Analogical Translation of Medical Words in Different Languages

Philippe Langlais[1], François Yvon[2], and Pierre Zweigenbaum[2]

[1] Université de Montréal, Dept I.R.O., Québec, H3C 3J7 Canada
[2] CNRS, LIMSI, Orsay, F-91403 France
felipe@iro.umontreal.ca yvon@limsi.fr, pz@limsi.fr

Abstract. Term translation has become a recurring need in many domains. This creates an interest for robust methods which can translate 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. We evaluate our approach on medical terms. To study the robustness of the method, we evaluate it on a series of datasets taken from different language groups and using different scripts. We investigate to which extend the approach can cope directly with multiword terms, and study its dependency to the size of the training set.

1 Introduction

New words are coined all the time, especially in technical domains. Among others, medicine is well-known for its propension to create new words to describe new diseases (*cardiomyopathy perivesiculitis*), interventions (*cystectomy*), microorganisms (*autoantibodies*), substances (*thiogalactosides*), etc. Many of these are named with complex words, built from existing morphemes: neoclassical compounds are probably the most characteristic type, but other word formation devices are also productive, *e.g.*, words derived from person names (*Wolffian*). The previous sentences list examples in English, but the same observation applies to a number of other languages [1,2], seemingly with a great degree of parallelism. For instance, one finds in Swedish: *tiogalaktosider* (*thiogalactosides*); Finnish: *kystektomia* (*cholecystectomy*); Russian: *аутоантитела* (*autoantibodies*); French: *périvésiculite* (*perivesiculitis*).

The question we address in this paper is then: are medical word formation devices parallel enough in different languages for it to be possible to guess the translation of a new word? For instance, given knowledge of a number of medical words in a source language L_S and their translations in a target language L_T, can one *generate* the translation w_T in L_T of an unseen word w_S in L_S?

This problem has been addressed by Schulz and colleagues [3], who wrote rules to *generate* words in Spanish from Portuguese medical words. Claveau [4] went further by using machine learning techniques to learn transducers from examples to generate French words from English medical words (and the reverse).

These methods can be called "generative" as they build new target words from previously unseen source words. They can rely on human expertise [3] or on machine learning methods [4].

Quite different, non-generative methods can also be used to *identify* word translations in parallel corpora [5] if such corpora can be found which contain the desired source words (and their translations). These word-alignment methods take advantage of the prior existence of translations, but are intrinsically limited by the availability of parallel corpora. They can be called "identification-based" methods, as they must be provided with data which contain the solutions to the problem (the target translations). Comparable corpora can also be used when parallel corpora are scarce (*e.g.*, [6]), but they make the task of translation identification more difficult and error-prone.

Both kinds of methods can be helped if a morphological analyzer of the source and/or target languages is available [7,8]: in that case, complex source words can be decomposed and the generation or identification of target translations is reduced to that of correspondences between component morphemes. However, it requires a substantial human investment to obtain a precise morphological analysis of derived and compound words and to specify the mapping between component morphemes in source and target languages (even though it may be partially helped by machine learning methods, *e.g.*, [9]).

The present work explores the use of a different generative method: analogical learning [10,11,12]. As the above-mentioned methods of this type [4] 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 this kind of method performs on medical words. We evaluate it on a series of datasets and compare it to an identification-based, non-generative method based on edit-distance.

This paper is organized as follows. We first present the datasets used for testing the method. We introduce the principles of analogical learning on which our system relies. We describe a series of evaluations which test different features of the datasets. We discuss their respective results, which show that the method performs as well on the different language and script pairs, in different translation directions, on both uni- and multi-terms, but depends to some extent on the size of the training set.

2 Datasets

We ran our experiments with several goals in mind. First, we wanted to check whether analogical learning is better suited for specific language pairs. Second, we were interested in observing whether it is more suited to translate into a morphologic rich language (such as Finnish) or the other way round. Third, we wanted to appreciate whether analogical learning is equally efficient when translating multiterms (terms with several words) as when translating uniterms.

