ARTICLE IN PRESS IJIMAI journal

Biomedical Term Extraction: NLP Techniques in Computational Medicine

Antonio Moreno Sandoval[1], Julia Díaz[1], Leonardo Campillos Llanos[2], Teófilo Redondo[3]*

[1] Universidad Autónoma de Madrid (UAM) / Instituto de Ingeniería del Conocimiento (IIC) (Spain)
[2] Laboratoire d'Informatique pour la Mécanique et les Sciencies de l'Ingénieur (LIMSI-CNRS) (France)
[3] Ayming España (Spain)

Received 12 November 2017 | Accepted 4 February 2018 | Published 6 April 2018

Abstract

Artificial Intelligence (AI) and its branch Natural Language Processing (NLP) in particular are main contributors to recent advances in classifying documentation and extracting information from assorted fields, Medicine being one that has gathered a lot of attention due to the amount of information generated in public professional journals and other means of communication within the medical profession. The typical information extraction task from technical texts is performed via an automatic term recognition extractor. Automatic Term Recognition (ATR) from technical texts is applied for the identification of key concepts for information retrieval and, secondarily, for machine translation. Term recognition depends on the subject domain and the lexical patterns of a given language, in our case, Spanish, Arabic and Japanese. In this article, we present the methods and techniques for creating a biomedical corpus of validated terms, with several tools for optimal exploitation of the information therewith contained in said corpus. This paper also shows how these techniques and tools have been used in a prototype.

Keywords

Biomedical Terminology, Natural Language Processing, Term Recognition, Information Extraction.

DOI: 10.9781/ijimai.2018.04.001

I. Introduction

TERMINOLOGY is a branch of Applied Linguistics whose main goal is the creation of specialized or technical language. Thematic domains are by themselves the realm of a specific sublanguage, adapted to designing the concepts in each topic or knowledge area. In this sublanguage, many exclusive terms coexist with those that have acquired meanings other than those common to the general language. Elaborating a terminological dictionary is a multidisciplinary task that requires contributions from both lexicographers and subject matter experts in order to define a specific term in the most precise way.

Some fields, that show a rapid evolution in the area, need to include new concepts at a very fast pace and require constant work in detecting those concepts and proceeding to normalize or standardize. Medical terminology is one such field where the sheer number of specialized terms exceeds the usual number of specialized terms in other knowledge areas, when taking into account both simple lemmas and compound forms. New terms and concepts are generated in a very dynamic fashion and this needs computing tools such as automatic recognizers (as part of the information extraction process). These applications analyze digital texts and identify candidates that can be terms of a given domain, so it can be validated by an expert (akin to a supervised learning process).

* Corresponding author.
E-mail address: teo.redondo@gmail.com

II. Basic Concepts and Techniques

A. Automatic Recognition of Terms and Concepts in Digital Texts

1. Objectives

Term Extraction or Automatic Term Recognition (ATR) is a field in language technology that involves "extraction of technical

Please cite this article in press as:
A. Moreno Sandoval, J. Díaz, L. Campillos Llanos, T. Redondo. Biomedical Term Extraction: NLP Techniques in Computational Medicine, International Journal of Interactive Multimedia and Artificial Intelligence, (2018), http://dx.doi.org/10.9781/ijimai.2018.04.001

terms from domain-specific language corpora" [1], or identifying term candidates in texts of lists of words [2]. The original interest lies not in creating terminology resources, but in extracting words or expressions that identify topics in a document. This use is typical when working with medical texts [3], as a tool for information extraction and text mining [4]. Different NLP techniques are described in detail in Moreno Sandoval and Redondo, 2016 [5].

In order to detect new terms and concepts, texts that are recent and also representative are required. **Corpus Linguistics**, with an ever-growing influence in recent years due to the availability of large datasets, has the compilation of texts of a given domain as one of the main objectives. Documents must be digital, so searches or other computational handling can be performed, such as morphosyntactic annotation and statistical analysis. Once the medical corpus is created, the automatic recognizer will extract a number of candidate terms.

In Terminology there are well-established methodological traditions to enhance lexicography resources and build data banks following standard procedures [6]. However, the speed at which new terms (neologisms) are created in certain knowledge areas makes this approach extremely costly. It is at this precise point where systems for automatic extraction of terms are of great help, but always considering that the final "word" lies in the hands of the area expert.

2. Domain and Difficulties

In the classical definition of Terminology, a *term* or *terminological unit* is a linguistic expression of a concept in a specialized domain [7]. From the perspective of ATR, the task consists in identifying how a term is defined under the following lines [8]:

- *Unithood*: the degree of cohesion or stability of words in an expression.
- *Termhood*: the degree of specificity of the term with respect to the knowledge area. For instance, *hepatic* is related to a medical domain, not to aeronautics or space.

