Morphological Analysis of Biomedical Terminology with Analogy-Based Alignment

Vincent Claveau
IRISA-CNRS
Vincent.Claveau@irisa.fr

Ewa Kijak
IRISA-Univ. Rennes 1
Ewa.Kijak@irisa.fr

Abstract

In the biomedical domain, many terms are neoclassical compounds (composed of several Greek or Latin roots). The study of their morphology is important for numerous applications since it makes it possible to structure, translate, retrieve them efficiently...

In this paper, we propose an original yet fruitful approach to carry out this morphological analysis by relying on Japanese, more precisely on terms written in kanjis, as a pivot language. In order to do so, we have developed a specially crafted alignment algorithm relying on analogy learning. Aligning terms with their kanji-based counterparts provides at the same time a decomposition of the term into morphs, and a kanji label for each morph.

Evaluated on a dataset of French terms, our approach yields a precision greater than 70% and shows its relevance compared with existing techniques. We also illustrate the interest of this approach through two direct applications of the produced alignments: translating unknown terms and discovering relationships between morphs for terminological structuring.

1 Introduction

In many domains, accessing the information in documents or collections of documents is guided by the use of well-defined terms, which form a terminology of the domain. This is particularly true in the biomedical domain where there is a long tradition of terminologies development for structuring the knowledge as well as accessing it. An example is the MeSH (Medical Subject Headings) www.nlm.nih.gov/mesh terminology which is used to index the very popular PubMED database (www.pubmed.gov). Knowing how to handle these terms, understanding them, translating them or building semantic relationships between them are thus essential operations for applications like enrichment of bilingual lexicons, or more generally machine translation, information retrieval...

In this framework, the work presented here is interested in the morphology of simple terms from the biomedical domain as a basis for the terminological analysis. More precisely, we present a technique aiming at breaking up a term into its morphological components, namely morphs, and associating in the same time semantic knowledge to these morphs. Note that in this paper, we distinguish morphs, elementary linguistic signs (segments), from morphemes, equivalence classes with identical signified and close significants (Mel'čuk, 2006). We therefore tackle the same issue already raised in some studies (Deléger et al., 2008; Markó et al., 2005, for example), but we try here to suppress the costly human operations required by these studies.

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 decomposing 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 words which are easier to find in general language dictionaries. For example, the term photochimiotherapy can be translated in Japanese by 光化学療法; splitting and aligning these two terms gives: photo ↔ 光 ('light'), chimio ↔ 化学 ('chemistry'), thérapie ↔ 療法 ('therapy'). Our approach chiefly relies

Proceedings of Recent Advances in Natural Language Processing, pages 347–354,
Hissar, Bulgaria, 12-14 September 2011.

on the hypothesis that the composition of terms in kanjis is the same than those of English or French simple terms. This hypothesis can be seen as peremptory, but the results presented below in this paper show that it is a reasonable hypothesis. Finally, our approach provides, at the same time 1) an effective way to split terms into morphs, 2) the semantic meaning of each morph as they are actually used.

This morphological analysis thus relies on an essential step which consists in aligning English or French terms with Japanese ones taken from a multilingual terminology. To do so, we propose a new alignment technique, particularly suited to this kind of data, which mixes *Forward-Backward* algorithm and analogy-based machine learning. After a presentation of related work in Section 2, either in terms of applications or methods, we describe this alignment technique in Section 3. Results of the morphological analysis are detailed in Section 4. In Section 5, we illustrate the interest of such analysis through two applications. The first one shows that our technique can be used to translate and analyse never-seen-before terms. The second application illustrates how the morphs and their obtained semantic labels can be used from a terminological point of view.

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, for example). 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, as the already mentioned ones (Deléger et al., 2008; Markó et al., 2005) in which morphs and combination rules are provided by an expert, or on more automatic approaches. The latter usually try to find recurrent letter patterns as morph-candidate. But such techniques cannot associate a semantic meaning with these morphs. To our knowledge, no existing work makes the most of a pivot language to perform an automatic morphological analysis, as we propose in this study.

