edit distance) among the known translations is chosen; this L1-term and the Kanji term are then aligned and the alignment is propagated to the L2 pair. This step is done by representing alignment as characters in the L1 term and by adding the marks in L2 so that it minimizes the edit distance.
– if the Kanji is not known, the most probable alignment, based on the previous iteration probabilities is proposed.

5 Experiments

5.1 Evaluation Data

The data used in the experiments presented below come from the UMLS MetaThesaurus (Tuttle et al., 1990). The MetaThesaurus groups several terminologies for several languages and associates to each term a concept identifier (CUI). The CUI are language independent and thus make it easy to build lists of terms in the spotted language with their Japanese equivalents. In this paper, we present experiments for English and French. In both cases, we only considered Japanese terms composed of Kanjis, and only simple (one-word) French or English terms. About 14,000 English-Kanjis pairs and 8,000 French-Kanjis ones are formed this way. An ending mark (';') is added to each French or English term.

We randomly selected 1,600 pairs for French and 500 for English in order to evaluate the performance of our alignment technique. These pairs have been aligned manually to serve as gold standard.

5.2 Alignment results

In the different experiments presented below, the performance is evaluated in terms of precision: an alignment is counted as correct only if all the components of the pair are correctly aligned (thus, it is equivalent to the sentence error rate in standard machine translation).

For each pair, the EM algorithm indicates the probability of the proposed alignment. Therefore, it is possible to only consider alignments having a probability greater than a given threshold. By varying this threshold, we can compute a precision according to the number of terms aligned. Figures 1 and 2 respectively present the results obtained on the French and English test pairs. We indicate the curves produced by the EM algorithm with and without our morphemic normalization. For comparison purpose, we also report the results of giza++ (Och and Ney, 2003), a reference tool in machine translation. The different IBM models and sets of parameters available in giza++ were tested; the results reported are the best ones (obtained with IBM model 4 without distortion).

As expected, the interest of the morphemic normalization appears clearly in these two experiments; in the worst case (that is, when all the terms are kept for alignment), it yields a 70% precision for French and 80% for English. Indeed, the normalization brings a 10% improvement whatever the number of aligned pairs. Normalization also has an interest

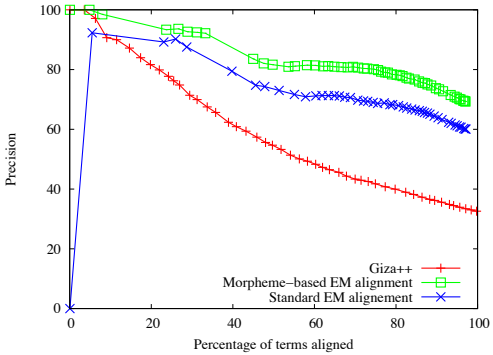

Figure 1: Precision of French-Kanji alignment according to the number of test pairs aligned

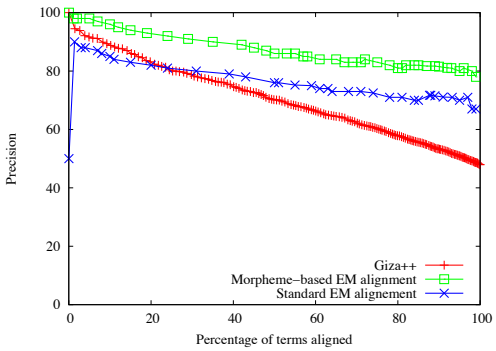

Figure 2: Precision of English-Kanji alignment according to the number of test pairs aligned

in terms of complexity since it reduces the needed iteration number by improving the convergence of the EM estimation.

A manual examination of the results shows that most of the errors are caused by the falsification of our hypothesis: some French-Japanese pairs cannot be decomposed in a similar way. For example, the French term anxiolytiques (anxiolytics) is translated by a sequence of Kanjis meaning literally 'drugs for depression'. Among these errors, some pairs imply terms that are not neo-classical compounds in French, Japanese or both (eg. méninges (meninges) is translated by 膜 'brain membrane'). Other errors are caused by a lack of training data: some morphs or sequences only appear once, or only combined with another morph, which mislead the segmentation.

