An Automatic Extractor for Biomedical Terms in Spanish

Leonardo Campillos-Llanos*, Antonio Moreno-Sandoval*, José Mª. Guirao-Miras†
*Computational Linguistics Laboratory - Universidad Autónoma de Madrid
†ETS Ingeniería Informática y de Telecomunicación - Universidad de Granada
{leonardo.campillos, antonio.msandoval}@uam.es
jmguirao@ugr.es

Introduction

- Terms are linguistic realizations of concepts in a specific domain [1, 2].
- Automatic Term Recognition (ATR) aims at identifying candidate words in a text.
- ATR involves indentifying two features of a term: its *termhood* and its *unithood* [2].
- Further difficulties arise when systems deal with variations [3] and homonymous words.
- We present a system that uses lexically-based, tagger-based, and ruled-based methods.

System architecture

- The system consists of four steps (Fig. 1), each selecting different types of candidates:
  - □ **High reliability**: we use a gold standard list of terms curated from dictionaries [10-13].
  - □ **Medium reliability**:
    - ‣ **Single-word terms**: there are two types:
      - Items registered in a silver standard list.
      - Words that were not in the silver standard are proposed as candidate terms if:
        1. a Part-of-Speech tagger for Spanish (GRAMPAL, [14]) does not recognize them; and 2. a list of biomedical roots, stems and affixes matches any unrecognized word.
    - ‣ **Multi-word terms**: we use rules of multi-word term formation and phrase patterns.

The PoS tagger and the list of stems, roots and affixes

- GRAMPAL [14] contains more than 50,000 lemmas and generates over 500,000 words.
- The list of Biomedical stems, roots and affixes is made up of:
  - □ **Graeco-Latin affixes** (e.g. *cardio*-) and **roots** (e.g. *pancrea*-) gathered from studies on medical terminology [22-24].
  - □ **Stems/affixes** for the recognition of **pharmacological** and **biological substances** (e.g. *–cavir*). They were compiled from lists approved by medical institutions [25-28].
- We excluded general affixes: e.g. *pre*-

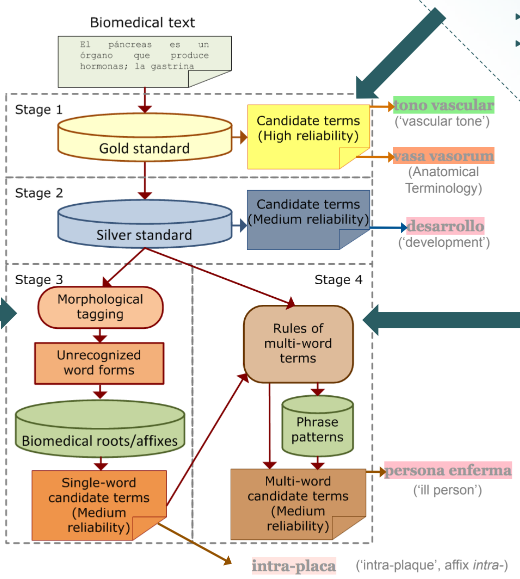


Figure 1. Processing pipeline and examples

Background

- Several non-commercial term extractors have been applied to the medical domain:
  - □ For English: TerMine [4] and the systems exposed in [5]
  - □ For French: TERMINO [6] and FASTR [7]
  - □ For Spanish: YATE [8] and TExtractor [9]
- At this stage, our system only focuses on term classification [1].

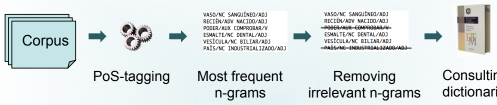
The gold standard and the silver standard lists

- For the lists of biomedical terms, two types of resources were used:
- □ **Corpora**: terms extracted semi-automatically [15] from the MultiMedica corpus [16].
- □ **Lexical resources**: terms that were not found in the corpus were semi-automatically curated from general and specialized resources [17-20].
- The **gold standard list** gathers terms registered in leading medical dictionaries [10-13].
- The **silver standard list** includes:
  - □ Terms not registered in medical dictionaries, but found in leading books and papers. We used the Google Books corpus [21] to reference each item.
  - □ Terms that were registered in medical dictionaries, but have:
    - ‣ A very general sense: e.g. *posibilidad*, 'probability'.
    - ‣ Some senses not restricted to medicine:
      e.g. *valorar* ('to titrate', chemistry) ~ 'to assess'.
  - Lists include inflected forms to cope with variants of terms:

crónico ('chronic')	→	*crónico, crónica, crónicos, crónicas*
curar ('to heal')	→	*curado* ('healed'), *curando* ('healing')…