Last, we also wanted to gauge how important the quantity of training material is to the overall approach.

The UMLS Metathesaurus [13] is a large repository of medical terminologies, with over 1.5 million distinct concepts and over 5 million distinct terms in 17 languages (version 2008AA, March 2008).[3] A given concept may be labelled with terms from different languages. It is therefore a very interesting resource to extract bilingual medical lexicons. A difficulty however is that terms which label the same concept in different languages are not always linguistically translations of each other: they may correspond to different ways of referring to the same entity. For instance, UMLS concept C0027051 is labelled with *Myocardial infarction*, *Heart attack*, *Infarto miocardico*, *Infarto del Miocardio*, *Ataque al corazon*, among a total of 105 distinct strings; each term is tagged with its language, but there is no systematic tagging of which term is a translation of which other term. Therefore, we designed a series of filters to extract sets of bilingual term pairs from the UMLS Metathesaurus. Depending on the source terminologies, datasets of different sizes could be obtained.

Small size datasets: MeSH thesaurus. 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.[4] Its preferred terms are called "Main Headings" (synonym terms are called "Entry Terms"). We collected pairs of source and target Main Headings (TTY[5] = 'MH') with the same MeSH identifiers (SDUI). 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.

Russian MeSH is normally written in Cyrillic, but some terms are simply English terms written in uppercase Latin script (e.g., ACHROMOBACTER for English Achromobacter). We filtered out these terms (1,366), retaining only Cyrillic Russian MeSH terms (23,394).

Medium size datasets: MedDRA thesaurus. The Medical Drug Regulatory Activities thesaurus (MedDRA) is intended to describe adverse effects of drugs and other related terms. It contains different term types: high-level group terms (TTY = 'HG', 332 terms in English or Spanish), hierarchical terms (TTY = 'HT', 1682 terms) and lower-level terms (TTY = 'LT', 56580 terms). MedDRA also has preferred terms (TTY = 'PT', 17867 terms). We collected pairs of source and target terms of the same types (TTY = 'MH') with the same MedDRA identifiers (SDUI).

[3] The UMLS can be obtained at no cost from the National Library of Medicine at `http://www.nlm.nih.gov/research/umls/`.

[4] The MeSH thesaurus and its translations are included in the UMLS Metathesaurus. Independently from the UMLS, the MeSH can also be browsed online at `http://www.nlm.nih.gov/mesh/MBrowser.html`. The French-English bilingual version can be seen at `http://ist.inserm.fr/basismesh/mesh.html` or at `http://www.chu-rouen.fr/ssf/arborescences.html`.

[5] In the UMLS Metathesaurus tables, the TTY field codes the type of the term, with values depending on the source terminology.

Large size dataset: SNOMED CT nomenclature.

The Systematic Nomenclature of Medicine (SNOMED CT) has a large coverage of signs and symptoms, but also of anatomy, diseases and other medical concepts. SNOMED CT has full-form descriptor terms (TTY = 'FN', 311,313 terms in English / 310,311 in Spanish), preferred terms (TTY = 'PT', 311,313 / 310,311), synonymous terms (TTY = 'SY', 141,474 / 102,929). As in MeSH, we required that only preferred terms should appear in term pairs.

Data preparation. In each source, each word was lowercased, and pairs of identical words were discarded. Table 1 shows the number of terms for each source (column 2, *All terms*). We also prepared for each source its subset consisting of uniterms (terms composed of exactly one word, i.e., with no space) made only of alphabetic characters and possibly dashes, containing at least one lowercase character (column 3 of Table 1). It can be seen that MedDRA and SNOMED have a very small proportion of uniterms.

Table 1. Data sources: bilingual term lists. EN = English, FI = Finnish, FR = French, RU = Russian, SP = Spanish, SW = Swedish.