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, let us cite the work of Morin and Daille (2010). They 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). Even if the goal is different here, such approaches share some similarities with the one presented here. Indeed, they all require aligning the words at the letter level. In most cases, this is performed with 1-1 alignment algorithm, that is, algorithm only capable to align one character, which can be empty, of the source language word with one another character of the target language word. Yet, in recent work about phonetization (Jiampojamarn et al., 2007), authors have shown that *many-to-many* alignment could yield interesting results.

3 Analogy for alignment

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 2): 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 4). $\beta_{t,v}$ 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 $\delta$ do
  initialization of $\gamma$ to 0
  for all pair $(x^T, y^V)$ do
    $\gamma$ = Expectation($x^T, y^V, maxX, maxY, \gamma$)
  $\delta$ = Maximization($\gamma$)
return $\delta$
```

Algorithm 2 *Expectation*

```
Input: $(x^T, y^V)$, $maxX, maxY, \gamma$
$\alpha :=$ Forward-many2many( $x^T, y^V, maxX, maxY$ )
$\beta :=$ 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
          $\gamma(x_{t-i+1}^t, y_{v-j+1}^v) \mathrel{+}= \frac{\alpha_{t-i,v-j}\delta(x_{t-i+1}^t, y_{v-j+1}^v)\beta_{t,v}}{\alpha^{T,V}}$
return $\gamma$
```

Algorithm 3 *Maximization*

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

Algorithm 4 *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
          $\alpha_{t,v} \mathrel{+}= \delta(x_{t-i+1}^t, y_{v-j+1}^v)\alpha_{t-i,v-j}$
return $\alpha$
```

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 5.1).

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 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 $a = f_a^1 \oplus f_a^2 \oplus ... \oplus f_a^n$, where $\oplus$ denotes the concatenation operator. 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.*

As for most European languages, French 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 = lcss(morph$_1$,morph$_2$) $\ominus$ ato $\oplus$ 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.

Algorithm 5 *Maximization* with analogical normalization

```
Input: $\gamma$
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 $\mathcal{M}$ of all morphs associated to $b$ with the help of the $n$ most frequent analogical rules from the previous iteration
    $\delta(a, b) = \frac{\sum_{c \in \mathcal{M}} \gamma(a, c)}{\sum_x \gamma(a, x)}$
return $\delta$
```

4 Experiments

4.1 Evaluation Data

The data used for our experiments are extracted from the UMLS MetaThesaurus (Tuttle et al., 1990), which group several terminologies for several languages. In the MetaThesaurus, each term is associated with a concept identifier (CUI) which facilitates the Japanese/French pairs extraction. We only consider Japanese terms composed of kanjis, and only simple (one-word) French terms. About 8,000 pairs are formed this way. An ending mark (';') is added to each term.

We randomly selected 1,600 pairs among these 8,000 pairs in order to evaluate the performance of our alignment technique. These 1,600 pairs have been aligned manually to serve as gold standard.

4.2 Alignment results

We evaluate our approach in terms of precision: an alignment is considered 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. Figure 1 presents the results obtained on the 1,600 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).

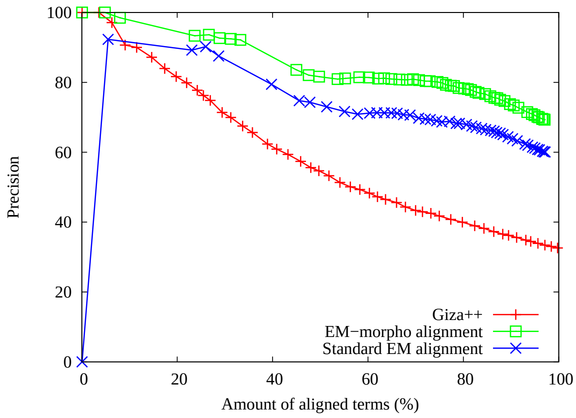

Figure 1: Precision of alignment according to the number of test pairs aligned

As expected, the interest of the morphemic normalization appears clearly in this figure; it yields a 70% precision in the worst case (that is, when all the terms are kept for alignment). Indeed, the normalization brings a 10% improvement whatever the number of aligned pairs.