The main difficulties in *Unithood* are located in recognizing syntagmatic structures and the boundaries between words in compounds (*multiword terms*). For instance, the ATR should detect as candidate terms *infarto* (*infarct* or *heart attack*), *infarto de miocardio* (*myocardial infarct*) and *infarto agudo de miocardio* (*acute myocardial infarct*), but not *posible infarto* (*possible infarct*).

In *Termhood* it is typical to find polysemic terms that do belong in different knowledge areas. For instance, *nuclear* is a term both in Physics and in Genetics or Biology. Using resources of terms in other areas can lead to achieving wrong results.

In addition, there are two phenomena that make things more complicated in recognizing biomedical terms: *variation* and *homonymy*. In the former case, the problem appears when a knowledge area holds a great number of formal variations of the same term. This affects both simple terms (*aterosclerosis ~ ateroesclerosis*) and compound terms (*carcinoma microcítico de pulmón ~ carcinoma microcítico pulmonar*). Ananiadou and Nenadic [9] distinguish five types of terminological variation, that are basically just formal alternatives:

- Ortography: *alfa-amilasas ~ amilasa alfa ~ -amilasa*
- Morphology: *obsesiva-compulsiva ~ obsesivo-compulsivas*
- Lexicon: *infarto de corazón ~ infarto cardíaco*
- Structure: *virus del papiloma humano ~ papilomavirus humano*
- Acronyms and abbreviations: *SST ~ ST*, both referring to *somatostatina*

In addition to constant creation of neologisms in the biomedical area, foreign influence is sourcing new variations. Linguistic calques or loan translations with little or no adaptation to the new language are one such example. In biomedical texts in Spanish, terms like *bypass*, *by pass* and *baipás* appear quite naturally. Another example is the increasing inclusion of modifiers to already existing terms: *deficiencia de hexosaminidasa A ~ deficiencia total de hexosaminidasa A*. An essential task for both human experts and ATR is to normalize formal variations representing the same concept. The existence of multilingual ontologies and metathesaurus, such as those integrated in UMLS (*Unified Medical Language System*) [10], provide an essential contribution. This resource includes several thesaurus and terminological works: *Medical Subject Headings* (MeSH) [11], *Systematized Nomenclature of Medicine – Clinical Terms* (SNOMED-CT) [12], or version 10 of the International Classification of Diseases (ICD-10) [13]. UMLS contains unique identification codes associated to each terminology variation in different resources. For example, code C0817096 refers to *breast* or *thoracic cavity* in MeSH and also the term *thoracic* or *thorax* in SNOMED-CT.

On the other hand, term homonymy, especially acronyms, is another challenge for ATR. For instance, *IM* can refer to both *insuficiencia mitral* and *infarto de miocardio*. Without the contribution in contextual and domain knowledge from terminology experts it is very difficult to decide in which concept the acronym belongs. Some systems try to solve this by restricting the lexicon to a specific field [14], but in several cases, this presents problems since limits or boundaries between biomedical areas are rather fuzzy.

3. Approaches and Methods

Although several authors distinguish basically between linguistic techniques and statistical techniques [15], in term recognition several heterogeneous methods are combined so as to achieve the best results, as will be shown below. In a conventional way, the different approaches towards ATR are classified along four types: a) dictionary-based, b) rule-based, c) statistics-based and machine learning, d) hybrid [16].

- Dictionary-based approaches use digital resources such as grammar words without content (also known as *stop words*), as well as ontologies, glossaries and domain thesaurus. These lists allow the filtering of the text: with the former, words of no interest get eliminated and with the latter, terms are singularly identified. This approach is the most efficient and simple, but it tends to be rather incomplete and it is not available in all domains nor for all researchers. An example is detailed in Segura-Bedmar et al [17], where the UMLS metathesaurus and other name lists of generic drugs were used, with the objective of identifying and classifying pharmacological names in biomedicine texts.
- Rule-based approaches use pattern analysis of the term creation (for example, compounds by addition, hyphenated compounds, syntagmatic patterns) and grammar knowledge (morphological analysis of the terms, lists of lemmas and affixes). This approach has abundantly been used from 1990 onwards. Morphological description of lemmas and affixes, for instance, has been used to detect medical terms [18], and other researchers used concatenated category pattern-based algorithms [19]. For Spanish, noun phrases (or nominal syntagmas) have been used for medical terms extraction [20]. In general, an effective strategy can be achieved if work focuses on a language with Greek and Latin bases to create new terms. This, however, is not the case in all domains nor all languages [21].