Dataset	*All terms*	*Uniterms*
mesh-SW-EN	19090	5928
mesh-FR-EN	19230	5091
mesh-SP-EN	21021	6240
mesh-FI-EN	21787	7013
mesh-RU-EN	23394	7842
meddra-SP-EN	67523	3598
snomedct-SP-EN	284255	10921

3 Analogical learning

An *analogical proportion* is a relation between four items $[x : y = z : t]$ where x is to y what z is to t in a sense to be specified (see Lepage [10] or Stroppa and Yvon [11] for more detail). Here, formal relations between strings of characters are considered, *e.g.*, [*aortotomy : aortitis = spondylotomy : spondylitis*]. *An analogical equation* is an analogical proportion where an item is unknown, *e.g.*, $[x : y = z :?]$. Stroppa and Yvon [11] propose a method to solve analogical equations, *i.e.*, to generate the missing fourth item. Complex objects may also be considered in an analogical proportion, *e.g.*, pairs of words of the form (*source, target*) where *target* is the translation of *source* (these are entries in an existing bilingual lexicon). Given such an object with a missing part (*e.g.*, missing *target*), *analogical inference* can predict it by solving analogical equations. It proceeds in three steps:

(i) collecting triplets of word pairs whose first elements define with *source* an analogy;

(ii) solving the analogical equations between the corresponding second elements;

(iii) selecting the best candidate among these solutions.

Let us illustrate this with the word pair (*spondylitis, ?*) where we want to find as second term the French translation of *spondylitis*. The following analogical proportions are identified in (i): 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 (ii), producing solutions among which *spondylite* (the correct translation). The same solution may be generated through multiple equations, therefore the frequency of each solution can be used to rank the solutions generated in (iii).

The main difficulties in this method stem from the very large number of analogical proportions that must be considered in (i) (it is cubic in the number of input objects), and have been addressed by sampling and by using suitable data structures.

4 Experiments

4.1 Experimental setup

For each experimental condition, we computed the following measures [14]:

Coverage : 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 $R_k = \frac{N_k}{N}$.

Edit-distance [15] computes a distance between two words based on their common and distinct characters. 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 this method is one where all correct translations are included in the list of potential target words. We built such a list by using the target part of each of our input bilingual lexicons, an extremely favorable situation for this method.

To study the applicability of the method to any medical term, not only those made of a single word (uniterms), we tested the methods both using the whole bilingual term lists and using their subsets consisting of only uniterms.

4.2 Results

The algorithm was applied to translate the different test sets, each consisting of a random 10% split of the prepared source bilingual term lists searching analogies (step *i*) in the SEARCH set, solving the resulting analogical equations (step *ii*) then ranking solutions according to frequency (step *iii*).

Analogy. Table 2 shows the coverage, precision and recall obtained on all types of terms from each language to English, then the same data for some of the reverse language pairs. P_1 and R_1 stand for precision and recall at rank 1, *i.e.*, when looking at the top candidate translation proposed by the algorithm. P_{25} and R_{25} refer to precision and recall at rank 25: this provides an idea of whether using a classifier to rerank candidate translations could find the correct translation among the top ones proposed by the present simple frequency ordering. Similar data is also displayed for uniterms only in Table 3.

Table 2. Generating translations through analogy for all types of terms. Coverage, precision and recall are shown as percentages. Correct is the percentage of terms that receive a reference translation by analogy. Because of the huge sizes of the full MedDRA and SNOMED terminologies, tests were only performed on a subset of the test material. However 90% of the whole terminologies were used to build analogies.

Dataset	Test	Coverage	Correct	P_1	R_1	P_{25}	R_{25}
All types of terms, Language X *to English.*							
mesh-FI-EN	2178	44.3	32.5	38.3	17.0	63.7	28.2
mesh-FR-EN	1923	38.2	29.5	45.5	17.4	69.3	26.5
mesh-RU-EN	2340	40.0	30.8	49.2	19.7	69.2	27.6
mesh-SP-EN	2102	43.3	35.1	50.5	21.9	73.1	31.6
mesh-SW-EN	1907	44.2	33.6	44.6	19.7	68.2	30.2
meddra-SP-EN	1589	73.4	62.9	19.0	13.9	53.4	39.2
snomedct-SP-EN	2000	60.1	49.0	35.0	21.1	62.9	37.8
All types of terms, English to Language X							
mesh-FI-EN	2178	46.5	31.7	34.0	15.8	54.7	25.4
mesh-FR-EN	1923	42.6	27.9	34.1	14.5	56.7	24.1
mesh-RU-EN	2340	46.7	33.4	36.8	17.2	60.7	28.3
mesh-SP-EN	2102	48.1	40.9	19.6	64.2	30.9	36.3
mesh-SW-EN	1909	43.8	31.9	38.0	16.7	64.2	28.1
meddra-SP-EN	1644	79.7	60.3	20.1	16.0	46.0	36.7
snomedct-SP-EN	1806	68.5	48.8	20.3	13.9	45.6	31.3