Rules for multi-word terms and phrase patterns

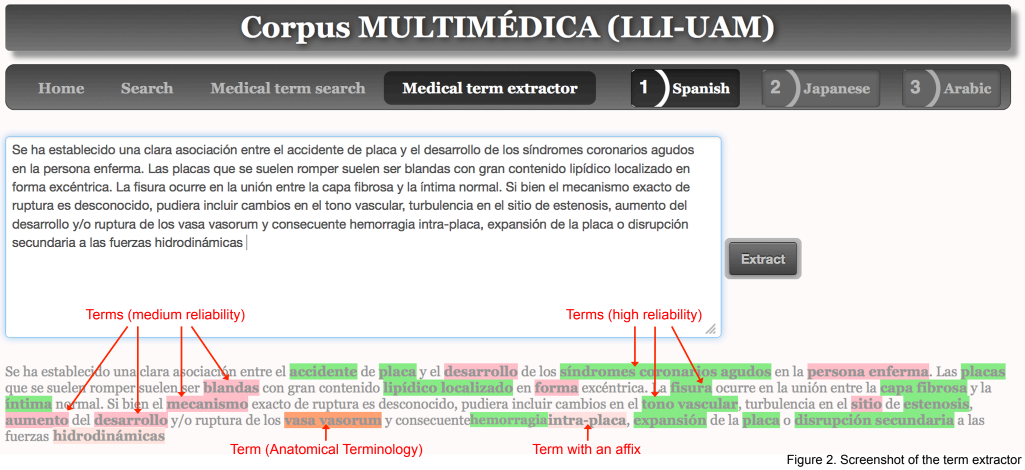
- They were obtained semi-automatically in the following way:



- Rules cover, among others, the following combinations:
  - □ N + ADJ: e.g. *cólico nefrítico* ('renal colic').
  - □ N + PREP + N: e.g. *enfemedad por depósito* ('storage disease').
  - □ N + N: e.g. *virus Coxsackie* ('Coxsackie virus').
- Context/phrase patterns serve as templates where terms appear: e.g. *persona con* + noun: *persona con* <u>*demencia*</u> ('person with <u>dementia</u>')

The interface

- http://cartago.lllf.uam.es/corpus3/extractor.pl?menu=extractor



Se ha establecido una clara asociación entre el accidente de placa y el desarrollo de los síndromes coronarios agudos en la persona enferma. Las placas que se suelen romper suelen ser blandas con gran contenido lipídico localizado en forma excéntrica. La fisura ocurre en la unión entre la capa fibrosa y la íntima normal. Si bien el mecanismo exacto de ruptura es desconocido, pudiera incluir cambios en el tono vascular, turbulencia en el sitio de estenosis, aumento del desarrollo y/o ruptura de los vasa vasorum y consecuente hemorragia intra-placa, expansión de la placa o disrupción secundaria a las fuerzas hidrodinámicas

Se ha establecido una clara asociación entre el **accidente** de **placa** y el **desarrollo** de los **síndromes coronarios agudos** en la **persona enferma**. Las **placas** que se suelen romper suelen ser **blandas** con gran contenido **lipídico localizado** en **forma** excéntrica. La **fisura** ocurre en la unión entre la **capa fibrosa** y la **íntima** normal. Si bien el **mecanismo** exacto de ruptura es desconocido, pudiera incluir cambios en el **tono vascular**, turbulencia en el **sitio** de **estenosis**, **aumento** del **desarrollo** y/o ruptura de los **vasa vasorum** y consecuente**hemorragiaintra-placa**, **expansión** de la **placa** o **disrupción secundaria** a las fuerzas **hidrodinámicas**

Figure 2. Screenshot of the term extractor

Conclusions

- Our extractor is complementary to existing systems for Spanish, but depends more on the lexicon and the tagger.
- The next step will be to evaluate its performance.
- To improve the usability, domain experts should assess other user-related aspects such as:
  - □ The intuitive use of the interface.
  - □ The understandability and the quality of the results.

Acknowledgements

This work was funded by the MultiMedica project (TIN2010-20644-C03-03) and the MA2VICMR project (S2009/TIC1542).

