An affix-based method for automatic term recognition from a medical corpus of Spanish

Antonio Moreno-Sandoval  
Autonomous University of Madrid  
antonio.msandoval@uam.es

Leonardo Campillos-Llanos  
Autonomous University of Madrid  
leonardo.campillos@uam.es

Alicia González-Martínez  
Autonomous University of Madrid  
a.gonzalez@uam.es

José M. Guirao-Miras  
University of Granada  
jmguirao@ugr.es

1 Introduction

Automatic Term Recognition (ATR) aims to develop software to identify a list of candidate words in a text that are likely to be technical terms. After this process, specialists have to assess the results in order to validate the final list of terms.

Current literature on ATR describes basically three approaches: statistical, linguistic and hybrid. First, statistical techniques rely on measuring how distinctive is a word or lemma in a specialized context when compared with a general corpus. This approach is well represented by the log-likelihood statistic (Dunning 1993) included in Wordsmith Tools (Scott 2008), or the logDice metric used in The Sketch Engine (Kilgarriff et al. 2004).1

Secondly, linguistic approaches focus on using language resources such as dictionaries, lexicons, and ontologies. For further references of each approach, we refer to Ananiadou and Nenadic (2006).

This paper presents a hybrid approach with two stages. Firstly, automatic methods are applied to construct the list of candidates. Secondly, a list of affixes is used to select proper medical terms, in order to reduce the human assessment.

We conducted an experiment on comparing three different automatic methods to obtain lists of candidates. For that goal, we evaluated the accuracy of the medical terms selected by each approach.

2 Overall description of the method

Our ATR method consists of two phases (see Figure 1):  
1. Extracting lists of candidate terms.  
2. Matching candidate terms against the affix list.

3 The Spanish MultiMedica corpus

The MultiMedica corpus has been compiled for the homonymous project2 by the Computational Linguistics Laboratory at the Autonomous University of Madrid (LLI-UAM). This is a comparable corpus of Spanish, Arabic and Japanese texts about health topics. One of the foreseen applications is an ATR tool aimed at translators and terminologists in the health domain. The Spanish corpus consists of three resources:

- **Harrison:**3 it includes professional and scientific texts written by medical doctors, and gathers over 3800 documents
- **OCU-Salud:**4 it is a collection of journalistic texts written by medical doctors, but edited and reviewed by journalists.
- **Website Tu otro medico:**5 it assembles encyclopedic articles written by professional doctors for non-specialists.

The corpus was filtered by hand in order to avoid information redundancy. In total, it covers 4200 texts and over 4 million words, and reflects a balanced combination for most medical specialties.

4 Extraction of the three lists

To obtain the lists of candidate terms, we tested three approaches: a tagger-based technique (by

---

1 www.sketchengine.co.uk/ [24/05/2013]

2 http://labda.info.uc3m.es/multimedica/

3 http://www.harrisonmedicina.com/

4 http://www.oci.org/oci-salud/

5 http://www.tuotromedico.com/
using GRAMPAL tagger, Moreno and Guirao (2006), a corpus-based technique (with a frequency list of word forms in the CREA Corpus of Contemporary Spanish) 1, and a statistical technique (the Log-likelihood statistic, henceforth LLH):

- The “Unknown-by-GRAMPAL” list contains the words that were not recognized by the GRAMPAL lexicon, which contains over 50000 lemmas. This lists includes 22413 tokens.
- The “Not-in-CREA” list is formed by all the words in MultiMedica corpus that are not found in the CREA corpus. This list has 23239 tokens.
- The LLH list was extracted by comparison against the CREA. Only the words ranked over 10 in the LLH statistic are considered. This list assembles 8667 word forms.

Almost 50% of the items in the CREA list are noisy words: 350000 out of over 700000 tokens are non-Spanish words (mostly foreign words, but also proper names and misspellings). On the other hand, GRAMPAL can analyze over 500000 correct word forms.