5.3 Semi-supervised approaches

To compare the two active alignment strategies and the bootstrapped one presented in Section 4, we are interested in the alignment performance and the convergence speed. Thus, Figure 4 presents, for these three semi-supervised approaches together with the original unsupervised version, the precision on the French dataset after different number of iterations of the EM loop. For comparison purposes, we also report the results of the original unsupervised version. For a fair comparison between the unsupervised versions requiring the oracle, the same amount of pairs (20) is presented to the oracle at each iteration. The bootstrapped version is based on the alignment probabilities gathered from the English-Kanji alignment task. Of course, to prevent any bias, none of the pairs processed by the oracle are used as test pairs.

As expected, the three active alignment strategies converge faster than the original one, but also yield better overall results. Adding information at each iteration clearly helps the alignment to produce more relevant association between Kanjis and morphs. The difficulty-driven strategy obtains good results. In particular, it outperforms the random strategy after a few iterations. Before that, when iteration < 5, the probabilities collected are not reliable enough to propose interesting pairs to the oracle, and the oracle even seems to decrease the performance of the alignment. The combination of this strategy with the

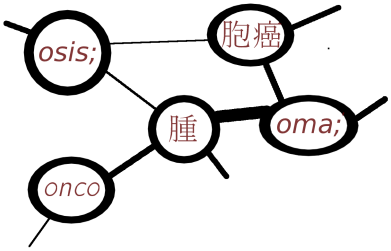

Figure 3: Morpheme-Kanji graph

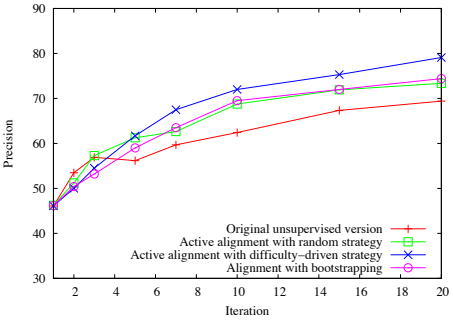

Figure 4: Precision of semi-supervised and unsupervised alignments according to the number of iterations

bootstrapping also performs better than the original version, but not as well as the real oracle one. This result can be explained by the fact that the difficulty-driven strategy asks for pairs that are difficult to align, not only for the studied language (here, French), but also for the one serving as bootstrap (English). Nonetheless, it still remains a good option to improve the results without any human supervision.

6 Morpho-semantic analysis for Information Retrieval

As it was previously said, biomedical Information Retrieval (IR) has some important characteristics due to the use of specialized terms. In that respect, taking into account rich morphological information has already been proved useful, but only with hand-crafted resources (Markó et al., 2005b). Beside the intrinsic evaluation of our approach presented in the previous section, we evaluate its use in a large scale IR experiment in English.

6.1 Morpho-semantic graphs

Once all the terms are aligned, one can study the recurrent correspondences between English morphs and Kanjis. The more a morph is aligned with a sequence of Kanjis, the more they are 'semantically' related. All these links can be represented as a graph: the vertices represent Kanjis and morphemes (i.e a set of morphs grouped during the analogical step of the alignment), and the edges are weighted according to the number of times that a particular morpheme is aligned with a Kanji sequence among the 14,000 training pairs from the umls. Figure 3 shows a toy example of such a graph. The size of the edge lines is proportional to the associated weight.

