A Finite-State Approach to Translate SNOMED CT Terms into Basque Using Medical Prefixes and Suffixes

Olatz Perez-de-Viñaspre, Maite Oronoz, Manex Agirrezabal and Mikel Lersundi
IXA NLP Group
University of the Basque Country UPV/EHU
operezdevina001@ikasle.ehu.es

Abstract

This paper presents a system that generates Basque equivalents to terms that describe disorders in SNOMED CT. This task has been performed using Finite-State transducers and a medical prefixes and suffixes lexicon. This lexicon is composed of English-Basque translation pairs, and it is used both for the identification of the affixes of the English term and for the translation of them into Basque. The translated affixes are composed using morphotactic rules. We evaluated the system with a Gold Standard obtaining promising results (0.93 of precision). This system is part of a more general system which aim is the translation of SNOMED CT into Basque.

1 Introduction

SNOMED Clinical Terms (SNOMED CT) (College of American Pathologists, 1993) is considered the most comprehensive, multilingual clinical healthcare terminology in the world. It does not exist in Basque language, and we think that the semi-automatic translation of SNOMED CT terms into Basque will help to fill the gap of this type of medical terminology in our language. By its translation we have a double objective: i) to offer a medical lexicon in Basque to the bio-medical personnel to try to enforce its use in the bio-sanitary area, and ii) to access multilingual medical resources as the UMLS (***U****nified* ***M****edical* ***L****anguage* ***S****ystem*) (Bodenreider, 2004) in our language.

Basque is a minority language in its standardization process and persists between two powerful languages, Spanish and French. Although today Basque holds co-official language status in the Basque Autonomy Community, during centuries it was out of educational and sanitary systems, media, and industry.

We have defined a general algorithm (see section 2) based on Natural Language Processing (NLP) resources that tries to achieve the translation with an incremental approach. The first step of the algorithm is based on the mapping of some lexical resources and has been already developed. Considering the huge size of SNOMED CT (296,000 active concepts and around 1,000,000 descriptions in the English version dated 31-01-2012) the contribution of the specialized dictionaries has been limited. In the second step that is specified in this paper, we have used Finite State Machines (FSM) in the form of transducers to generate one-word-terms in Basque taking as a basis terms from the English release of SNOMED CT mentioned before. The generation is based on the translation by means of medical suffixes (i.e. *-dipsia*, *-megaly*) and prefixes (i.e. *episio-*, *aesthesi-*) and in their correct composition, considering morphotactic rules. (Lovis et al., 1995) stated that a big group of medical terms can be created by neologisms, that is, concatenations of existing morphosemantic units understood by anybody. This units usually have Greek and Latin origins and their meaning is known by the specialists. (Banay, 1948) specified that about three-fourths of the medical terminology is of Greek origin.

In this work we take advantage of these features to try to translate terms from the *Disorder* subhierarchy of SNOMED CT. This corresponds to one of the 19 top level hierarchies of SNOMED CT, to the one called *Clinical Finding/Disorder*. In our general approach, we prioritized the translation of the most populated hierarchies: *Clinical Finding/Disorder* (139,643 concepts), *Procedure* (75,078 concepts) and *Body Structure* (26,960 concepts). Using lexical resources, we obtained the equivalents in Basque of the 19.32 % of the disorders. In this work we will try to obtain the one-word-terms that are not found in dictionaries.

There are several general-purpose libraries for

Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing, pages 99–103,
St Andrews–Sctotland, July 15–17, 2013. ©2013 Association for Computational Linguistics

the creation of transducers as XFST (Karttunen et al., 1997), Nooj[1] or AT&T's FSM (Mohri et al., 2006). We have used Foma, a free software tool to specify finite-state automata and transducers (Hulden, 2009).

In the rest of the paper the translation algorithm is briefly described in section 2. The use of finite state machines in order to obtain Basque equivalents is explained in section 3. Finally, some conclusions and future work are listed in section 4.

2 Translation of SNOMED CT

The general algorithm (see figure 1) is language-independent. It could be used to translate any term if the linguistic resources for the input and output languages are available.

