Translating Medical Words by Analogy

Philippe Langlais
Université de Montréal, Dept I.R.O.,
H3C 3J7 Montreal, QC, Canada
felipe@iro.umontreal.ca

François Yvon
Université Paris-Sud 11 &
CNRS, LIMSI, F-91403 Orsay
yvon@limsi.fr

Pierre Zweigenbaum
CNRS, LIMSI, F-91403 Orsay &
Inalco, ERTIM, F-75012 Paris
pz@limsi.fr

Abstract

Term translation has become a recurring need in medical informatics. This creates an interest for robust methods which can translate medical words in various languages. We propose a novel, analogy-based method to generate word translations. It relies on a partial bilingual lexicon and solves bilingual analogical equations to create candidate translations. To evaluate the potential of this method, we tested it on several datasets for five language pairs (10 translation directions). At best could the approach propose a correct translation for up to 67.9% of the input words, with a precision of up to 80.2% depending on the number of selected candidates. We compare it to previous methods including word alignment in parallel corpora and edit distance to a list of words, and show that these methods can complement each other.

1 Introduction

Term translation has become a recurring need in medical informatics. With the expansion of multilingual societies, the same information needs to be described in several languages (see, *e.g.*, the development of Spanish in the United States). Cross-language information retrieval enables users to search in a language and obtain documents in a different language, generally relying on query translation [Hersh and Donohoe, 1998]. To keep pace with the evolution of international medical terminologies, local terms must be provided for new or changed concepts (generally originally described in English). For instance, MeSH thesaurus updates are translated worldwide into multiple languages [Nelson *et al.*, 2004]. The SNOMED CT nomenclature is being translated into several languages, including French (`http://sl.infoway-inforoute.ca/content/dispPage.asp?cw_page=snomedct_e#8`). This requires a continuous effort to find the most suitable translations for new terms, often linked to new concepts.

A number of Natural Language Processing methods have been proposed to help find translations of words and terms. As suggested for another lexical task [McDonald, 1993], they can be divided into internal and external methods.

Internal methods rely on the words themselves, *i.e.*, their morphology. Observing that many European languages have similar medical words of Greek and Latin origins, a transducer may be designed to translate, *e.g.*, English words into French words. Such a transducer may be learnt by induction on a sample of *{word, translation}* pairs [Claveau and Zweigenbaum, 2005]. Morphological knowledge can be leveraged to decompose words into morphemes or *subwords* so that translation is only needed for this smaller set of items [Markó *et al.*, 2006; Namer and Baud, 2005]. A limitation however is that the set of subwords must be specified to begin with.

External methods take advantage of the context in which words occur. In a multilingual context, a suitable context is given by parallel text corpora: sets of bitexts, *i.e.*, *{text, translation}* pairs. Aligning sentences and words in these parallel texts performs a kind of reverse engineering which tracks back the lexical decisions and knowledge of human translators, down to word translations. This kind of method has been tested on multilingual medical terminologies such as ICD-10 [Baud *et al.*, 1998; Nyström *et al.*, 2006] and on documents extracted from a multilingual web site [Deléger *et al.*, 2006]. A well-known limitation is the relative scarcity of parallel corpora, which motivates the recourse to the more readily available "comparable" corpora: corpora of texts which are generally in two different languages, but cover the same theme (*e.g.*, smoking cessation). However, finding word translations in comparable corpora [Chiao and Zweigenbaum, 2002] is much more difficult than in parallel corpora.

The present work explores the use of a different internal method: analogical learning [Lepage, 1998; Stroppa and Yvon, 2005]. As the above-mentioned methods of this type [Claveau and Zweigenbaum, 2005; Markó *et al.*, 2006; Namer and Baud, 2005], it is trained on an initial bilingual lexicon and relies on the formal similarity of medical words in some languages to propose new translations; in contrast to external methods, it can generate translations for unseen words. In this paper, we examine how analogical learning performs on medical words. We evaluate it on a similar dataset as earlier, comparable work [Claveau and Zweigenbaum, 2005], and study its complementarity to an external method such as the above [Deléger *et al.*, 2006] and to a non-generative internal method based on edit distance. We also investigate the viability of the approach for translating from and into various languages, including morphologically rich languages such as Finnish.