This representation allows us to shed light on different types of semantic relations between the morphemes. It is done by exploring the neighborhood of each morpheme: each vertex receives an amount of energy which is propagated to the connected vertices proportionally to the edge's weight. For instance, Figure 5 presents the closest morphemes reached, in the form of tag clouds, for the French morpheme ome (oma in English, a suffix for cancer-related terms). The size and color represent the energy that reach the neighboring morpheme vertices. The reached vertices are expected to be conceptually related and to exhibit synonym or quasi-synonym morphemes of the suffix ome. It is interesting to see that other

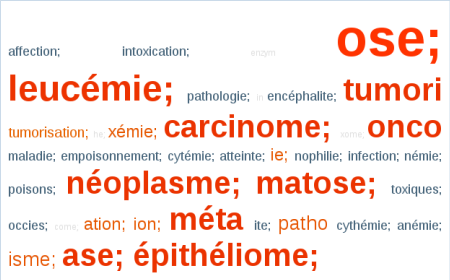

Figure 5: Cloud of 1st order affinities for French morpheme ome

Figure 6: Cloud of 2nd order affinities for French morpheme gastro

related suffixes are found, but also prefixes like onco.

The alignment and the segmentation produced by our algorithm also make it possible to study the co-occurrences of morphemes in English (or French) terms. One can study first-order affinities (which morphemes are frequently associated with other morphemes) and, more interesting, second order affinities (morphemes sharing the same co-occurring morphemes). The second-order affinity allows us to group morpheme according to their paradigm. For instance, the tag cloud in Figure 6 illustrates the morphemes associated with gastro (morpheme for stomach) according to this second order affinity. Most of the morphemes identify organs, and the closest ones are for biologically close organs.

This information of different nature makes it possible to identify relationships between terms, or build synonyms, or explore the termbase using these morphological elements. Yet, to our knowledge, such specialized morpho-semantic resources do not exist. It makes a direct evaluation of these three different uses of the alignment results not possible. But in the remaining of this paper, we propose to evaluate them in an Information Retrieval framework.

6.2 Morphemic representation for Information Retrieval

In order to integrate the morphological information in an IR system, we adopt a simple indexing representation: the documents are considered as bags-of-morphemes and words. The morphemes are those obtained by decomposing biomedical terms, and for some experiments those associated to the former ones as second order affinities. The goal is of course to be able to associate a query containing stomachalgia with a document containing gastrodynia.

Thus, when indexing the collection, terms are decomposed. Two cases may occur: the term is either known as it appears in the alignment pairs, either not. In the first case, we simply use the decomposition produced by the alignment algorithm. In the second case, we make the most of the δ probabilities to generate the most probable translation. To do so, we use a simple approach: the translation probability in δ are used by a Viterbi algorithm to generate the most probable Kanji translation. We do not use language modeling. It is

important to note that this translation process produces at the same time the decomposition of the initial term by associating each morph with its Kanji translation. This translation process is thus equivalent to the desired morpho-semantic analysis for unknown terms (i.e. absent from the pair list used in the alignment step).

In these two cases, another alignment product is also used: the analogical rewriting rules collected at the last alignment iteration. They allow us to detect the morphs belonging to a same morpheme. Such information makes it possible to match a query containing hemo with a document containing haemo, hemato or even emia;.

A *baseline* system and four indexing systems using the morphological information are proposed. They all rely on a standard IR approach, namely a vector space model with an Okapi BM25 weighting scheme (Robertson et al., 1998, for details, see) and a tokenizer similar to Terrier's one (Ounis et al., 2006); the standard values of the BM25 parameters b, k_1, k_3 have been kept. The *baseline* system performs a standard indexing of the documents with a Porter stemming (Porter, 1980).
1 – The first morphologically-enhanced system is morpheme-based. It simply considers the morphemes produced by the decomposition of the term in the documents (and queries) as words to index (the original terms are also kept for indexing, along with their morphemes). The morpheme weights take the decomposition probability into account; it is defined as the product of this probability and the BM-25 weight.
2 – The second system is Kanji-based. Here again, the terms of the documents are decomposed, and the closest Kanjis are used as indexing words. These Kanjis are those identified in the neighborhood of the morphs produced when decomposing the terms (see Section 6.1).
3 – The third system adopts the same morpheme-based representation as the first system, but expands the queries with first-order affinities of their morphemes. The morphemes used as expansion are weighted according to their proximity in the graph and the weight of the morphemes that they expand.
4 – The last system is similar to the third one but uses the second-order affinities to expand the queries.

