Translating Biomedical Terms by Inferring Transducers

Vincent Claveau[1] and Pierre Zweigenbaum[2]

[1] OLST - University of Montreal, Montreal QC H3T 1N8, Canada,
Vincent.Claveau@umontreal.ca,
WWW home page: http://olst.ling.umontreal.ca/~vincent
[2] AP-HP, STIM/DSI; INSERM, U729; INALCO, TIM, Paris, France
pz@biomath.jussieu.fr,
WWW home page: http://www-new.biomath.jussieu.fr/~pz

Abstract. This paper presents a method to automatically translate a large class of terms in the biomedical domain from one language to another; it is evaluated on translations between French and English. It relies on a machine-learning technique that infers transducers from examples of bilingual word pairs; no additional resource or knowledge is needed. Then, these transducers, making the most of the high regularity of translation discovered in the examples, can be used to translate unseen French terms into English or vice versa. We report evaluations that show that this technique achieves high precision, reaching up to 85% of correct translations for both French to English and English to French tasks.

1 Introduction

In the biomedical domain, the international research framework and fast knowledge update make producing, managing and updating multilingual resources an important issue. Within this context, this paper presents and evaluates an original method to automatically translate a large class of biomedical simple terms (*i.e.*, composed of one word) from one language to another; it is tested on translations from French into English and English into French.

Our approach relies on two major hypotheses: (*i*) a large class of French and English terms are morphologically related; (*ii*) differences between French and English terms are regular enough to be automatically learned. These two hypotheses make the most of the fact that biomedical terms often share a common Greek or Latin basis, and that their morphological derivations are very regular (*e.g. ophtalmorragie/ophthalmorrhagia*, *leucorragie/leukorrhagia*...). Our technique relies on a supervised machine-learning algorithm, called OSTIA[3] (Oncina, 1991), that infers transducers (*cf.* next section) from examples of bilingual term pairs. Such transducers, when given a new term in English (respectively French), must propose the corresponding French (resp. English) term.

Only few researches aim at directly translating terms from one language to another (Schulz *et al.*, 2004). Nonetheless, closely related problems are often

[3] The authors wish to thank J. Oncina for giving them access to the code of OSTIA.

addressed in the domain of automatic corpus translation (cognate detection or statistical word alignement in bitexts (Véronis, 2000)). These approaches heavily rely on bitexts, which are not always available, and they look for an existing translation in a text (*i.e.* a relational problem) while we are willing to produce the translation of a term without other information (a generation problem).

In the next section, we present the machine-learning technique we use to infer transducers. In Section 3, we describes the methodology and the data used in our experiments. Section 4 details the results obtained for both French to English and English to French translation tasks. Last, some perspectives for this work are given in Section 5.

2 Inferring transducers

2.1 Sub-sequential transducers

The transducers inferred with OSTIA and used here to translate biomedical terms are an extension of the classical transducers called sub-sequential transducers. We give readers a basic background of these sub-sequential transducers; for formal definitions, please refer to (Oncina *et al.*, 1993).

Transducers are finite-state machines that can be seen as graphs in which an input symbol and an output string are associated to each edge and having one initial state I and one or more final states. An input sequence E is said to be *accepted* if there exists a path of edges from I to a final state such that the concatenation of the input symbols of these edges gives E. The *transduction*, or *translation* of an input sequence E corresponds to the concatenation of the output strings associated to the edges used to accept E. A sub-sequential transducer is a deterministic transducer (two edges with the same input symbol cannot emerge from the same state) in which all the states are final, and having an output string associated to each state. This string is produced when the input sequence to be accepted by the transducer ends on the state it is associated with.

Figure 1 presents a simple sub-sequential transducer with the usual notations of the automata. It represents the transduction function that translates ϵ (the empty word) into D, $a(bc)^n$ into $A(BC)^nE$ and $a(bc)^nb$ into $A(BC)^nBF$. A word like *abca* is not accepted and thus not translated by this transducer.

