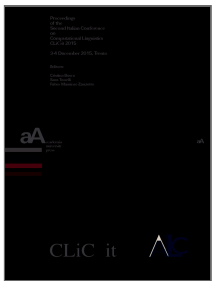

Cristina Bosco, Sara Tonelli and Fabio Massimo Zanzotto (dir.)

Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015
3-4 December 2015, Trento

Accademia University Press

New wine in old wineskins: a morphology-based approach to translate medical terminology

Raffaele Guarasci and Alessandro Maisto

DOI: 10.4000/books.aaccademia.1488
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2015
Published on OpenEdition Books: November 11, 2016
Series: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic EAN: 9788899200008

New wine in old wineskins: a morphology-based approach to translate medical terminology

Raffaele Guarasci, Alessandro Maisto
Department of Political, Social and Communication Sciences
University of Salerno
Via Giovanni Paolo II, 132,
84084 Fisciano (SA)
{rguarasci,amaisto}@unisa.it

Abstract

English. In this work we introduce the first steps toward the development of a machine translation system for medical terminology. We explore the possibility of basing a machine translation task in the medical domain on morphology. Starting from neoclassical formative elements, or *confixes*, we started building MedIta, a cross-language ontology of medical morphemes, aiming to offer a standardized medical consistent resource that includes distributional and semantic information of medical morphemes. Using this information, we have built an ontology-driven Italian-English machine translation prototype, based on a set of Finite State Transducers, and we have carried out an experiment on *Orphanet* medical corpus to evaluate the feasibility of this approach.

Italiano. *In questo lavoro si introduce lo sviluppo di un sistema per la traduzione automatica della terminologia medica. Si propone un approcio morfologico, che utilizza gli elementi formativi neoclassici, i* confissi*. Si introduce* MedIta*, un'ontologia multilingua di morfemi del dominio medico, che mira ad offrire una risorsa validata secondo gli standard medici e che contiene informazioni semantiche e statistiche. La fattibilit della risorsa viene valutata tramite un prototipo di sistema di traduzione italiano-inglese basato su Trasduttori a Stati Finiti.L'applicazione viene poi testata su un campione estratto dal corpus medico* Orphanet.

1 Introduction

Automating Machine Translation (MT) of a technical language is a challenging task that requires an in-depth analysis both from a linguistic point of view and as regards the implementation of a complex system. This becomes even more complex in medical language. Indeed the translation of medical terminology must always be validated by a domain expert following official classification standards. For this reason currently there are no translation support tools specifically created for the medical domain. In this work we propose an MT system based on a set of Finite State Transducers that uses cross-language morpheme information provided by a lexical resource. The underlying idea is that in a technical language a morpho-semantic approach (Dujols et al., 1991) may be more effective than a probabilistic one in term-by-term translation tasks. Even though our approach could seem a bit "old fashioned", we must consider that proper nature of medical language, fully based by morphemes derived from neoclassical formative elements (Thornton, 2005). Neoclassical formative elements are morphological elements that come into being from Latin and Greek words, they combine with each other following compositional morphology rules. Due to the heterogeneous nature of these elements, they have received different definitions, we prefer to use the term *confixes*, a morpheme with full semantic value, which has been predominantly used in the literature (Sgroi, 2003; D'Achille, 2003; De Mauro, 2003). In this work we focused only on word formation related to the medical domain.

2 Related Work

In the following section we briefly present the most relevant studies or applications regarding the use of a morpho-semantic approach, and studies that exploited morphological rules in machine

translation tasks. Morpho-semantic approaches have already been applied to the medical domain in many languages. Works that deserve to be mentioned are those by (Lovis et al., 1998) that identified the ICD[1] (International Classification Diseases) codes in diagnoses written in different languages; (Hahn et al., 2001) that segmented the subwords in order to recognise and extract medical documents; and (Grabar and Zweigenbaum, 2000) that used machine learning methods on the morphological data of the SNOWMED[2] nomenaclature (French, Russian, English). As regards morphological approaches in machine translation tasks, we mention a lexical morphology based Italian-French MT tool (Cartoni, 2009); MT models for morphologically rich languages, like Russian and Arabic (Toutanova et al., 2008; Minkov et al., 2007), a German-English biomedical terms MT tool (Daumke et al., 2006) and an approach based on finite state technologies (Amtrup, 2003). Furthermore we notice an unsupervised morph-tokens analysis applied to MT tasks (Virpioja et al., 2007) and an approach that applies morphological analysis to statistical MT systems (Lee, 2004).