7 Biomedical IR experiments

7.1 Experimental setting

For the following experiments, we use the dataset built for the filtering track of the TREC-9 conference. This dataset is itself based on the document collection ohsumed, which is composed of about 350,000 medline abstracts. In addition to this, about 4,000 queries and the corresponding relevance judgments were developed for TREC-9. The queries are composed of several fields: the subject, which a MeSH term, and a definition of this term. Although the collection was initially built for evaluating filtering systems, here we use this dataset as a standard IR collection, and only consider the subject field as the query.

7.2 Results

Table 1 presents the results of the baseline IR system and the system based on the morphological analysis. The performance of the systems are evaluated using the standard IR evaluation measures: we compute precision on top 5, 10... 1,000 documents (P@x), mean average precision (MAP), interpolated average precision (IAP) and the R-precision

	baseline (BM-25 + stemming)	System 1 morpheme-based	System 2 Kanji-based
MAP	29.93	33.94 (+13.4 %)	32.76 (+9.5 %)
IAP	31.74	35.55 (+12 %)	34.49 (+8.6 %)
R-prec	35.28	39.64 (+12.3 %)	38.59 (+9.4 %)
P@5	69.87	73.45 (+5.1 %)	71.70 (+2.6 %)
P@10	67.99	71.31 (+4.9 %)	*69.65 (+2.4 %)*
P@50	52.98	56.90 (+7.4 %)	55.24 (+4.3 %)
P@100	40.86	44.56 (+9.1 %)	43.39 (+6.2 %)
P@500	15.11	17.21 (+13.9 %)	16.92 (+12 %)
P@1000	8.72	10.10 (+15.86 %)	9.95 (+14.2 %)

Table 1: Performance of the morpheme and Kanji based IR systems on the OHSUMED collection, with the TREC queries

(R-prec).In order to assess if the differences between the two systems are statistically significant, we run a Wilcoxon test ($p = 0.05$) (Hull, 1993); those differences with the *baseline* that are not judged statistically significant are italicized.

The morpheme-based system, only relying on decomposing term and grouping its morph into morpheme, yields very good results with a 13% MAP gain. As expected, decomposing the terms improves more specifically the performance at the end of the retrieved document list (P@100 and higher), since it makes it possible to retrieve relevant documents even though they do not contain the exact terms of the queries. The Kanji-based system yields very similar results. Although the Kanjis were expected to be a more generic representation, no additional gain is obtained. In practice, in some queries, the Kanjis are too generic to capture the specific meaning expected or bring no additional information compared with the original morphemes. Moreover, no selection is performed on the morpheme to be translated into Kanjis, and some Kanjis have properties (document frequencies) that highly differ from the source morpheme since they can be translation of different morphemes. A weighting scheme taking the initial document frequencies into account seems an important foreseen work.

Table 2 presents the results of the two last systems, based on query expansion. The two expansion-based systems have more contrasted results. On the one hand, expanding the queries with first-order affinities gives good results; although it yields a lower precision at the top of list than system 1, it obtains a slightly better recall. On the other hand, second-order affinities produce bad results compared with morphological decomposition alone. The affinities added to the query, most of the time, break the specificity of the information asked; it makes the system retrieve too much non-related documents.

8 Conclusion and future work

The original idea of making the most of another language like Japanese in order to help the morphological decomposition and analysis of compounds offers many new opportunities to automatically handle biomedical terms. The new alignment approach based on analogy that we propose takes the particularities of the data into account, and also offers different ways to balance quality, convergence speed and human intervention through the semi-supervised approaches proposed. The high quality results obtained allow us to address IR problems