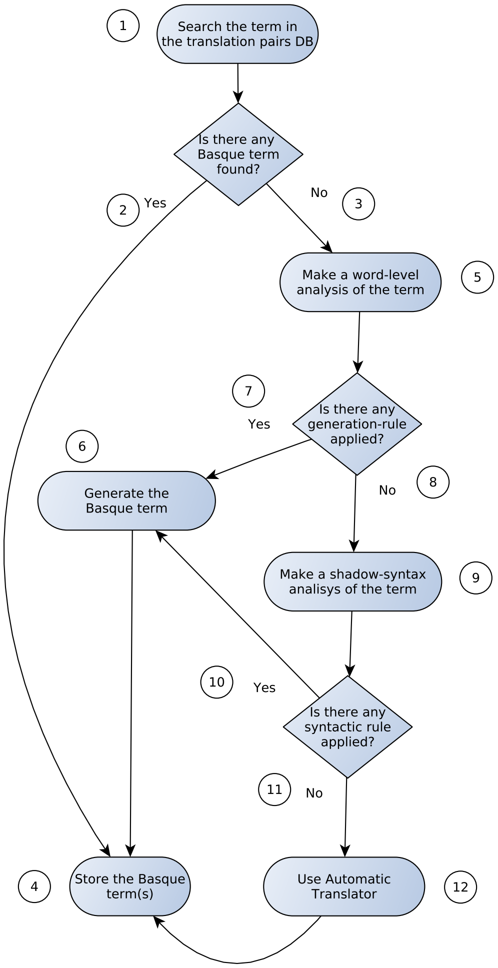
1
Search the term in the translation pairs DB
Is there any Basque term found?
2
Yes
No
3
Make a word-level analysis of the term
5
7
Yes
Is there any generation-rule applied?
6
Generate the Basque term
No
8
Make a shadow-syntax analisys of the term
9
10
Yes
Is there any syntactic rule applied?
11
No
4
Store the Basque term(s)
Use Automatic Translator
12

Figure 1: Schema of the Algorithm.

In the first step of the algorithm (see numbers 1-2-4 in Figure 1), some specialized dictionaries and the English, Spanish and Basque versions of the *International Statistical Classification of Diseases and Related Health in its 10th version* (ICD-10) are used. For example for the input term "abortus" all its Basque equivalents *"abortu"*, *"abortatze"* and *"hilaurtze"* are obtained.

The second phase of the algorithm is described in this paper in section 3. When a term is not found in the dictionaries (number 3 in Figure 1) generation-rules are used to create the translation.

In the case that an output is not obtained in the previous phases (number 8 in the algorithm), chunk-level generation rules are used. Our hypothesis is that some chunks of the term will be already translated. The application should generate the entire term using the translated components.

In the last step, we want to adapt a rule-based automatic translation system called *Matxin* (Mayor et al., 2011) to the medical domain.

We want to remark that all the processes finish in the 4th step. That is, we store the generated translations with the intention of using them to translate new terms.

3 Finite-State Models and Translation

This section exposes the system that obtains Basque equivalent terms from English one-word-terms based on FSMs.

3.1 Translation process

The generation of Basque equivalents is performed in two phases: the identification of the affixes first, and the translation and composition of the affixes secondly. All the linguistic information is stored in lexica and 31 rules are written for the process (1 for identification, 1 for translation and 28 for morphotactics).

Figure 2 shows the Finite State Transducer for the identification of the affixes. The lexica of the affixes is loaded (1-6) and then any prefix (the "*" symbol indicates 0 or more times) followed by one unique suffix is identified. The letter "o" may be also identified as it is used to join medical affixes. The "+" symbol is used for splitting the term.

```
1 read lexc prefixes.lex
2 define PREFALL
3 define PREF PREFALL.u ;
4 read lexc suffixes.lex
5 define SUFALL
6 define SUFF SUFALL.u ;
7 regex [[[PREF 0:%+] (o 0:%+)]* SUFF] ;
```

Figure 2: Rules for the affix identification.

The combination of the finite state transducers

[1] http://www.nooj4nlp.net/NooJManual.pdf

for the translation and for the composition using morphotactics is shown in Figure 3. First, the lexica for the translation task is loaded (1-4), then 28 rules for the morphotactics are defined (simplified in the rule numbered 5). The translation rule (shown in rule number 6) is composed of the word-start mark (the ˆ symbol), the prefix followed by the optional linking "o" letter zero or more times, and a single compulsory suffix; finally the transducer combines the translation and the morphotactic finite state transducers (7).

```
1 read lexc prefixes.lex
2 define TRANSPRE
3 read lexc suffixes.lex
4 define TRANSSUF
5 define MORPHO ...
6 define TRANS (%^ ) [[[TRANSPRE %+] (o:o %+)]*
TRANSSUF] ;
7 regex TRANS .o.  MORPH ;
```

Figure 3: Rules for the affix translation.