This paper is organized as follows. We first introduce

the principles of analogical learning on which our system relies, and describe the corpora used to test the method. We then present its evaluation, with a comparison to two other methods. We discuss the results and suggest an articulation of these different types of method, summarize our contribution and conclude with perspectives for further work.

2 Methods

2.1 Formal Analogy

A *proportional analogy*, or analogy for short, is a relation between four items noted $[x : y = z : t]$ which reads as "*x* is to *y* as *z* is to *t*". Among proportional analogies, we distinguish *formal analogies*, that is, those we can identify at a graphemic level, such as [*adrenergic beta-agonists*, *adrenergic beta-antagonists*, *adrenergic alpha-agonists*, *adrenergic alpha-antagonists*]. Formal analogies can be defined in terms of factorization [Stroppa and Yvon, 2005].

Définition 2.1 *Let x be a string over an alphabet Σ, a* factorization *of x, noted f_X, is a sequence of n factors $f_X = (f_X^1, \ldots, f_X^n)$, such that $x = f_X^1 \bullet f_X^2 \bullet f_X^n$, where $\bullet$ denotes the concatenation operator.*

We thus define a formal analogy as follows. Intuitively, this definition states that (x, y, z, t) are made up of a common set of alternating substrings.

Définition 2.2 *$\forall (x, y, z, t) \in \Sigma^{\star^4}$, $[x : y = z : t]$ **iff** there exists factorizations $(f_X, f_Y, f_Z, f_T) \in (\Sigma^{\star^d})^4$ of (x, y, z, t) such that, $\forall i \in [1, d]$, $(f_Y^i, f_Z^i) \in \{(f_X^i, f_T^i), (f_T^i, f_X^i)\}$.*

It is routine to check that it captures the example analogy introduced above, based on the following factorizations:

$$\begin{array}{rcl} f_x & \equiv & \text{(adrenergic bet, a-agonists)} \\ f_y & \equiv & \text{(adrenergic bet, a-antagonists)} \\ f_z & \equiv & \text{(adrenergic alph, a-agonists)} \\ f_t & \equiv & \text{(adrenergic alph, a-antagonists)} \end{array}$$

In the sequel, we call an *analogical equation* an analogy where one item (usually the fourth) is missing and we note it $[x : y = z : ?]$. Lepage [1998] proposes a method to solve analogical equations, that is, to generate the missing fourth item. Stroppa and Yvon [2005] describe a generalization of this algorithm, which accounts for the definition of formal analogy we gave above. More precisely, they show that the set of solutions to an analogical equation is a rational language; that is, we can build a finite-state machine to recognize them.

We implemented such a solver in this work, the details of which are beyond the scope of this paper. It is important to realize though that very often, there is not one single solution to an analogical equation, but many of them. For instance, *spondylite*, *ispondylte*, *ispndylote*, *spndyloite*, and *itespondyl* are 5 of the 110 solutions to the equation [*chondropathie* : *spondylopathie* = *chondrite* : *?*]. Though all these forms verify this equation, only the first is a French word (and the *natural* solution to the equation).

2.2 Analogical learning

Let $\mathcal{L} = \{(i, o) \mid i \in \mathcal{I}, o \in \mathcal{O}\}$ be a learning set of observations, where $\mathcal{I}$ (resp. $\mathcal{O}$) is the set of possible forms of the input (resp. output) linguistic system of the application. We denote $I(u)$ (resp. $O(u)$) the projection of u into the input (resp. output) space; that is, if $u = (i, o)$, then $I(u) \equiv i$ and $O(u) \equiv o$. For an incomplete observation $u = (i, ?)$, the inference procedure consists in:

1. building $\mathcal{E}_{\mathcal{I}}(u) = \{(x, y, z) \in \mathcal{L}^3 \mid [I(x) : I(y) = I(z) : I(u)]\}$, the set of input triplets that define an analogy with $I(u)$.
2. building $\mathcal{E}_{\mathcal{O}}(u) = \{o \in \mathcal{O} \mid \exists (x, y, z) \in \mathcal{E}_{\mathcal{I}}(u)$ s.t. $[O(x) : O(y) = O(z) : o]\}$ the set of solutions to the equations obtained by projecting the triplets of $\mathcal{E}_{\mathcal{I}}(u)$ into the output space.
3. selecting candidates among $\mathcal{E}_{\mathcal{O}}(u)$.