	baseline (BM-25 + stemming)	System with 1st order affinities	System with 2nd order affinities
MAP	29.93	34.40 (+14.9 %)	28.74 (-3.9 %)
IAP	31.74	36.63 (+15.4 %)	30.80 (-2.9 %)
R-prec	35.28	39.92 (+13.2 %)	34.38 (-2.6 %)
P@5	69.87	71.76 (+2.7 %)	*68.65 (-1.7 %)*
P@10	67.99	70.46 (+3.6 %)	*66.20 (-2.6 %)*
P@50	52.98	56.30 (+6.7 %)	50.50 (-4.68 %)
P@100	40.86	44.69 (+9.4 %)	39.07 (-4.38 %)
P@500	15.11	17.98 (+18.9 %)	*15.01 (-0.64 %)*
P@1000	8.72	10.56 (+21.1 %)	*8.77 +0.66 %)*

Table 2: Performance of the expansion based IR systems on the OHSUMED collection, with the TREC queries.

caused by the morphological complexity of the biomedical terminology that could not be addressed with usual IR tools like stemmers. In this respect, our concerns about the role of morphology to access information are similar to existing studies (Markó et al., 2005a; Deléger et al., 2008), but to our knowledge, we are the first to propose an automatic process, directly available for many languages. Of course, our approach chiefly relies on the availability of multilingual terminologies, but such databases like UMLS are now widely developed, on the contrary of usable morphological resources.

Many perspectives are foreseen from this work. First, from a technical point of view, we plan to consider more complex segmentation than the linear one we implemented. Indeed, the syntactic properties of the Kanjis (some of them expect an agent or object), could help to better structure the different morphemes. One could also exploit the semantic relations between Kanjis that can be easily found in general Japanese dictionaries.

Concerning the analysis aspects illustrated in the last section, many possibilities are also under consideration. As the links between morphs that we produce are not typed, the use of heuristics (such as string inclusion used by Grabar and Zweigenbaum (2002)) or techniques from distributional analysis could provide useful additional information to better characterize the relationships. Yet, the problem of evaluating this type of work arises, especially the ground truth construction, since such resources do not exist. The IR setting used in this paper could be used again, possibly with more biomedical collections such as the TREC Genomics ones (Hersch and Voorhees, 2009).

Finally, an adaptation of these principles for complex terms is under study. The main difficulty in this case is to manage the reordering of the words composing these terms, and thus manage the distortion in the alignment algorithm. This issue is important for IR since these multiword terms are known to occurs with many variants and thus prevent to match queries and documents with different variants of the same term (Nenadic et al., 2005).

Acknowledgments

This work was partly funded by OSEO, the French innovation agency, in the framework of the Quæro project.

References

Baud, R., Rassinoux, A.-M., Ruch, P., Lovis, C., and Scherrer, J.-R. (1999). The power and limits of a rule-based morpho-syntactic parser. In *Proceedings of the 1999 Annual Symposium of the American Medical Informatics Association. Transforming Health Care through Informatics. AMIA*, pages 22–26, Washington, DC, USA.

Claveau, V. (2009). Translation of biomedical terms by inferring rewriting rules. In Prince, V. and Roche, M., editors, *Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration*. IGI - Global.

Claveau, V. and L'Homme, M.-C. (2005). Structuring terminology by analogy-based machine learning. In *Proc. of the 7th International Conference on Terminology and Knowledge Engineering, TKE'05*, Copenhaguen, Denmark.

Deléger, L., Namer, F., and Zweigenbaum, P. (2008). Morphosemantic parsing of medical compound words: Transferring a french analyzer to english. *International Journal of Medical Informatics*, 78(Supplement 1):48–55.

Grabar, N. and Zweigenbaum, P. (2002). Lexically-based terminology structuring: Some inherent limits. In *Proc. of International Workshop on Computational Terminology, COMPUTERM*, Taipei, Taiwan.

Hathout, N. (2009). Acquisition morphologique à partir d'un dictionnaire informatisé. In *Actes de la 16e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles*, Senlis, France.

Hersch, W. and Voorhees, E. (2009). TREC genomics special issue overview. *Information Retrieval*, 12(1):1–15.