Figure 4 shows the whole process with an example. First, we identify the prefixes and suffixes of the English input term by means of the transducer that marks those affixes (schiz+encephal+y). Then, we obtain the corresponding Basque equivalent for each part and we form the term (eskiz+entzefal+ia).

Input term: schizencephaly
Identified affixes: schiz+encephal+y
Translated affixes: *eskiz+entzefal+ia*
Output. Basque term: *eskizentzefalia*

Figure 4: Basque term generation.

As we said before, in order to obtain a well formed Basque term, we apply different morphotactic rules. For example, in Basque, words starting with the "r" letter are not allowed, and an "e" is needed at the beginning. Figure 5 shows an example where the translated prefix *"radio"* needs of the mentioned rule, obtaining *"erradio"*.

Input term: radionecrosis
Identified affixes: radio+necr+osis
Translated affixes: *radio+nekr+osi*
Basque term: *erradionekrosi*

Figure 5: Morphotactic rule application.

3.2 Resources

In order to identify the English medical suffixes and prefixes we have joined two lists: the "Medical Prefixes, Suffixes, and Combining Forms" from Stedman's Medical Dictionary (Stedman's, 2005) and the "List of medical roots, suffixes and prefixes" from Wikipedia (Wikipedia, 2013). We obtained a list of 826 prefixes and 143 suffixes.

For the translation task, we have manually checked the Basque equivalents of the previously mentioned medical suffixes and prefixes list in specialized dictionaries such as *Zientzia eta Teknologiaren Hiztegi Entziklopedikoa* (Dictionary of Science and Technology) (Elhuyar, 2009), Euskalterm (UZEI, 2004) and *Erizaintzako Hiztegia* (Nursing Dictionary) (EHUko Euskara Zerbitzua and Donostiako Erizaintza Eskola, 2005).

By means of checking the behavior of the prefixes and suffixes in the English and Basque terms we have manually deduced the appropriate Basque equivalent. Table 1 shows an example of obtaining the equivalent of the "encephal" prefix, deducing that *"entzefal"* is the most appropriate equivalent.

English terms	**Basque terms**
echo**encephal**ogram	*eko**entzefal**ograma*
encephalitis	***entzefal**itis*
encephalomyelitis	***entzefal**omielitis*
leuko**encephal**itis	*leuko**entzefal**itis*
...	...

Table 1: The translation of the "encephal" prefix.

From all the prefixes and suffixes listed, we are able to deduce 812 prefixes and 139 suffixes for Basque. Those are currently being supervised by an expert to give them the highest confidence possible. This technique allows the inferring of new medical terms not appearing in dictionaries.

3.3 Results

We selected the one-word-terms of the *Disorder* sub-hierarchy of SNOMED CT. This sub-hierarchy with terms representing disorders or diseases is formed by 107,448 descriptions, being 3,979 one-word-terms. Even this last quantity is low considering the whole sub-hierarchy, we must take into account that the influence of those one-word-terms is very high, appearing around 79,000 times among all the descriptions.

The total one-word-term set has been split into two sets, one for defining and developing the system and another one for evaluating it. The evaluation set is composed of the 885 one-word-terms that have been previously translated in the first step of the algorithm (see section 2). That is we have the correct English-Basque pairs as Gold Standard. For the development set we have selected the remaining 3,094 one-word-terms.

As mentioned before, in this paper we show the results obtained from the translation of the medical prefixes and suffixes forming the terms. That is, we have only translated the terms that have been completely identified with the medical prefixes and suffixes. For example, terms with the suffix “thorax” have not been translated as it does not appear in the prefixes and suffixes list. That is, the “hydropneumothorax” term has not been translated even though the “hydro” and “pneumo” prefixes have been identified.

In Table 2 we show the quantities and percentages of the terms that have been completely identified in both sets. Our set of the one-word-terms has not been cleaned up to remove the words without any medical affix. Thus, the percentages from the table will never reach 100 per cent.

	Total	**Identified**	**Percent**
Development	3,094	834	26.96%
Evaluation	885	309	34.92%

Table 2: Quantities of completely identified terms.

From the 885 terms in the evaluation set, 728 terms contain at least one medical prefix or suffix, being 309 completely identified. The results obtained in this fist approach are shown in Table 3 by means of True Positives (TP), False Negatives (FN), False Positives (FP), Precision (Prec.), Recall (Rec.) and F-measure (F-M). A recall of 0.41 is obtained (287 correctly identified from 706 TP and FN) and a precision of 0.93 (287 out of 309). The recall will be increased in the future, including not completely identified terms in the system. Thus, we can conclude that the results obtained are very good concerning precision.