Since the first two steps of this inference procedure are generating candidate solutions, we call them the *generator*. Step 3 is responsible for selecting candidates, and is therefore called the *selector*. Let us illustrate this with the word pair (*spondylitis*, *?*) whose second term should be the French translation of *spondylitis*. The following analogical proportions are identified in (1): that written above, [*adenomalacia : adenitis = spondylomalacia : spondylitis*], [*arthropathy : arthritis = spondylopathy : spondylitis*], etc., where (*adenomalacia*, *adénomalacie*), (*adenitis*, *adénite*), (*spondylomalacia*, *spondylomalacie*), etc., are in our bilingual lexicon, but not (*spondylitis*, *?*). Analogical equations such as [*adénomalacie : adénite = spondylomalacie : ?*] are thereby formed and solved in (2), producing solutions among which (3) correctly selects *spondylite*.

2.3 Practical issues

Although simple in principle, analogical learning involves important practical issues. There are basically two problems that must be solved for the approach to work. The first one stems for the identification of the triplets in the input space that form with the unknown term an analogy (step 1). This is an operation a priori cubic in the number of input objects, about 10,000 here). Langlais and Patry [2007] describe an approach where the problem is turned into solving a quadratic number of analogical equations, which is still too time consuming in applications such as ours. To alleviate time issues, the authors propose to sample forms in the input space. In this work, we applied a technique described in [Langlais and Yvon, 2008] which allows to solve the problem in a time roughly linear in the input space size. Here again, the description of this technique is beyond the scope of this paper. Suffice it to say, that thanks to this technique, we can solve step 1 of analogical learning exactly, that is, we can identify *all* the analogies (involving the form to translate) present in the input space.

The second problem stems for the potentially high number of forms produced by the generator. These forms arise in part because the solver generates many solutions, as we already discussed. The fact that several input analogies, and in turn, several target equations are being considered while translating a single form exacerbates the problem. To our knowledge, there is no known satisfactory solutions to this issue yet. In this work, we simply keep the count with which a given solution is generated. The top-ranked solutions are those proposed by the analogical system. This simple selector has the advantage that it allows to investi-

gate how far down the list we must go to find the oracle solution (see next section).

3 Experimental protocol

3.1 Material

We ran our experiments with several goals in mind. First, we wanted to test how our approach is impacted by the size of the training material. Therefore, we collected two different sized corpora (MASSON and MESH) for the French-English language pair. Second, we wanted to check whether analogical learning is better suited for specific language pairs. We were also interested in observing whether it is more suited to translate into a morphologically rich language (such as Finnish) or the other way round. We therefore considered a bench of language-specific datasets (MESH). Last, we also compared the analogical method to a corpus-based alignment method, and compiled for this purpose the HEALTH dataset.

MASSON We used a list of French medical words and their English translations obtained from the Masson medical dictionary (`http://www.atmedica.com/`). The same initial list was used in the work of Claveau and Zweigenbaum [2005], but we did not keep words which were identical in French and English. We selected 13,392 word pairs with a normalized edit distance between 0.02 (differing by 1 character) and 0.67 (rather distant, such as (*toux, cough*). This list was randomly split into SEARCH (80%), DEV (10%) and TEST sets (10%, *i.e.*, 1,306 words) (DEV was not used in the part of the work presented here).

MESH The Medical Subject Headings (MESH) is the thesaurus used by the US National Library of Medicine to index the biomedical scientific literature in the MEDLINE database. Its preferred terms are called "Main Headings" (synonym terms are called "Entry Terms"). We collected pairs of source and target Main Headings (TTY[1] = 'MH') with the same MeSH identifiers (SDUI). [2] We considered five language pairs: two European close ones (English-French and English-Spanish), two distant ones (Finnish-English and Swedish-English) and one pair involving different scripts (Russian-English).[3] The resulting MESH datasets contain roughly half the pairs of terms we collected in MASSON. We randomly split each dataset in two parts: 90% into SEARCH and the remaining 10% into TEST.

HEALTH To investigate the performance of a corpus-based alignment method, we compiled two parallel corpora: one was obtained from the Health Canada English-French bilingual website (`http://www.hc-sc.gc.ca/`, 142,441 distinct target words), the other consists of 7,260 pairs of English and French abstracts of French journal articles published in about 350 French medical journals (about 3 million words in total, 107,441 distinct target words).