Hull, D. (1993). Using Statistical Testing in the Evaluation of Retrieval Experiments. In *Proceedings of the 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'93*, Pittsburgh, États-Unis.

Jiampojamarn, S., Kondrak, G., , and Sherif, T. (2007). Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In *Proc. of the conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, New York, USA.

Jiang, J. and Zhai, C. (2007). An empirical study of tokenization strategies for biomedical information retrieval. *Information Retrieval*, 10(4-5):341–363.

Knight, K. and Graehl, J. (1998). Machine transliteration. *Computational Linguistics*, 24(4):599–612.

Kurimo, M., Creutz, M., and Turunen, V. (2009). Morpho challenge evaluation by information retrieval experiments. In *Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*, CLEF'08, pages 991–998, Berlin, Heidelberg. Springer-Verlag.

Kurimo, M., Virpioja, S., and Turunen, V. T. (2010). *(Eds), Proceedings of the MorphoChallenge 2010*. Espoo, Finlande.

Langlais, P. and Patry, A. (2007). Translating unknown words by analogical learning. In *Proc. of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 877–886, Prague, Czech Republic.

Langlais, P., Yvon, F., and Zweigenbaum, P. (2008). Translating medical words by analogy. In *Proc. of the workshop on Intelligent Data Analysis in bioMedicine and Pharmacology (IDAMAP) 2008*, Washington, DC.

Lepage, Y. (2000). Languages of analogical strings. In *Proc. of the 18th conference on Computational linguistics, COLING'00*, Universität des Saarlandes, Saarbrücken, Germany.

Markó, K., Schulz, S., and Han, U. (2005a). Morphosaurus - design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. *Methods of Information in Medicine*, 44(4).

Markó, K., Schulz, S., Medelyan, O., and Hahn, U. (2005b). Bootstrapping dictionaries for cross-language information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*, Salvador, Brésil.

Mel'čuk, I. (2006). *Aspects of the Theory of Morphology.* Trends in Linguistics. Studies and Monographs. Mouton de Gruyter, Berlin.

Moreau, F. and Sébillot, P. (2005). Contributions des techniques du traitement automatique des langues à la recherche d'information. Technical Report 1690, IRISA.

Morin, E. and Daille, B. (2010). Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation (LRE)*, 44.

Nenadic, G., Spasic, I., , and Ananiadou, S. (2005). Mining biomedical abstracts: What's in a term? In *Proceedings of the IJCNLP 2004*, volume 3248 of *Lecture Notes in Artificial Intelligence*, pages 797–806. Springer-Verlag.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., and Lioma, C. (2006). Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006).*

Porter, M. F. (1980). An Algorithm for Suffix Stripping. *Program*, 14:130–137.

Robertson, S. E., Walker, S., and Hancock-Beaulieu, M. (1998). Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive. In *Proceedings of the 7th Text Retrieval Conference, TREC-7*, pages 199–210.

Settles, B. (2009). Active learning literature survey. Technical report, University of Wisconsin–Madison.

Stroppa, N. and Yvon, F. (2005). An analogical learner for morphological analysis. In *Proceeedings of the 9th CoNLL*, pages 120–127, Ann Arbor, MI, USA.

Trieschnigg, D., Kraaij, W., and de Jong, F. (2007). The influence of basic tokenization on biomedical document retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 803–804, New York, NY, USA. ACM.

Tsuji, K., Daille, B., and Kageura, K. (2002). Extracting French-Japanese word pairs from bilingual corpora based on transliteration rules. In *Proc. of the 3rd International Conference on Language Resources and Evaluation, LREC'02*, Las Palmas de Gran Canaria, Spain.

Tuttle, M., Sherertz, D., Olson, N., Erlbaum, M., Sperzel, D., Fuller, L., and Neslon, S. (1990). Using meta-1 – the 1st version of the UMLS metathesaurus. In *Proc. of the 14th annual Symposium on Computer Applications in Medical Care (SCAMC)*, pages 131–135, Washington, USA.