3 Proposed approach

The proposed approach can be divided in two main phases:

- the creation of a lexical resource: an ontology of morphemes belonging to the medical domain to be used as a knowledge base. This ontology represents medical morphemes and provide both semantic and statistical (e.g. distributional profiles) information about them.
- the implementation of a MT prototype that exploits information provided by this lexical resource to perform an effective medical term translation.

Currently tested languages are Italian and English, but one of the advantages of the morpho-semantic method is that linguistic analyses designed for a language can often be transferred to other languages that share the common basis of neoclassical formative elements (Deléger et al., 2007).

[1] http://www.cdc.gov/nchs/icd/icd10cm.htm

[2] http://www.ihtsdo.org/snomed-ct

3.1 Medical morphemes ontology (MedITA)

Our starting point is an ontology of medical morphemes (prefixes, suffixes and confixes), that includes various kinds of information for each morpheme, like distributional profiles extracted from medical corpora, medical classifications and definitions. This resource is made possible by the formative elements underlying medical terms: morphemes may detect and describe the semantic relations existing between those words that share portions of meaning. Relying on words sharing morphemes endowed with a particular meaning (e.g. *-acusia*, hearing disorders) it is not difficult to find sets of near-synonyms (Namer, 2005). Moreover, we can infer the medical subdomain to which the synonym set belongs (e.g."*otolaryngology*") and we can differentiate any item of the set by exploiting the meaning of the other morphemes involved in the words.

- **synset**: *iper-acusia*, *ipo-acusia*, *presbi-acusia*, *dipl-acusia*;
- **subdomain**: *-acusia* "otolaryngology";
- **description**: *ipo-* "lack", *iper-* "excess", *presbi-* "old age", *diplo-* "double".

On the basis of the morphemes meaning, we can also infer relations between words that are not morphologically related, but which are composed of morphemes that share at least one semantic feature and/or the medical subdomain (see Table 1). This is made possible using formative elements, that do not represent mere terminations, but possess their own semantic self-sufficiency (Iacobini, 2004).

Related to	Morpheme	Subdomain
Tumors	*cancero-*, *carcino-*,	oncology
Stomach	*stomac-*, *gastro-*	gastroenterology
Skin fungus	*fung-*, *miceto-*, *mico-*	dermatology

Table 1: Morphemes that share semantic features

To start building the ontology we used a top-down approach: first of all we have divided the medical specialties into 22 categories (e.g. "internal medicine", "cardiology", "traumatology", etc...), with the support of a domain expert. The lexical resource used as source is the electronic version of the GRADIT[3] (De Mauro, 1999). Using

[3] Electronic version of *Grande Dizionario Italiano dell'Uso*

the GRADIT it has been possible to extract every kind of morpheme related to the medical domain and group them on the basis of their subdomains. Each morpheme has been compared with the morphemes included in the Open Dictionary of English[4]. The respective English translation has been manually added to each element. The resulting set of medical morphemes have been formalized into a resource that specifies their category:

- Confixes (*cfx*): neoclassical formative elements with a full semantic value (i.e. *pupillo-*, *mammo-*, *-cefalia*);
- Prefixes (*pfx*): morphemes in the first part of the word, able to connote it with a specific meaning (i.e. *-ipo*, *-iper*);
- Suffixes (*sfx*): morphemes in the final part of the word, able to connote it with a specific meaning (i.e. *-oma*, *-ite*);

Subsequently a set of semantic information has been added to every morpheme. These semantic labels provide descriptions about the meaning they confer to the words composed with them and information about morpheme classification. Such semantic information regards the three following aspects:

- Meaning: describes the specific meaning of the morpheme;
- Medical Class: gives information regarding the medical subdomain to which the morpheme belongs;
- Translation: presents the corresponding morpheme in the English language.

3.2 MT System

We built a Morphology-based Machine Translation prototype that works in two steps. The system is composed of a set of Finite State Automata to find approximate morpheme matching and a set of Finite-State Transducers[5] able to translate the Italian term into the English one. In the first step a partial matching to recognize Italian medical terms from text was performed, after that each recognized morpheme that composes the word was tagged with semantic information. To maximize the morphological recognition with minimum effort a set of patterns able to recognize different sequences of morphemes are identified (e.g. : *cfx-cfx*; *cfxs-sfx*; etc.) These patterns are derived from distributional profiles of morphemes: the most frequent compositions of morphemes extracted from a sample of 1000 words from ICD-10 for Italian and UMLS[6] (Unified Medical Language System) for English. A new category named *cfxs* is needed to reduce systematic kinds of errors in specific cases. *cfxs* identifies all the confixes that can appear before a suffix, with its correspondent English morpheme deprived of the final part, to avoid repetition in case of suffixation (i.e. *cystoitis*, *cfx-sfx*, is not valid, but *cystitis*, *cfxs-sfx* is valid). After that, the Transducer takes as input the morphemes and produces the corresponding translations. In the end, using the same morpheme sequences, it tags every Italian Medical Term with the respective English translation.