3.2 Evaluation

We computed the following measures to evaluate our analogical translation device:

Coverage is the proportion of input words for which the system can generate translations. If N_t words receive translations among N, coverage is defined as $\frac{N_t}{N}$.

Precision : among the N_t words for which the system proposes an answer, precision is the proportion of those for which a correct translation is output. The system proposes a ranked list of translations for each input word. Depending on the number of output translations k that one is willing to examine, a correct translation will be output for N_k input words. Precision at rank k is thus defined as $P_k = \frac{N_k}{N_t}$.

Recall is the proportion of the N input words for which a correct translation is output. Recall at rank k is defined as $R_k = \frac{N_k}{N}$.

We additionally compared, in ideal conditions, the recall of our (internal) method to that of an (external) word alignment method and to a non-generative internal method based on edit distance. *Word alignment* takes a parallel corpus of texts and their translations and aims to determine which pairs of (source, target) words are in a translation relation in the corpus. Ideally, if an input word is in the source corpus, its translation can be identified in the target corpus. Our ideal test therefore consists in checking whether a given input word occurs in the source part of our parallel corpora. *Edit distance* computes a distance between two words based on their common and distinct characters [Levenshtein, 1966]. Since in our setting, source and target words are often formally similar, given a list of potential target words, a candidate translation of an input word is the target word which is closest to it in terms of edit distance. An ideal situation for that method is one where all correct translations are included in the list of potential target words. We built such a list by adding the target part of our test MASSON set to the list of words in the English part of the HEALTH corpora (total of 229,695 unique words).

4 Results

The algorithm was applied to translate the terms of the TEST material, searching analogies (step 1) in the SEARCH set, solving the resulting analogical equations (step 2) then ranking solutions according to frequency (step 3).

Influence of the corpus size The contrast between the small (MESH) and large (MASSON) French-English datasets can be observed in Table 1. Out of the 1,306 terms of MASSON, 1,092 source words obtained translations, which yields a coverage of 83.6%. Precision ranges from $P_1 = 34.8\%$ to $P_{25} = 80.2\%$, while recall ranges from $R_1 = 29.1\%$ to $R_{25} = 67.9\%$. For the MESH test set, only 199 terms out of 509 ones received translations,

[1] In the UMLS Metathesaurus tables, the TTY field codes the type of the term. Its values depend on the source terminology.

[2] We did not collect pairs of entry terms because we do not know how to pair actual translations among the possibly numerous entry terms of a given main heading.

[3] Russian MeSH is normally written in Cyrillic, but some terms are simply English terms written in uppercase Latin script (e.g., *ACHROMOBACTER*). We removed those terms.

Table 1: Performance of the approach on MESH and MASSON (FR→EN).

	$nb.$	$Cov.$	P_1	R_1	P_{25}	R_{25}
MASSON	1306	83.6	34.8	29.1	80.2	67.9
MESH	509	39.1	46.2	18.1	61.3	24.0

yielding a coverage of 39.1%. The precision ranges from 46.2% to 61.3%, while recall ranges from 18.1% to 24%.

Clearly, the training size impacts the approach. Analogical learning can identify more input analogies in larger datasets, therefore proposing translations for more terms. This eventually comes at a price at rank 1: more noisy translations are produced for the largest dataset (see P_1). But allowing the system to propose more solutions clearly shows the advantage of searching through a larger datset.

Influence of the language pair We investigated on the MESH datasets the influence of the language pair and the translation direction. In total, we ran 10 translation sessions that are summarized in Table 2.

Table 2: Performance of analogical learning as a function of the translation direction (MESH).

	nb	$Cov.$	P_1	R_1	P_{25}	R_{25}
FI→EN	701	44.2	49.0	21.7	65.5	29.0
FR→EN	509	34.4	46.3	15.9	63.4	21.8
RU→EN	784	48.6	38.1	18.5	61.7	30.0
SP→EN	624	46.0	42.5	19.6	60.6	27.9
SW→EN	592	41.0	46.1	18.9	64.2	26.4
FI←EN	701	42.8	44.3	19.0	63.7	27.2
FR←EN	509	39.1	46.2	18.1	61.3	24.0
RU←EN	784	47.1	44.4	20.9	67.2	31.6
SP←EN	624	39.7	44.0	17.5	66.1	26.3
SW←EN	592	40.9	45.0	18.4	64.5	26.4

Overall, we observe that analogical learning offers comparable performances for all translation directions, although some fluctuations are observed. Somehow surprisingly, the largest coverage rates are observed when translating from and into Russian. This shows that analogical learning is not bounded to translate closely related languages only, not even is it designed to treat languages that share the same scripts. We also note that it is not affected by translating into a morphologically rich language, such as Finnish or Swedish.