Total	**TP**	**FN**	**FP**	**Prec.**	**Rec.**	**F-M**
728	287	419	22	0.93	0.41	0.56

Table 3: Precision and recall of the evaluating set.

Moreover, the quality of the results obtained is also very good. We have been able to give correct equivalents to complex terms such as “hyperprolactinemia”, that has five medical prefixes and suffixes (“hyper+pro+lact+in+emia”).

We have also analyzed the incorrect results in order to be able to improve the system. For example, the prefix “myc” has been translated as “miz”, but we realized that whenever the prefix is followed by an “o”, it should be “mik” in order to generate a correct Basque term. Many of the mistakes are easily rectifiable for the final purpose of translating SNOMED CT.

4 Conclusions and future work

We implemented an application that generates Basque terms for diseases in English, by means of finite-state transducers. This application is one of the phases in the way to translate SNOMED CT into Basque. In order to translate the medical prefixes and suffixes, we have manually generated the translation pairs for 951 prefixes and suffixes, obtaining a very useful resource for Basque.

The FSTs exposed in this paper could be easily applicably to other languages whether an affix lexicon with its translation is defined and the morphotactic rules adapted to the target language.

As we have seen in section 3.3, most of the English terms have not been identified completely and that prevented the translation of them. To cope with this problem we have two developing paths: the deduction of new suffixes and prefixes from specialized dictionaries (Hulden et al., 2011); and the implementation of transliteration transformations to those parts (Alegria et al., 2006).

We have only applied the transducers to the *Disorder* sub-hierarchy, and we will have to check the results we can obtain applying it to the *Finding* sub-hierarchy and to the *Procedure* and *Body Structure* hierarchies. We found terms such as “electroencephalography” or “oligomenorrhea” in those hierarchies, formed with medical prefixes and suffixes identified for this task.

The promising results obtained will contribute to the translation of the whole SNOMED CT, but also to the normalization of Basque in the biosanitary domain, as new terms are generated.

References

I. Alegria, N. Ezeiza, and I. Fernandez. 2006. Named Entities Translation Based on Comparable Corpora. In *Multi-Word-Expressions in a Multilingual Context Workshop on EACL06*, pages 1–8.

G. Banay. 1948. An introduction to medical terminology, Greek and Latin derivations. *Bulletin of the Medical Library Association*, 36(1):1–27, Jan.

O. Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:267–270.

College of American Pathologists. 1993. The Systematized Nomenclature of Human and Veterinary Medicine: SNOMED International.

EHUko Euskara Zerbitzua and Donostiako Erizaintza Eskola. 2005. *Erizaintzako Hiztegia*. EHU. Argitalpen Zerbitzua.

Elhuyar. 2009. *Elhuyar Zientzia eta Teknologiaren Hiztegi Entziklopedikoa*. Elhuyar Edizioak & Euskal Herriko Unibertsitatea.

M. Hulden, I. Alegria, I. Etxeberria, and M. Maritxalar. 2011. Learning word-level dialectal variation as phonological replacement rules using a limited parallel corpus. In *EMLP 2011: Dialects2011*, pages 39–48.

M. Hulden. 2009. Foma: a Finite-State Compiler and Library. In *Proceedings of EACL 2009*, pages 29–32, Stroudsburg, PA, USA.

L. Karttunen, T. Gaál, and A. Kempe. 1997. *Xerox Finite State Tool*.

C. Lovis, Pa. Michel, R. Baud, and Jr. Scherrer. 1995. Word Segmentation Processing: A Way To Exponentially Extend Medical Dictionaries. *MEDINFO*, 8:28–32.

A. Mayor, I. Alegria, A. Díaz de Ilarraza, G. Labaka, M. Lersundi, and K. Sarasola. 2011. Matxin, an Open-source Rule-based Machine Translation System for Basque. *Machine Translation*, 25:53–82.

M. Mohri, F. Pereira, M. Riley, and C. Allauzen. 2006. AT & T FSM Library Finite-State Machine Library. Technical report, AT&T Labs-Research, NJ, USA.

Stedman's, 2005. *Stedman's Medical Dictionary*, chapter Medical Prefixes, Suffixes, and Combining Forms. Lippincott Williams & Wilkins, twenty-eighth edition edition.

UZEI. 2004. Euskalterm Terminologia Banku Publikoa. `http://www.euskadi.net/euskalterm`.

Wikipedia. 2013. List of medical roots, suffixes and prefixes – Wikipedia, The Free Encyclopedia. `http://en.wikipedia.org/w/index.php?title=List_of_medical_roots,_suffixes_and_prefixeso`.