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 neoclassical compounds in French, Japanese or both (eg. méninges (meninges) is translated by 脳膜 '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 Using the morph/kanji alignments

In this section, we present two ways of exploiting the results produced by our morphological analysis technique. The first one aims at translating unseen terms and the second one aims at structuring terminologies by finding related terms or morphs.

5.1 Translating and analysing unknown terms

The alignment technique that we propose can be used as a first step to translate an unknown term (i.e a term absent from the training data of our alignment algorithm). Translating terms has already been tackled in several studies, mostly to reduce the *out-of-vocabulary* errors in machine translation tasks. Most of these studies look for translations in textual resources: parallel or comparable corpora (Chiao and Zweigenbaum, 2002; Fung and Yee, 1998), Web (Lu et al., 2005). Others have considered this problem without external resources; in this case, the approach rely on the similarities between the terms in the two languages (cognates) (Schulz et al., 2004, for example), or on the similarities of rewriting operations to go from one term to its equivalent in the other language (Langlais and Patry, 2007; Claveau, 2009). Our work falls into this category.

In the experiment reported here, we translate French terms into Japanese. In practice, we use the probabilities from δ to generate the most probable translation. The approach is straightforward: the morph translation probabilities in δ are used in a Viterbi-like algorithm; thus, we do not use a language model in addition to the translation model.

It is important to note that this translation process also produce the alignment of the source term into its translation. As a result, it also segments the initial term and label them with the corresponding kanjis. Therefore, it corresponds to the morphosemantic analysis of the unknown term.

For the need of this experiment, 128 terms and their kanji translations have been selected at random to form the test set (of course, they have been removed from the alignment training set). These French terms are translated as explained above with the help of the *delta* table, and the generated translations are compared with the expected ones.

Reference	UMLS	Web
Correctly translated (and segmented)	58	82
Incorrectly translated (or segmented)	34	10
Not translated	36	36

Table 1: Unknown terms translation results

The results of this small experiment are presented in Table 1. 58 of 128 terms, that is 45%, have been correctly translated and segmented. There are two types of errors: either a wrong translation has been proposed (it concerns 34 terms), or no translation was found (36 terms). When examining these untranslated terms, we find without any surprise that they are either words which are not neo-classical compounds, or compounds having one or several components that do not appear in the training data of the alignment algorithm. The precision on the terms for which a translation is proposed is thus 63%; this result is very promising given the simplicity of our implementation of the translation. It is also worth noting that, among the errors, most of the proposed translations are correct paraphrase, absent from the UMLS but attested on the Web in bio-medical Japanese websites; with this wider reference, the precision on translated terms reaches 89 %.

5.2 Morph analysis

Once all the terms are aligned, one can study the recurrent correspondences between French morphs and kanjis. These correspondences can be shed into light through different techniques: Galois lattices (kanjis would be the intention and morph the extension), in a distributional analysis manner, or by analysing the kanji-morph graph with small-world, connected components... In this paper we propose to use such a graph representation: 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 8,000 training pairs from the UMLS. Figure 2 shows a small excerpt of the resulting graph. The size of the edge lines is proportional to the associated weight.

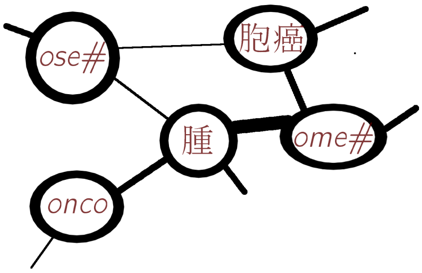

Figure 2: Morpheme-kanji graph

This representation allows us to easily explore the different kinds of neighbourhood of a morpheme: each vertex receives an amount of energy which is propagated to the connected vertices proportionally to the edge's weight. Figures 3 and 4 respectively present the kanjis (manually translated in English in this figure) and the 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 neighbouring kanji (respectively the morpheme) vertices. The reached vertices are expected to be conceptually related and to exhibit translation relations or synonymy, as one can see in these examples. Thus, Figure 3 represents a sort of semantic profile of the morpheme ome, in which the kanjis are used as semantic tags, while Figure 4 proposes synonyms and quasi-synonyms morphemes of the suffix ome. It is interesting to see that other related suffixes are found, but also prefixes like onco.