Comparison to corpus-based alignment Among the 1,306 source words of MASSON, 262 can be found in the source part of the Health Canada corpus and 233 in the source part of the abstracts corpus (Table 3): in total 479 source words could ideally be translated by word alignment in these two corpora, *i.e.*, a recall of 36.7%, assuming the presence of suitable target words and a perfect aligner.

Comparison to edit-distance Comparing the source (test) words of the MESH dataset to the target words in the TEST material by edit distance (Table 4), we measured that

Table 3: Ideal recall by word alignment in two medical corpora.

	HEALTH	Abstracts	Total
Number of words	262	233	479
Recall (%)	20.1	17.8	36.7

93.7% of the correct translations were found in top position (MASSON). This shows that French and English terms are very close. When TEST set target words are merged into the list of target Health Canada words, this ideal recall decreased to 75.9%, and to 73.3% when adding the words of the abstracts corpus.

Table 4: Ideal recall by edit distance in TEST words and two medical corpora (MASSON). HealthC stands for the Health Canada corpus.

	TEST	+HealthC	+Abstracts
Corpus size	1,306	143,433	229,695
Found	1224	991	957
Recall (%)	93.7	75.9	73.3

5 Discussion

On MASSON (FR→EN), analogical learning could identify a correct translation for up to 67.9% of the source test words, with a corresponding precision of 80.2%. Given the simplicity of frequency ordering used in place of step 3 in the present experiments, we expect the system to perform better in terms of precision if a better strategy is devised. Ongoing work on using a classifier to select candidate solutions shows that we can boost the precision of candidates to 90% with little or no loss in recall.

Even if we used the same dataset (MASSON), a precise comparison with Claveau and Zweigenbaum [2005] is difficult however, since their TEST set, although taken from the same superset, was quite different from ours as it contained pairs of identical words. Their best attainable precision was 75% when test words were randomly selected as in the present work, but included 10–12% of identical words. They do not report the corresponding recall.

Examples of successful analogies on MASSON are shown in the first part of Table 5. Example 1 shows how a translation where a word ending is involved (*-ie* / *-ia*) leverages an example with a prefix switch (*exo-* ↦ *ecto-*), itself licensed by another word pair (*exosquelette* ↦ *ectosquelette*). Example 2 ilustrates how an (*-ic* ↦ *-oid*) change in English is generated for *dermic* by analogy to (*lupic* ↦ *lupoid*), thereby producing a translation for French *dermoïde*. Examples 3–4 and 5–6 show multiple paths to support the same candidate translations, with analogies based on different words and suffixes. Note that "suffixes" and "prefixes" as mentioned in this paragraph are only an a posteriori description of the results of the algorithm: no morphemic knowledge was given to the system.

Word alignment in our two parallel medical corpora could at best identify 36.7% of the source words. Indeed, additional or larger parallel corpora might be found, but it is

Table 5: Example analogies supporting correct translations

	source	triplets for analogical equations	target
1	exocardie FR→EN	<ectosquelette,ectocardie,exosquelette> <ectoskeleton,ectocardia,exoskeleton>	exocardia
2	dermoïde FR→EN	<lupique,lupoïde,dermique> <lupic,lupoid,dermic>	dermoid
3	immunisation FR→EN	<volatil,immun,volatilisation> <volatile,immune,volatilization>	immunization
4	immunisation FR→EN	<neutralité,immunité,neutralisation> <neutrality,immunity,neutralization>	immunization
5	périvésiculite FR→EN	<ombilical,périombilical,vésiculite> <umbilical,periumbilical,vesiculitis>	perivesiculitis
6	périvésiculite FR→EN	<odontogramme,vésiculogramme,périodontite> <odontogram,vesiculogram,periodontitis>	perivesiculitis
7	alpha-cyclodextrins EN→SW	<beta-endorphin,alpha-endorphin,beta-cyclodextrins> <betaendorfin,alfaendorfin,betacyklodextriner>	alfacyklodextriner
8	iodoproteins EN→FI	<phosphates,iodates,phosphoproteins> <fosfaatit,jodaatit,fosfoproteiinit>	jodoproteiinit
9	pneumopericardium EN→SP	< hydrothorax,hydropneumothorax,pericardium> < hidrotórax,hidroneumotórax,pericardio>	neumopericardio
10	polysaccharides EN→FR	<liposarcoma,sarcoma,lipopolysaccharides> <liposarcome,sarcomes,lipopolyoside>	polyoside
11	bronchoscopy EN→RU	<arthrography,arthroscopy,bronchography> <артрография,артроскопия,бронхография>	бронхоскопия
12	buxus EN→RU	<cistaceae,buxaceae,cistus> <ладанниковые,самшитовые,ладанник>	самшит