Edit-distance. Table 4 provides similar information collected with the edit-distance method. To complement the investigation of edit-distance, we observed

Table 3. Generating translations through analogy for uniterms.

Dataset	Test	Coverage	Correct	P_1	R_1	P_{25}	R_{25}
Uniterms, Language X *to English*							
mesh-FI-EN	701	44.2	30.4	49.0	21.7	65.5	29.0
mesh-FR-EN	509	34.4	23.0	46.3	15.9	63.4	21.8
mesh-RU-EN	784	48.6	32.1	38.1	18.5	61.7	30.0
mesh-SP-EN	624	46.0	29.8	42.5	19.6	60.6	27.9
mesh-SW-EN	592	41.0	29.1	46.1	18.9	64.2	26.4
meddra-SP-EN	361	50.8	41.7	48.5	24.6	77.3	39.3
snomedct-SP-EN	1094	57.8	34.6	34.6	20.0	54.8	31.7
Uniterms, Language English to X.							
mesh-FI-EN	701	42.8	29.7	44.3	19.0	63.7	27.2
mesh-FR-EN	509	39.1	25.1	46.2	18.1	61.3	24.0
mesh-RU-EN	784	47.1	33.0	44.4	20.9	67.2	31.6
mesh-SP-EN	624	39.7	28.0	44.0	17.5	66.1	26.3
mesh-SW-EN	592	40.9	28.4	45.0	18.4	64.5	26.4
meddra-SP-EN	359	56.3	45.7	33.2	18.7	63.4	35.7
snomed-SP-EN	1094	56.6	34.2	33.1	18.7	55.6	31.5

that uniterms and their translations in close languages (such as French and English or Spanish and English) are very similar (less than 3 edit-operations on average). Differences can be substantial for more distant language pairs (such as Finnish and Swedish into/from English). Of course, for languages that do not share the same alphabet, terms differ drastically, which plugs edit-distance based approaches. In some exceptional instances though, the correct match may happen to be found; for instance, the unique case in mesh-RU-EN uniterms where edit-distance provides the correct translation is *инь-янь* (*yin-yang*), where *инь-янь* is the only Russian term in the MeSH thesaurus made of two sequences of three letters separated by a hyphen.

Multi-terms and their translations are much less correlated in terms of edit-distance. We computed that an average of 8 to 12 edit-operations distinguish multi-terms from their translations in the different language pairs that share the same alphabet. The SNOMED and MedDRA tasks (all terms) involve a more important deviation of the source terms and their translations. Therefore, we can expect edit-distance variants to perform very badly on these tasks. Besides, as underlined earlier, the proportion of (easier-to-handle) uniterms in these two large terminologies is much lower.

5 Discussion

Examples of successful analogies are shown in Table 5. Example 1 (fr-en) shows how a translation where a word ending is involved (*-ie* / *-ia*) leverages an example with a prefix switch (*exo-* $\mapsto$ *ecto-*), itself licensed by another word

Table 4. Identifying translations through edit-distance for the Language X to English translation direction. As edit distance always proposes candidate translations, its coverage is always 100% and $P = R$, so we simplify the table accordingly and only show values for precision at ranks 1 and 25.