The alignment and the segmentation produced by our algorithm also make it possible to study

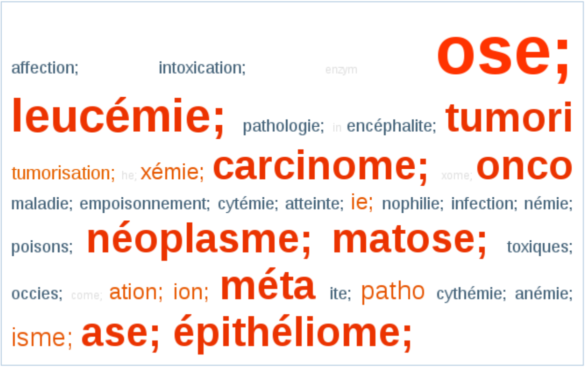

Figure 4: Morpheme cloud for morpheme ome

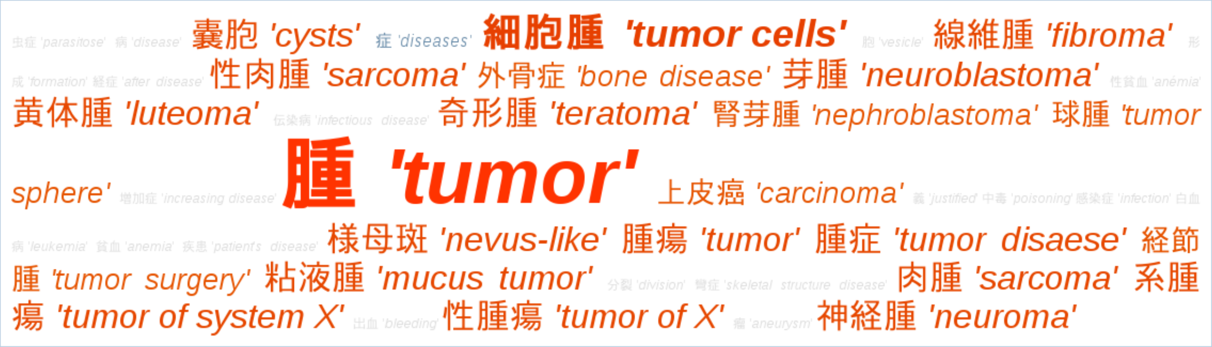
虫症 'parasitose' 病 'disease' 嚢胞 'cysts' 症 'diseases' 細胞腫 'tumor cells' 胞 'vesicle' 線維腫 'fibroma' 形
成 'formation' 経症 'after disease' 性肉腫 'sarcoma' 外骨症 'bone disease' 芽腫 'neuroblastoma' 性貧血 'anémia'
黄体腫 'luteoma' 伝染病 'infectious disease' 奇形腫 'teratoma' 腎芽腫 'nephroblastoma' 球腫 'tumor
sphere' 増加症 'increasing disease' 腫 'tumor' 上皮癌 'carcinoma' 義 'justified' 中毒 'poisoning' 感染症 'infection' 白血
病 'leukemia' 貧血 'anemia' 疾患 'patients disease' 様母斑 'nevus-like' 腫瘍 'tumor' 腫症 'tumor disaese' 経節
腫 'tumor surgery' 粘液腫 'mucus tumor' 分裂 'division' 彎症 'skeletal structure disease' 肉腫 'sarcoma' 系腫
瘍 'tumor of system X' 出血 'bleeding' 性腫瘍 'tumor of X' 瘤 'aneurysm' 神経腫 'neuroma'