known that their size is not indefinitely extendable. This illustrates that methods which rely on spotting known words are limited by the available material where such words can be found. An interesting point however is that the intersection between the words that can be translated by word alignment and those that can be translated by analogy only counts 161 words: assuming each method performed optimally, their union could translate 1,205 words and obtain 92.3% recall. For instance, the following words were translated only by analogy: *ablépharie, abrachiocéphalie, acroesthésie, actinocongestine*, while the following were found only in the parallel corpus: *acuité, acyclique, acétylation, acétyltransférase*.

Compared to alignment, edit distance had an easier task since all target words are included in the search list. Had we not added the list of target words, edit distance would have had a much lower potential recall. A more realistic test would consist in using for a candidate list a large corpus such as the target part of our parallel corpora. This would then come close to the above word alignment experiment, where the test of identity with a source word would be replaced with one of edit distance with a target word—although a monolingual corpus would be sufficient, and could therefore be much larger.

On the MESH datasets, we observed that analogical learning is not impacted by the language pair being treated, nor by the translation direction. In particular, translating into a morphologically rich language does not seem to be a problem. This contrasts with statistical machine translation which is known to perform poorly when translating into a highly inflected language such as Finnish. Examples of successful analogies on MESH for various language pairs are shown in the bottom part of Table 5. Russian correspondences often happen to rely on transliterations (11), but not systematically (12, *buxus*/самшит). The latter shows that the method does not rely on character correspondences between translated word pairs, but only on the synchronous existence of analogical equations in both languages.

As it is often the case with corpus-based approaches, analogical learning is impacted by the size of the training material. Based on our contrastive experiment on MASSON and MESH, we observe this trend for analogical learning as well. More examples allow the approach to identify a larger set of input analogies, yielding in turn a better coverage.

The analogical method is the only one of those tested here which can generate translations for unseen words. The resolution of an analogical equation combines the known words in the equation to create a new, hypothetical word which solves it. Identifying and solving a large number of such analogical equations builds cumulative support for the most promising hypotheses.

A way to improve the analogical method would be to provide it with knowledge on morphemes or "subwords," as prepared in previous work [Namer and Baud, 2005; Deléger *et al.*, 2007]. This could be used to enforce morphemic boundaries when generating analogical equation solutions and therefore reduce the number of generated forms, or to perform a posteriori filtering of candidate translations in step 3.

Analogy-based word translations were used to help machine translation in [Langlais and Patry, 2007]. The top candidate translation was kept and directly fed to the MT

system. The authors report small but consistent improvements. Candidate translations can also be used to help specialized terminologists prepare term translations. In that setting, a terminologist would examine the top n translation candidates and select the most relevant translation (or produce another one) to include in the target terminology.

6 Conclusion

We introduced an analogy-based method to generate word translations and evaluated its potential on medical words. Its precision can be quite good once a stronger selection component is integrated in its last step. Its recall is lower, with an upper bound at 68% (MASSON) in the current experiments. We saw that it can be increased by a combination with complementary, existing methods based on attested words, such as word alignment in parallel corpora or edit distance with a large word list. It has the distinctive ability to generate translations for unseen words.

We tested our method on different language pairs involving morphologically rich languages (such as Finnish and Swedish) as well as languages with different scripts. We observed that the approach does not seem to be impacted by the language pair considered. Lepage [1998] gave evidence that the approach works as well for Asian languages. We therefore plan to test the present method on more distant language pairs including Japanese or Chinese.

Another perspective is to tackle the direct translation of multiword terms: our analogical solver can work directly on such terms without having to first segment them into words. This should be particularly interesting in the context of medical terminologies.