2.2 The OSTIA algorithm

OSTIA is the machine learning technique inferring sub-sequential transducers from examples of French/English pairs of biomedical terms. OSTIA is formally

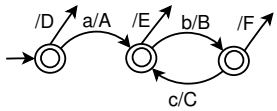

Fig. 1. A simple sub-sequential transducer

presented by J. Oncina (1991); here, we only give an outline of its principles. It is illustrated with an example: we want to learn the transducer presented above in Figure 1 from the training set T containing the 6 following input/output examples: $\{\epsilon/D, a/AE, ab/ABF, abc/ABCE, abcb/ABCBF, abcbc/ABCBCE\}$. OSTIA works in three steps (Oncina, 1998):

1. a prefix tree of every input sequence in T is built up. Empty output strings are associated to each internal state and edge of the tree and the complete output strings are associated at the leaves of the branch accepting the corresponding input sequence (*cf.* Figure 2).
2. every common prefix of the output sequences is moved up from the leaves towards the root of the tree (Figure 3).
3. last, starting from the root, every possible pair of states of the transducer is considered and is merged if the resulting transducer does not contradict the training data (Figures 4 and 5). When no further merge can be done, the algorithm ends.

3 Experiments

3.1 Learning data and evaluation set

The data we use to train and test this translation technique are taken from an on-line French medical dictionary (Dictionnaire Médical Masson) in which some of the entries contain English equivalent terms. We selected those entries that were simple terms both in French and English, avoiding proper nouns and acronyms. About 12,000 bilingual term pairs were collected this way.

In order to focus on morphologically related pairs, a formal similarity was computed for each pair through the string edit distance. Pairs were then ordered in a list according to their scores in descending order.

3.2 Methodology

It is important to provide OSTIA only with training pairs that are actually morphologically related (*i.e.* from the top part of the list presented above). In contrast, the data used to evaluate our technique can be taken from any part of this

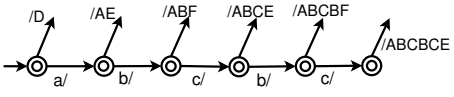

Fig. 2. Transducer at the end of step 1

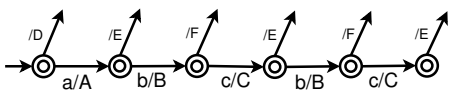

Fig. 3. Transducer at the end of step 2

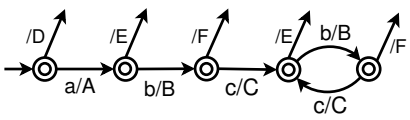

Fig. 4. Transducer after one merge

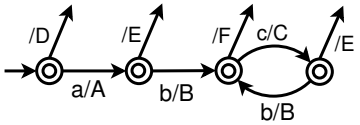

Fig. 5. Transducer after two merges

list, even though obviously no automatic system can provide good translations of terms at the end of the list. To take this point into account and provide a fair and complete evaluation, we run two experiments:

exp. 1. training pairs and testing pairs are taken from the first half of the list;

exp. 2. training pairs are taken from the first half of the list and testing pairs from the whole list.

For each experiment, we test our approach for the translation of terms from French to English and from English to French. The inference process is repeated 10 times: the initial set is divided in 10 folds and OSTIA uses 9 of these folds; the tenth fold is different each time. Thus, ten transducers are inferred; this allows us to average their results (see Section 4). Each test set comprises 2000 pairs (of course different from the training pairs).

Since ten transducers are inferred each time, it is possible to compare the translations they propose for each term in the testing set. The more frequently a translation is proposed, the more likely it is to be the correct one. Thus, we also implement a simple voting process: for each testing term, we keep the translation that was proposed most often by the ten transducers. In case of a tie between several translations, one of them is chosen at random.