Figure 3: Kanji cloud for ome

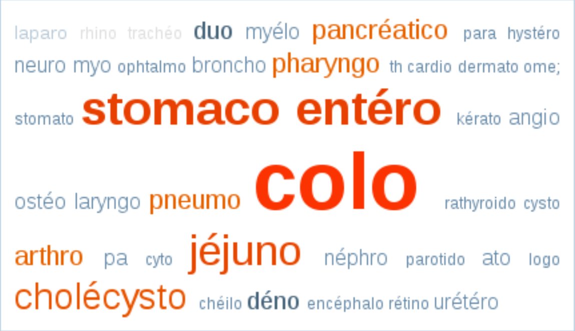
laparo rhino trachéo duo myélo pancréatico para hystéro
neuro myo ophtalmo broncho pharyngo th cardio dermato ome;
stomato stomaco entéro kérato angio
ostéo laryngo pneumo colo rathyroido cysto
arthro pa cyto jéjuno néphro parotido ato logo
cholécysto chéilo déno encéphalo rétino urétéro

Figure 5: Morpheme cloud for gastro second-order affinities

the co-occurrences of morphemes in 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 5 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 (other benefits from these alignments can be derived) 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 impossible.

6 Conclusion

The original idea of making the most of another language like Japanese in order to help the morphologically 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 in order to yield high quality results. Since this whole process is entirely automatic, it makes it possible to overcome the limits of terminological systems, like the one of Deléger et al. (2008), which heavily rely on manually populating a morphological database.

Many perspectives are foreseen for 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.

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.

References

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for French-English translations in comparable medical corpora. Journal of the American Medical Informatics Association, 8(suppl).

Vincent Claveau and Marie-Claude L'Homme. 2005. Structuring terminology by analogy-based machine learning. In Proc. of the 7th International Conference on Terminology and Knowledge Engineering, TKE'05, Copenhaguen, Denmark.

Vincent Claveau. 2009. Translation of biomedical terms by inferring rewriting rules. In Violaine Prince and Mathieu Roche, editors, Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration. IGI - Global.

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

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from non-parallel, comparable texts. In Proc. of 36th Annual Meeting of the Association for Computational Linguistics ACL, Montréal, Canada.

Natalia Grabar and Pierre Zweigenbaum. 2002. Lexically-based terminology structuring: Some inherent limits. In Proc. of International Workshop on Computational Terminology, COMPUTERM, Taipei, Taiwan.

Sittichai Jiampojamarn, Grzegorz Kondrak, , and Tarek Sherif. 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.

Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. Computational Linguistics, 24(4):599–612.

Philippe Langlais and Alexandre Patry. 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, June.

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

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

Wen-Hsiang Lu, Shih-Jui Lin, Yi-Che Chan, and Kuan-Hsi Chen. 2005. Semi-automatic construction of the Chinese-English MeSH using web-based term translation method. In Proc. of AMIA annual symposium.

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

Igor Mel'čuk. 2006. Aspects of the Theory of Morphology. Trends in Linguistics. Studies and Monographs. Mouton de Gruyter, Berlin, March.

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

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1):19–51.

Stefan Schulz, Kornel Markó, Eduardo Sbrissia, Percy Nohama, and Udo Hahn. 2004. Cognate Mapping - A Heuristic Strategy for the Semi-Supervised Acquisition of a Spanish Lexicon from a Portuguese Seed Lexicon. In Proc. of the 20th International Conference on Computational Linguistics, COLING'04, Geneva, Switzerland.

Nicolas Stroppa and François Yvon. 2005. An analogical learner for morphological analysis. In Proceeedings of the 9th CoNLL, pages 120–127, Ann Arbor, MI, USA.

Keita Tsuji, Béatrice Daille, and Kyo Kageura. 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.

Mark Tuttle, David Sherertz, Nels Olson, Mark Erlbaum, David Sperzel, Lloyd Fuller, and Stuart Neslon. 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.