References

[1] Krauthammer, M., and G. Nenadic. 2004. Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37: 512-526.
[2] Kageura, K., and B. Umino. 1996. Methods of automatic term recognition: A review. *Terminology*, 3(2): 259-289.
[3] Ananiadou, S., and S. G. Nenadic. 2006. Automatic terminology management in biomedicine. In S. Ananiadou and J. McNaught (eds.) *Text Mining for Biology and Biomedicine*, chapter 4, 67-97. Boston, MA: Artech House.
[4] Frantzi, K., Ananiadou, S. and Mima, H. 2000. Automatic recognition of multi-word terms. *Intern. Journal of Digital Libraries*, 3(2): 117-132.
[5] Alexopoulou, D., T. Wächter, L. Pickersgill, C. Eyre, and M. Schroeder. 2008. Terminologies for text-mining; an experiment in the lipoprotein metabolism domain. *Bioinformatics,* 9(4).
[6] Plante, P., and L. Dumas. 1989. Le dépoulliment terminologique assisté par ordinateur. *Terminogramme,* 46: 24-28.
[7] Jacquemin, C. 1996. A symbolic and surgical acquisition of terms through variation. *Connectionist, Statistical and Symbolic Approaches to Learning for NLP*. *LNCS,* vol. 1040, 425-438.
[8] Vivaldi J. 2001. *Extracción de candidatos a término mediante la combinación de estrategias heterogéneas*. Ph.D. thesis. Univ. Politécnica de Catalunya.
[9] Valderrábanos, A.S., A. Belskis, and L. Iraola. 2002. TExtractor: a multilingual terminology extraction tool. www.bitext.com/prensa/ART_EN_textractor_a_multilingual_terminology_extraction_tool.pdf [Accessed: 23/09/2013]
[10] RANM (Royal National Academy of Medicine) 2011. *Diccionario de términos médicos*. Madrid: Editorial Médica Panamericana.
[11] Dorland. 2005. *Diccionario enciclopédico ilustrado de medicina*. 30th edition. Madrid: Elsevier, D. L.
[12] Gonzalo Sanz, L. Mª. (coord.) 1999. *Diccionario Espasa Medicina.* Madrid: Espasa, S.L.
[13] Cortés-Gabaudan, F. (coord.) 2007-2013. *Dicciomed*. http://dicciomed.eusal.es [Accessed: 23/09/2013]
[14] Moreno-Sandoval, A., and J. M. Guirao-Miras. 2006. Morpho-syntactic Tagging of the Spanish C-ORAL-ROM Corpus. In Y. Kawaguchi et al. (eds.) *Spoken Language Corpus and Linguistic Informatics*, 199-218. Amsterdam: John Benjamins.
[15] Moreno-Sandoval, A., L. Campillos-Llanos, A. González-Martínez, J. Mª. Guirao-Miras. 2013. An affix-based method for automatic term recognition from a medical corpus of Spanish. *Proceedings of the 7th Corpus Linguistics Conference 2013. Lancaster University*.
[16] Moreno-Sandoval, A., and L. Campillos-Llanos. 2013. Design and Annotation of MultiMedica – A Multilingual Text Corpus of the Biomedical Domain. *Procedia - Social and Behavioral Sciences*, Vol. 95, 33-39. Amsterdam: Elsevier. http://dx.doi.org/10.1016/j.sbspro.2013.10.619.
[17] Stichele, R. V. (coord.) 1995. Multilingual Glossary of technical and popular medical terms in nine European Languages. Heymans Institute of Pharmacology, University of Gent. http://allserv.rug.ac.be/~rvdstich/eugloss/welcome.html, 1995 [Accessed: 23/09/2013]
[18] Federative Committee on Anatomical Terminology. 2001. *Terminología Anatómica – International Anatomic Terminology.* Madrid: Panamericana.
[19] Real Academia Española. 2001. *Diccionario de la Real Academia*. Electronic version.
[20] Yetano Laguna, J., and V. Alberola Cuñat. 2003. *Diccionario de siglas médicas*. Madrid: Publicaciones del Ministerio de Sanidad y Consumo.
[21] Google Books. http.//books.google.com
[22] Jiménez Arias, Mª. E. 2012. Afijos grecolatinos y de otra procedencia en términos médicos. *MEDISAN,* 16(6): 1005-1021.
[23] López Piñero, J. Mª., and Mª. L. Terrada Ferrandis 2005. *Introducción a la terminología médica*. Barcelona: Masson.
[24] Sánchez González, M. A. 2012. *Historia de la medicina y humanidades médicas*. 2nd ed. Barcelona: Elsevier/Masson.
[25] WHO. 2011a. The use of stems in the selection of International Nonproprietary Names (INN) for pharmaceutical substances. http://apps.who.int/medicinedocs/documents/s19117en/s19117en.pdf [Accessed: 21/09/2013]
[26] WHO. 2011b. International Nonproprietary Names (INN) for biological and biotechnological substances (a review). http://apps.who.int/medicinedocs/documents/s19119en/s19119en.pdf [Accessed: 23/09/2013]
[27] American Medical Association. United States Adopted Names. 2013. www.ama-assn.org/resources/doc/usan/stem-list-cumulative.pdf
[28] American Medical Association. United States Adopted Names. 2013. www.ama-assn.org/resources/doc/usan/new-stem-list.pdf