5 The affix list

Linguistic and rule-based methods have already been applied to ATR tasks in the immunological (Ananiadou 1994) and the pharmacological domain (Segura et al. 2008). In our experiment, we used a list of 2168 affixes (considering spelling variations: e.g. aden-, adeno-). The list includes these data:

- Greek and Latin affixes from the health domain (e.g. cardio-, -itis); this list has been enriched with very frequent roots (e.g. pancrea-) taken from studies on medical terminology (López Piñero and Terrada 2005; Jiménez Arias 2012; Sánchez González 2012). We did not include very general affixes that are not always related to the medical domain (e.g. pre-).
- Stems for the recognition of pharmacological substances (e.g. -cavir), which were compiled from lists approved by the American Medical Association (AMA) for the nomenclature of clinical compounds; 2 and stems proposed by the World Health Organization (WHOa, WHOb 2011). We collected also affixes that refer to biochemical entities (e.g. -sterol). 3 Since most English affixes have a univocal correspondence with the Spanish terms, few needed an adaptation (especially, those ending with vowel; e.g. -ine > -ina, as in creatine > creatina).

6 The matching procedure

Prior to the matching procedure, we pre-processed the affix list to generate possible variants of each affix. This stage consisted of three tasks:

- In cases where the affix has an orthographic (or acute) accent, a variant without it is used. The affix may bear accent or not depending on the place of the stress within the word (e.g. in the term próstata the prefix is próst-, whereas in prostatico the prefix is prosta-).
- When the affix has an epenthetic vowel, two versions of it are used (e.g. from (ej)scoli- we have escoli- and scoli-).
- It is important to note that the candidates of the list are non-lemmatized forms. Thus, all inflected variants are generated for suffixes ending in -o (e.g. from suffix -génico we generate -génicos, -génica, and -génicas).

Subsequently, the affix list is classified in a sublist of prefixes and a sublist of suffixes. Both lists are sorted by length, so the larger affixes are placed at the beginning of the list.

Secondly, we carried out the matching procedure itself. We took each candidate from the three lists and compared it with each affix from the lists. If a substring from the beginning matches one prefix, we stopped the search and marked the prefix. We took the remainder of the candidate string and searched more prefixes until there are no matches. Then, we proceeded in the same manner with the suffixes. In the end, we had a list of prefixes and a list of suffixes for each candidate.

Finally, candidate terms with no affixes were classified as rejected, and items with at least one affix were classified as accepted. From this, we created a list of rejected and a list of accepted candidate terms for the three methods (Table 1).

Results in our corpus show that the list obtained by means of the GRAMPAL tagger has the highest recall, followed by the list obtained by comparing the MultiMedica corpus wordlist with the CREA.

---

1 The CREA Corpus contains over 150 million tokens and over 700000 different word forms. Frequency lists are available at: http://corpus.rae.es/frecuencias.html
3 We looked up the list of affixes collected by Michael Quinion (2008): http://www'affixes.org [accessed: 02/01/2012]
7 Evaluation

The lists of accepted and rejected terms by each method were manually evaluated in order to confirm the results. We accepted a candidate term if that item was registered in prestigious research articles and medical books or in an authorized medical dictionary (e.g. Dorland 2005; or Diccionario de términos médicos, Real Academia Nacional de Medicina 2011). Units of measure (e.g. kg) or foreign words were not accepted as terms, excepting those borrowings that health professionals tend to use (e.g. heparin was rejected, since it is adapted to heparina in Spanish; yet stent or bypass were accepted as terms). After having looked up the candidates in each list, we prepared contingency tables for each method (Tables 2-4).

<table>
<thead>
<tr>
<th></th>
<th>Accepted</th>
<th>Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unknown for GRAMPAL</strong></td>
<td>14551</td>
<td>7862</td>
</tr>
<tr>
<td><strong>Not in CREA</strong></td>
<td>12307</td>
<td>10932</td>
</tr>
<tr>
<td><strong>LLH</strong></td>
<td>3832</td>
<td>4835</td>
</tr>
</tbody>
</table>

Table 1. Accepted and rejected candidate terms.

<table>
<thead>
<tr>
<th></th>
<th>Accepted</th>
<th>Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Not in GRAMPAL</strong></td>
<td>88.56%</td>
<td>48.13%</td>
</tr>
<tr>
<td><strong>False</strong></td>
<td>11.44%</td>
<td>51.87%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Accepted</th>
<th>Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Not in CREA</strong></td>
<td>83.12%</td>
<td>34.80%</td>
</tr>
<tr>
<td><strong>False</strong></td>
<td>16.88%</td>
<td>65.20%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Accepted</th>
<th>Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Log-likelihood</strong></td>
<td>91.99%</td>
<td>69.74%</td>
</tr>
<tr>
<td><strong>False</strong></td>
<td>8.01%</td>
<td>30.26%</td>
</tr>
</tbody>
</table>

Tables 2-4. Contingency tables.