4 Experiment and Evaluation

To evaluate the approach described above and to assess its feasibility, we built a test dataset: a corpus of terms extracted from the Italian version of Orphanet[7], a resource that provides an inventory of more than 6000 rare diseases and a classification of diseases elaborated using existing published expert classifications. Orphanet has been chosen because the vast majority of rare diseases are composed of several morphemes (e.g. *hemimegalencephaly*, *acrocephalopolydactyly*). For each disease, Orphanet offers a brief summary with connections with other medical terminologies (MeSH[8], UMLS, MedDRA[9]) or standard classifications (ICD-10). In this early stage in order to test the performance of our morpho-semantic translator we evaluated the Precision score on a sample of 100 rare diseases extracted from Orphanet corpus. The "gold standard" taken into account is the translation provided from ICD-10. Our results were compared to those obtained using other MT systems widely used in recent years as a case study:

- **Google Translate**[10], the wildly popular MT service provided by Google. It uses a propri-

[4] https://www.learnthat.org/

[5] ASF and TSF are built using OpenFST Library (avaible at http://openfst.org/, in particular the python wrapper PyFst http://pyfst.github.io/

[6] http://www.nlm.nih.gov/research/umls/

[7] http://orpha.net/

[8] https://www.nlm.nih.gov/mesh/

[9] http://www.meddra.org/

[10] https://translate.google.com

etary statistical machine translation technology.

- **BabelNet**[11](Navigli and Ponzetto, 2010), a multilingual semantic network and ontology obtained as an integration of WordNet and Wikipedia.
- **HeTOP**[12](Grosjean et al., 2013), a controlled vocabulary that combines the best known biomedical terminology, vocabularies and classifications. It also integrates UMLS.

MT System	Precision
MedITA	91%
Google Translate	85%
BabelNet	73%
HeTOP	68%

Table 2: Precision comparison on Orphanet corpus

Although it must be considered that the system is based on an incomplete resource still in development and the test sample is quite small, this first analysis shows interesting results (see Table 2). In particular, a qualitative analysis of the results reveals some important aspects that deserve a deeper analysis. A brief summary and explanation of the most relevant aspects deriving from the Orphanet translation follows:

- On rare diseases the system has a precision higher than other systems, perhaps due to the intrinsic properties of the medical language, most evident in the case of rare diseases, as mentioned above. Notice that - in some cases - Google Translate and BabelNet provide a translation using a broader term (e.g. Google it: *"acromatopsia"* - en: *"colorblindess"*; it: *"iperargininemia"* - en: *"argininemia"*). Although in a broader context these translations could be considered as valid, in an extremely specific domain such as the medical one they are *de-facto* errors.
- In several cases the system proposes a translation that does not fit exactly with the standard: e.g. *polyendocrinopathia/polyendocrinopathy*. Many proposed translations can be considered *acceptable* because, although they are not yet formalized in the standard, they occur in other available resources, like technical papers, web pages, etc.
- The system never fails when other MT systems are wrong (see Table 3). This occurs with complex and extremely rare words; in these "extreme" cases we can argue that a morphological based translation could be better than a probabilistic one.

Another relevant aspect is that the system can work as spellchecker. This is a "side effect" of a morphological approach, despite that it may prove a useful function to improve precision, especially if it works on raw or uncontrolled data.

5 Conclusions

In this work we presented a morphology-based machine translation prototype specifically suited for medical terminology. The prototype uses ontologies of morphemes and Finite State Transducers. Even though the approach may seem a little out-of-date, the preliminary results showed that it can work as well as a probabilistic system in such a specific domain. It is worth mentioning that at this early stage we tested the prototype only on samples, since the evaluation is an extremely time-consuming task: every translated term must be manually compared with one or more medical standards. Medical standards are often not aligned, therefore an Orpha-number (disease id) does not necessarily match a disease listed in ICD-10. Moreover, these resources are not easily usable in an automated way, therefore the evaluation should entirely be done manually. Finally, even if at this preliminary stage there are many open issues, but the encouraging results suggest possible future developments: morpho-semantic approach, allows to easily extend the system to other languages; we can enrich the ontology to cover a bigger number of morphemes and we can take into account complex multiword expressions. A possible application of the system could be in the context of cross-border healthcare services in the European Union (Directive 2011/24/EU on patients' rights in cross-border healthcare)[13] and as a translation support tool for the international systems of coding diagnoses and disability (ICD and ICF[14]).