Test	P_1	P_{25}	P_1	P_{25}
		all terms		uniterms
mesh-SW-EN	33.8	37.8	70.0	74.8
mesh-FR-EN	71.8	77.1	84.6	89.6
mesh-SP-EN	81.5	89.1	85.8	89.7
mesh-FI-EN	33.6	38.0	71.2	76.8
mesh-RU-EN	1.0	1.1	0.1	0.8
snomedct-SP-EN	4.1	5.3	83.8	91.4
meddra-SP-EN	4.4	4.4	75.2	82.6

Table 5. Example analogies supporting correct translations.

	source/target	triplets for analogical equations
1	exocardie exocardia	<ectosquelette;ectocardie;exosquelette> <ectoskeleton;ectocardia;exoskeleton>
2	syöpägeenit oncogenes	<kasviproteiinit;syöpägeeniproteiinit;kasvit> <plant proteins;plants;oncogene proteins>
3	otsaontelo frontal sinus	<poskiontelotulehdus;poskiontelo;otsaontelotulehdus> <maxillary sinusitis;frontal sinusitis;maxillary sinus>
4	elintarviketeollisuus food industry	<lääkelainsäädäntö;elintarvikelainsäädäntö;lääketeollisuus> <legislation, drug;drug industry;legislation, food>
5	epha5-reseptori receptor, epha5	<akvaporiini 1;akvaporiini 5;epha1-reseptori> <aquaporin 1;receptor, epha1;aquaporin 5>
6	epha5-reseptori receptor, epha5	<alfa6beeta1-integriini;alfa5beeta1-integriini;epha6-reseptori> <integrin alpha6beta1;receptor, epha6;integrin alpha5beta1>

pair (*exosquelette* $\mapsto$ *ectosquelette*). This translation is indeed easy to find by edit-distance. The rest of those listed in Table 5 could not be found by edit-distance in our experiments. Example 2 (fi-en) pairs two formally unrelated words, *syöpägeenit* and *oncogenes*. Example 3 (fi-en) shows how an analogy on Finnish uniterms is parallelled by an analogy on English multiterms. In example 4 (fi-en), English terms involve commas and different word orders in analogy terms: *legislation, drug vs drug industry*. Example 5 (fi-en) has a hyphen and different word orders in Finnish and English. This example contains a digit, as does a rough 7% of the test terms in our dataset. Note that we do not treat digits in any specific way. Example 6 (fi-en) illustrates that different analogies can support the same translation.

Influence of parameters. The results do not evidence a strong influence of the language pair on analogical translation, whereas edit-distance is hindered by different scripts (Cyrillic) and (to a lesser extent) by more distant languages (Swedish, Finnish). This can be explained by several factors. A first factor is

linked with the analogy method, which does not rely on a comparison of the source and target terms. A second factor may come from the chosen domain, medicine, where a part of the vocabulary is built in a more or less systematic way. A third factor may come from the fact that most of the terms in our international terminologies are translations of an initial version, generally in English.

The translation direction has an impact on precision for some of the language pairs. For MeSH (all terms), precision is better when translating into English than the reverse. For MeSH uniterms, this is much less sensible. Globally though, analogy does not seem to be too much disturbed by translating into a rich morphological language.

Overall, analogical learning does equally well on uni- and multi-terms. This was expected since the method does not rely on the notion of word. For MeSH, between 23% (uniterms, French to English) to 40% (all terms, English to Spanish) of the test terms could be translated correctly by analogy. Many terms could not be translated because of a failure to identify analogies in the input space (step i).

For the MedDRA dataset, which is almost three times larger than the MeSH dataset, analogical learning could translate 63% of the Spanish to English *all terms* test set. Note however, that the precision is much smaller in that case. This is because many analogies are being identified during step (i), which in turn introduces many solutions. This clearly shows the need for a better filtering strategy (step iii) than the simple frequency-based ranking we considered in this study.

It is interesting to note, that for the SNOMED dataset, which is roughly four times larger than MedDRA, we witness a decrease of the number of correctly translated terms. If corpus size matters to a certain degree, what seems more important is the diversity of the phenomenon present in the search material.