4 Results

In order to evaluate the performances of the inferred transducers, the only measure we use is the precision of the produced translations. It is the rate of correctly translated terms from the source language into the target language. If a term is not accepted by the transducer, it is considered as incorrectly translated. This precision rate is computed for both experiments with respect to the number of training examples used by OSTIA. Figures 6 and 7 present the resulting graphs for experiment 1 and 2, for translation from English to French; similar graphs are obtained for French to English. We report the average precision of the 10 transducers inferred for each number of examples, as well as the precision obtained through the vote of the 10 transducers as described above. As a baseline, we compute the precision that would be obtained by a system systematically proposing the source term as its own translation.

The average precision is quite good and much better than the baselines: with 3,000 examples, about 64% of the terms are correctly translated in Experiment 1, and about 52% in Experiment 2. The precision obtained with the voting process is even better: in Experiment 1 it reaches about 85% of correct translations and 75% for Experiment 2. The good results for Experiment 2 are particularly interesting since it represents the performance of our technique on any term, even those with a non-morphologically related translation.

5 Perspectives

Many perspectives are foreseen for this work. From a technical point of view, future work is planned to improve the voting process between the inferred transducer. Indeed, the correct translation of a term is proposed by at least one of the

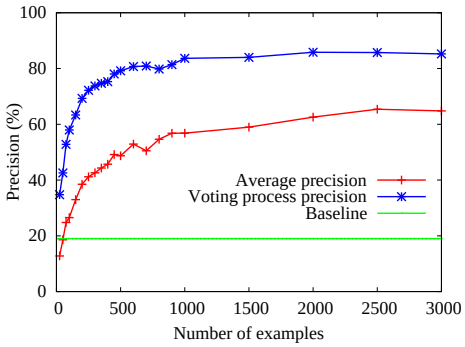

Fig. 6. Exp. 1 precision with respect to training set size En to Fr

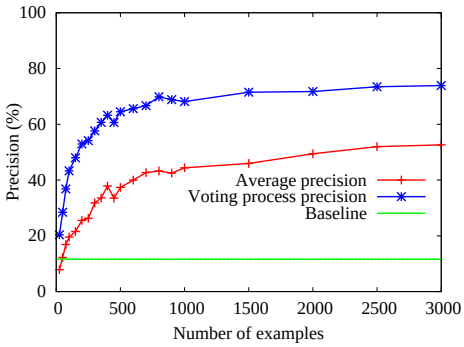

Fig. 7. Exp. 2 precision with respect to training set size En to Fr

10 inferred transducers 93% of the time for Exp. 1 and 85% for Exp. 2. Thus, these figures represent the upper limits an optimal voting method could reach. Secondly, considering the terms as sequences of morphs instead of sequences of letters (*e.g. broncho⊕pleuro⊕pneumo⊕nia*) could yield better results. Morphological analysis systems for French biomedical terms exist (Namer & Zweigenbaum, 2004) and could be used with OSTIA. Another possible extension concerns translation of complex terms (with several words). If we are able to translate word by word a complex term, it may be possible to produce a translation of the term as a whole, provided we are able to handle terminological variations (*virus de la variole/virus variolique*, *variola virus/variolic virus*) (Jacquemin, 2001).

References

Jacquemin C. *Spotting and Discovering Terms through NLP*. Cambridge: MIT Press (2001).

Namer F. & Zweigenbaum P. Acquiring meaning for French medical terminology: contribution of morpho-semantics. In *Proceedings of MEDINFO 2004*, San-Francisco, CA, USA (2004).

Oncina J. *Aprendizaje de lenguajes regulares y transducciones subsecuenciales*. PhD thesis, Universidad Politécnica de Valencia, Valence, Espagne (1991).

Oncina J. The data driven approach applied to the OSTIA algorithm. In *Proceedings of the Fourth International Colloquium on Grammatical Inference, ICGI'98*, p. 50–56, Ames, USA (1998).

Oncina J., García P. & Vidal E. Learning subsequential transducers for pattern regognition interpretation tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **15**(5), 448–458 (1993).

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 *Proceedings of the 20th International Conference on Computational Linguistics, COLING'04*, p. 813–819, Genève, Suisse (2004).

J. Véronis, Ed. *Parallel Text Processing*. Dordrecht: Kluwer Academic Publishers (2000).