References

[Baud *et al.*, 1998] Robert H. Baud, Christian Lovis, Anne-Marie Rassinoux, Pierre-André Michel, and Jean-Raoul Scherrer. Automatic extraction of linguistic knowledge from an international classification. In Branko Cesnik, Charles Safran, and Patrice Degoulet, editors, *Proceedings of the 9 th World Congress on Medical Informatics*, pages 581–585, Seoul, 1998.

[Chiao and Zweigenbaum, 2002] Yun-Chuang Chiao and Pierre Zweigenbaum. Looking for French-English translations in comparable medical corpora. *Journal of the American Medical Informatics Association*, 8(suppl):150–154, 2002.

[Claveau and Zweigenbaum, 2005] Vincent Claveau and Pierre Zweigenbaum. Translating biomedical terms by inferring transducers. In Elpida Keravnou Silvia Miksch, Jim Hunter, editor, *Proceedings 10th Conference on Artificial Intelligence in Medicine Europe*, volume 3581, Berlin / Heidelberg, 2005. Springer.

[Deléger *et al.*, 2006] Louise Deléger, Magnus Merkel, and Pierre Zweigenbaum. Contribution to terminology internationalization by word alignment in parallel corpora. In *Proceedings AMIA Annual Fall Symposium 2006*, pages 185–189, Washington, DC, November 2006. AMIA.

[Deléger *et al.*, 2007] Louise Deléger, Fiammetta Namer, and Pierre Zweigenbaum. Defining medical words: Transposing morphosemantic analysis from French to English. In *Proc MEDINFO 2007*, volume 129 of *Studies in Health Technology and Informatics*, pages 152–156, Amsterdam, 2007. IOS Press.

[Hersh and Donohoe, 1998] William R. Hersh and L. C. Donohoe. SAPHIRE International: a tool for cross-language information retrieval. *Journal of the American Medical Informatics Association*, 5(suppl):673–677, 1998.

[Langlais and Patry, 2007] Philippe Langlais and Alexandre Patry. Translating unknown words by analogical learning. In *EMNLP-CoNNL*, pages 877–886, Prague, Czech Republic, 2007.

[Langlais and Yvon, 2008] Philippe Langlais and François Yvon. Scaling up analogical learning. In *22nd COLING*, Manchester, England, 2008.

[Lepage, 1998] Yves Lepage. Solving analogies on words: an algorithm. In *COLING-ACL*, pages 728–734, Montréal, Canada, 1998.

[Levenshtein, 1966] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, pages 707–710, 1966.

[Markó *et al.*, 2006] Kornél Markó, Robert Baud, Pierre Zweigenbaum, Lars Borin, Magnus Merkel, and Stefan Schulz. Towards a multilingual medical lexicon. In *Proceedings AMIA Annual Fall Symposium 2006*, pages 534–538, Washington, DC, November 2006. AMIA.

[McDonald, 1993] David D. McDonald. Internal and external evidence in the identification and semantic categorization of proper names. In Branimir Boguraev and James Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*, pages 61–76. MIT Press, Cambridge, MA, 1993.

[Namer and Baud, 2005] Fiammetta Namer and Robert Baud. Predicting lexical relations between biomedical terms: towards a multilingual morphosemantics-based system. In *Studies in Health Technology Information*, volume 116, pages 793–798. IOS Press, Amsterdam, 2005.

[Nelson *et al.*, 2004] Stuart J. Nelson, Michael Schopen, Allan G. Savage, Jacque-Lynne Schulman, and Natalie Arluk. The mesh translation maintenance system: Structure, interface design, and implementation. In Marius Fieschi, Enrico Coiera, and Yu-Chuan Jack Li, editors, *Proceedings 10 th World Congress on Medical Informatics*, volume 107 of *Studies in Health Technology and Informatics*, pages 67–69, Amsterdam, 2004. IOS Press.

[Nyström *et al.*, 2006] Mikael Nyström, Magnus Merkel, Lars Ahrenberg, Pierre Zweigenbaum, Hokan Petersson, and Hans Ahlfeldt. Creating a medical English-Swedish dictionary using interactive word alignment. *BMC Medical Informatics and Decision Making*, 6(35), October 2006.

[Stroppa and Yvon, 2005] Nicolas Stroppa and François Yvon. An analogical learner for morphological analysis. In *9th Conf. on Computational Natural Language Learning (CoNLL)*, pages 120–127, Ann Arbor, MI, 2005.