Comparison with edit-distance. An interesting observation can be made when contrasting edit-distance and analogy variants. For uni-terms, edit-distance seems to be more appropriate when the languages share the same alphabet; it is the reverse for multi-terms. Translating multi-terms by analogy can lead to drastic improvements in precision and recall, as can be observed for the SNOMED and MEDDRA experiments, where edit-distance culminates at a recall of around 5% while analogy records a precision of 74% and a recall of 40% for the SNOMED dataset (rank = 25) and a precision of 55% and a recall of 31% for MedDRA (rank = 25). This clearly illustrates that analogy captures linguistic information that helps in translating multi-terms. The fact that it does not outperform edit-distance on single-terms (when using a single alphabet) is likely due to the nature of medical terms which share the same latin or greek roots, which facilitates the task of edit-distance-like approaches. Note however that edit-distance has access to the solution when translating, while analogy does not. Note also that for languages with different scripts (RU/EN), edit-distance simply fails to translate most of the terms. A transliteration step could alleviate this issue, but this would require specific resources for each language and script.

We investigated more closely whether the terms translated had a special configuration regarding edit-distance. We found out that the average edit-distance between terms and their reference translation is larger for the terms that we could not translate. The difference however, is not spectacular: in the order of one point for uniterms, and two points for multi-terms. This means that analogical learning is not especially biased toward translating "easy" terms.

Table 6. Average number of analogies found in the input space *nbi*, average number of target equations solved *nbe*, and average number of productive equations *nbp*, *i.e.*, equations with at least one solution. These figures are computed on the only words that received a translation by analogy for the X to English translation direction (similar figures are observed for the reverse direction).

	all terms			uniterms		
	nbi	*nbe*	*nbp*	*nbi*	*nbe*	*nbp*
mesh-FI-EN	55.5	28.3	25.4	7.8	6.3	5.2
mesh-FR-EN	63.2	26.2	23.7	6.4	5.8	4.9
mesh-RU-EN	43.4	28.6	25.4	30.3	8.1	6.8
mesh-RU-EN	37.5	29.9	26.3	30.3	8.1	6.8
mesh-SP-EN	30.2	27.4	25.3	15.8	6.7	5.5
mesh-SW-EN	60.3	18.8	16.5	17.8	7.5	5.9

Table 6 helps to appreciate the number of analogies identified in the input space, as well as the number of productive equations[6] formed in the output space. We call *productive* an equation which generates at least one solution. We observe that more analogies are identified while translating multi-terms. This might simply be due to the larger training datasets considered in this case. Another explanation could be that multi-terms exhibit strong construction patterns, as for instance in the case of *nervsystemets sjukdomar* in Swedish (*nervous system diseases*) that could be translated thanks to many analogies of the form:

$$[\textit{hypotalamustumörer}:\textit{nervsystemets tumörer} = \textit{hypotalamussjukdomar}:\textit{nervous system diseases}]$$
$$\Rightarrow [\textit{hypothalamic neoplasms}:\textit{hypothalamic diseases} = \textit{nervous system neoplasms}:?].$$

$$[\textit{ileumtumörer}:\textit{nervsystemets tumörer} = \textit{ileumsjukdomar}:\textit{nervsystemets sjukdomar}]$$
$$\Rightarrow [\textit{ileal neoplasms}:\textit{ileal diseases} = \textit{nervous system neoplasms}:?]$$

We also observe that most of the equations formed in the output space produce at least one solution, which indicates that the inductive bias of analogical learning (an input formal analogy corresponds to an output one) seems to be adequate.

[6] We only count the output equations that are being solved. In practice, many equations produced can be ruled out without solving them, thanks to properties on formal analogies.

Compared to analogy, 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 or word list such as can be found on the Web.

Synthesis and related work. A precise comparison with Claveau and Zweigenbaum [4] is difficult since their TEST set 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.

The analogical method 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. The frequency ordering used in this paper is a crude method for selecting the best translation; the use of a suitable classifier can boost selection (current experiments obtain a reduction of candidates by 90% with little or no loss in recall).