In the final step of the experiment, we collected a list of all candidate terms by gathering the results from the true accepted and the false rejected terms for each list, and deleting any repeated item. The following table shows the estimation of precision, recall, and F values for each method.

<table>
<thead>
<tr>
<th></th>
<th>GRAMPAL</th>
<th>CREA</th>
<th>LLH</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Precision</strong></td>
<td>88.56%</td>
<td>83.12%</td>
<td>91.99%</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>51.57%</td>
<td>41.44%</td>
<td>14.28%</td>
</tr>
<tr>
<td><strong>F measure</strong></td>
<td>65.18%</td>
<td>55.31%</td>
<td>24.72%</td>
</tr>
</tbody>
</table>

Table 5.

8 Discussion

In our experiment, the LLH statistical method stands out slightly above in terms of precision (91.99%), but not in terms of recall, due to the high rate of rejected items that are true terms (69.74%). The highest precision rate in this list may be explained by several reasons. First, the “Not-in-CREA” list is rather noisy for it includes many not recognized verb inflected forms. Secondly, the CREA corpus contains texts from the scientific and medical domain. When comparing the list of words from the MultiMedica corpus against the CREA wordlist, many terms included in the scientific texts from the CREA are not proposed as candidate terms – although they are medical terms. Similarly, the list obtained with GRAMPAL does not contain several medical terms which are included in the lexicon of the tagger. For example, GRAMPAL recognizes vacuna (‘vaccine’) as a valid word. Thus, it is not proposed as a candidate term, despite being a medical term.

9 Conclusions and future work

We have performed and evaluated three methods for ATR in the biomedical domain: a tagger-based, a corpus-based, and a statistical approach. The experiment in our corpus has shown that the LLH method achieved a high precision in retrieving candidate terms, at the expense of a low recall. For Spanish, which is an inflecting (or fusional) language, the combination of a tagger-based method with a comprehensive list of affixes provided better results.

Acknowledgments

This research has been supported by the Spanish Government (under the grant TIN2010-20644-C03-03) and by the Madrid Regional Government (under the grant MA2VICMR).

References


Jiménez Arias, M., E. (2012) “Afijos grecolatinos y de otra procedencia en términos médicos”. MEDISAN,
16(6): 1005-1021 [Accessed: 30/12/2012].


Real Academia Nacional de Medicina (2011) Diccionario de términos médicos. Madrid: Médica Panamericana


Longitudinal development of L2 English grammatical morphemes: A clustering approach

Akira Murakami
University of Cambridge
a.murakami39@gmail.com

1 Aim of the study

Based on a longitudinal learner corpus, this study aims at disclosing the developmental patterns of grammatical morphemes in English as the second language (L2). More specifically, the following research questions are addressed:

1. Does the longitudinal transition of accuracy within individual learners show systematic patterns such as linear increase or U-shaped development, or does the accuracy randomly fluctuate?

2. Are the patterns different depending on morphemes, learners’ native language (L1), and their proficiency?

2 Corpus

The EF-Cambridge Open English Learner Database (EFCamDat) is a learner corpus containing learners’ essays written in English as a second language (L2). A typical English course at Englishtown has 16 Lessons with eight Units each. At the end of each Unit is a free composition in which learners are asked to write on a certain topic.

Learners in Englishtown receive feedback from native-speaker “teachers” on each essay. The feedback includes identification and correction of grammatical morphemes, among other things. The present study views the feedback as error annotation and utilizes it in the calculation of accuracy scores explained later.

The EFCamDat includes for each essay such metadata as the learner’s country of residence, the title of the essay, the date and time of submission, and the Lesson and the Unit number at which the essay was written.

3 Target morpheme, L1 groups, and proficiency levels

The study targeted the following six grammatical morphemes; articles, past tense –ed, plural –s, possessive ‘s, progressive –ing, and third person –s.

Ten L1 groups were targeted: L1 Portuguese,