[11] http://babelnet.org/
[12] http://www.hetop.eu/hetop/
[13] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:088:0045:0065:EN:PDF
[14] http://www.who.int/classifications/icf/en/

Orphanet (it)	ICD-10 (en)	MedITA	Google	BabelNet	HeTOP
iperlisinemia	hyperlysinemia	✓	✗	✗	broader term
acrocefalopolisindattilia	acrocephalopolysyndactyly	✓	✓	✗	✗
polimicrogiria	Polymicrogyria	✓	✓	✓	✗
anisachiasi	anisakiasis	✓	✗	✗	✗
balantidiasi	balantidiasis	✓	✗	✗	✓
difillobotriasi	diphyllobothriasis	✓	✗	✗	✓
emimegalencefalia	hemimegalencephaly	✓	✗	✗	✓
poliembrioma	polyembryoma	✓	✗	✗	✗

Table 3: Translation comparison on rare diseases

References

Jan W Amtrup. 2003. Morphology in machine translation systems: Efficient integration of finite state transducers and feature structure descriptions. *Machine Translation*, 18(3):217–238.

Bruno Cartoni. 2009. Lexical morphology in machine translation: A feasibility study. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 130–138. Association for Computational Linguistics.

P. D'Achille. 2003. *L'italiano contemporaneo*. Il Mulino.

Philipp Daumke, Stefan Schulz, and Kornél Markó. 2006. Subword approach for acquiring and cross-linking multilingual specialized lexicons. *Programme Committee*, 1.

Tullio De Mauro. 1999. *Grande Dizionario Italiano dell'Uso*, volume 8. UTET.

Tullio De Mauro. 2003. *Nuove Parole Italiane dell'uso*, volume 7 of *GRADIT*. UTET.

Louise Deléger, Fiammetta Naner, Pierre Zweigenbaum, et al. 2007. Defining medical words: Transposing morphosemantic analysis from french to english. *Studies in Health Technology and Informatics*, pages 535–539.

Pierre Dujols, Pierre Aubas, Christian Baylon, and François Grémy. 1991. Morpho-semantic analysis and translation of medical compound terms. *Methods of Information in Medicine*, 30(1):30.

Natalia Grabar and Pierre Zweigenbaum. 2000. Automatic acquisition of domain-specific morphological resources from thesauri. In *Proceedings of RIAO*, pages 765–784. Citeseer.

Julien Grosjean, Tayeb Merabti, Lina F Soualmia, Catherine Letord, Jean Charlet, Peter N Robinson, Stéfan J Darmoni, et al. 2013. Integrating the human phenotype ontology into hetop terminology-ontology server. *Studies in health technology and informatics*, 192.

Udo Hahn, Martin Honeck, Michael Piotrowski, and Stefan Schulz. 2001. Subword segmentation–leveling out morphological variations for medical document retrieval. In *Proceedings of the AMIA Symposium*, page 229. American Medical Informatics Association.

Claudio Iacobini. 2004. Composizione con elementi neoclassici. In M. Grossmann & F. Rainer, editor, *La formazione delle parole in italiano*, pages 69–95. Niemeyer, T´ubingen.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 57–60. Association for Computational Linguistics.

Christian Lovis, Robert Baud, Anne-Marie Rassinoux, Pierre-André Michel, and Jean-Raoul Scherrer. 1998. Medical dictionaries for patient encoding systems: a methodology. *Artificial intelligence in medicine*, 14(1):201–214.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. 2007. Generating complex morphology for machine translation. In *ACL*, volume 7, pages 128–135.

Fiammetta Namer. 2005. Acquisizione automatica di semantica lessicale in francese: il sistema di trattamento computazionale della formazione delle parole dérif. In Anna Maria Thornton et Maria Grossmann, editor, *Atti del XXVII Congresso internazionale di studi Società di Linguistica Italiana: La Formazione delle parole*, pages 369–388.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

S. C. Sgroi. 2003. Per una ridenizione di "confisso": composti confissati, derivati confissati, parasintetici confissati vs etimi ibridi e incongrui. *Quaderni di semantica*, 24:81–153.

A. M. Thornton. 2005. *Morfologia*. Carocci.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying morphology generation models to machine translation. In *ACL*, pages 514–522.

Sami Virpioja, Jaakko J Väyrynen, Mathias Creutz, and Markus Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. *Machine Translation Summit XI*, 2007:491–498.