Another way to improve the analogical method would be to provide it knowledge on morphemes or "subwords," as prepared, *e.g.*, in [7]. 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 (iii).

6 Conclusion

We introduced an analogy-based method to generate word translations and tested it to evaluate its potential on medical words. Its precision can be quite good once a stronger selection component is integrated in its last step (current upper bound at 81%, MeSH, sp-en). Its recall is lower, with an upper bound at 55% (MedDRA, sp-en) in the current experiments. It can be increased by a combination with complementary, existing methods based on attested words, such as edit-distance with a large word list. It has the distinctive feature of being able to generate translations for unseen words.

We checked that the analogy method is robust on a series of language pairs, including distant languages (Finnish) and different scripts (Cyrillic). We also verified that it can tackle the direct translation of multiword terms without having to first segment them into words.

Acknowledgments

We thank the anonymous reviewers of this paper for their helpful comments. This work has been partially founded by the Natural Sciences and Engineering Research Council of Canada.

References

1. Iacobini, C.: Composizione con elementi neoclassici. In Grossmann, M., Rainer, F., eds.: La formazione delle parole in italiano. Niemeyer, Tübingen (2004) 69–95
2. Namer, F., Baud, R.: Predicting lexical relations between biomedical terms: towards a multilingual morphosemantics-based system. In: Stud Health Technol Inform. Volume 116. IOS Press, Amsterdam (2005) 793–798
3. Schulz, S., Markó, K., Sbrissia, E., Nohama, P., Hahn, U.: Cognate mapping - a heuristic strategy for the semi-supervised acquisition of a Spanish lexicon from a Portuguese seed lexicon. In: Proc 20th International Conference on Computational Linguistics (COLING 2004), Genève, Suisse (2004) 813–819
4. Claveau, V., Zweigenbaum, P.: Translating biomedical terms by inferring transducers. In Silvia Miksch, Jim Hunter, E.K., ed.: Proc 10th Conference on Artificial Intelligence in Medicine Europe. Volume 3581 of LNCS., Berlin / Heidelberg, Springer (2005)
5. Deléger, L., Merkel, M., Zweigenbaum, P.: Using word alignment to extend multilingual medical terminologies. In Zweigenbaum, P., Schulz, S., Ruch, P., eds.: Proc LREC Workshop Acquiring and representing multilingual, specialized lexicons: the case of biomedicine, Genoa, Italy, ELDA (2006) 9–14
6. Fung, P., Yee, L.Y.: An IR approach for translating new words from non-parallel, comparable texts. In: Proceedings of the 36^{th} ACL, Montréal (August 1998) 414–420
7. Hahn, U., Honeck, M., Piotrowski, M., Schulz, S.: Subword segmentation: Leveling out morphological variations for medical document retrieval. J Am Med Inform Assoc **8**(suppl) (2001) 229–233
8. Namer, F., Zweigenbaum, P.: Acquiring meaning for French medical terminology: contribution of morphosemantics. In Fieschi, M., Coiera, E., Li, Y.C.J., eds.: Proc 10^{th} World Congress on Medical Informatics. Volume 107 of Studies in Health Technology and Informatics., Amsterdam, IOS Press (2004) 535–539
9. Creutz, M., Lagus, K.: Morfessor in the morpho challenge. In: Proceedings of the PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes, Venice, Italy (2006)
10. Lepage, Y.: Solving analogies on words: an algorithm. In: COLING-ACL, Montréal, Canada (1998) 728–734
11. Stroppa, N., Yvon, F.: An analogical learner for morphological analysis. In: 9th Conf. on Computational Natural Language Learning (CoNLL), Ann Arbor, MI (2005) 120–127
12. Langlais, P., Patry, A.: Translating unknown words by analogical learning. In: Proc Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic 877–886
13. Lindberg, D.A.B., Humphreys, B.L., McCray, A.T.: The Unified Medical Language System. Methods Inf Med **32**(2) (1993) 81–91
14. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, New York (1999)
15. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady (1966) 707–710