With respect to statistics-based techniques, the foundation lies in measuring the degree of distinctiveness [22] of a word or lemma in a specialized context in contrast with their frequency in a general corpus. The two most common are the *log-likelihood ratio test* [23] and the *logDice* metrics used in The Sketch Engine [24]. The central idea of these techniques is to know which words or terms over- or under-used in the corpus for analysis when compared to the frequency of the same words in a reference corpus. In our case we take a corpus

of medical terms (*MultiMedica*) and compare it to the *Reference Corpus of Current Spanish* (Corpus de Referencia del Español Actual – CREA), that contains a balanced set of texts coming from different domains and linguistic registers. However, there are other statistics-based techniques, such as Mutual Information Metric [25] or the use of Distributional Semantics and lexical collocation [26]. For Spanish, the experiment for term detection has been run on a corpus of scientific texts by using n-grams and their likelihood and distribution in such corpus [27]. An algorithm to analyze lexical, morphological, syntactic features has been used to compare this with a reference corpus [28].

Machine Learning's approaches are a special type of using statistical techniques that consist in training algorithms with data from corpus that has been previously annotated by experts in the knowledge area. Machine Learning algorithms (among others, *Hidden Markov Models* – HMM, *Support Vector Machines* – SVM, or *Decision Trees*) identify features in the annotated terms and apply them to a new data set. The most basic type is called *classifier*, that divides words in a text between terms and non terms. Lastly, current advances in neural network research are yielding promising methods for sequence modeling tasks (such as PoS or NER). Biomedical entity recognition is being enhanced through *Recurrent Neural Network* (RNN) models, namely *Long-Short-Term Memory networks* [29] and hybrid architectures combining *Conditional Random Fields* (CRFs) [30], attention mechanisms and language modelling [31], among others. These kinds of approaches use vector representation of words along with their occurrence context or frequency distribution (word embeddings) [32] [33].

Hybrid techniques combine two or more techniques mentioned above. The most usual case uses a linguistic approach (dictionaries and rules of term formation) and a statistical metric, a hybrid method already developed for Spanish [34].

III. BIOmedical NLP Use Case – MultiMedica

MultiMedica (Multilingual Information Extraction in Health Domain and its Application to Scientific and Informative Documents) was a coordinated project between the LABDA research group (UC3M), the GSI group (UPM) and the LLI (UAM), the latter group being in charge of the following tasks:

- Compilation of a specialized corpus of texts about health topics. The corpus gathers documents in three languages with different genetic and typological features: Arabic, Japanese and Spanish
- Morpho-syntactic tagging of the corpora,
- Contrastive research on term formation,
- Development of an automatic term extractor,
- Design of a web-based search tool.

A. The Corpus

The initial experiment used a corpus of text in Spanish, a corpus that was later extended to include text in Japanese and Arabic. The subcorpus consists of 4,200 documents with a total of 4 million words. The textual typology covers from general articles written by doctors with a no-specialist audience in mind (typically reviewed and edited by journalists) up to scientific texts for a specialized audience (i.e. healthcare professionals). Technical / specialized texts prevail over general content (more than 80% correspond to technical texts), with most of the medical specialties represented in a balanced number. This qualifies the corpus as a reliable source to produce a list of valuable candidate terms. As an interesting addition, the corpus was morphosyntactically annotated (category and lemma), in order to allow for searches and agreement [35].

The MultiMedica corpus has gathered 51,476 biomedical texts in different genres (popular and technical texts) written in Spanish, Japanese and Arabic. The tool enables two main functions: queries in the medical corpus and medical term extraction of an input text. The tool presents a web interface for ease of use.

Table I outlines the composition of the corpus (number of texts and words/characters):

TABLE I. Summary of the MultiMedica Corpus Data

Subcorpus	Documents	Word or characters
Japanese	3,746	1,131,304
Arabic	43,526	2,559,323
Spanish	4,204	4,031,174
TOTAL	51,476	7,721,801

The Spanish corpus is made up of three subcollections: The *Harrison* subcorpus assembles professional and scientific texts written by medical doctors; the *OCU-Salud* subcollection gathers journalistic texts written by medical doctors and edited by journalists; and finally, the *Tu otro médico* subcorpus collects popularized texts from encyclopaedic articles written by professional doctors for non-specialists. Regarding the Arabic corpus, gathering documents was made difficult by the fact that most medical doctors in the Arabic-speaking world write articles in English. Most documents in this subcorpus were articles and popularized news collected from *Altibbi*, a Jordanian medical website equivalent to *Healthline* in the United States. The remaining texts were drawn from the health sections of the following journals: *Al-Awsat* (from Saudi Arabia), *Youm7* (from Egypt), and *El Khabar* (from Algeria).

In relation to the Japanese corpus, only abstracts of five medical journals were collected, due, again, to the lack of availability of data. However, the texts gather contents on different specialties: Oriental medicine in Japan (from the journal *Kampo Medicine*), infectious diseases (*Kansenshogaku Zasshi*), liver diseases (*Kanzo*), otolaryngology, (*ORLTokyo*), and obstetrics (*Sanfujinka no shinpo*).

B. Methodology and Pipeline

We summarize some experiments carried out on ATR of medical terms (full details are explained in another paper) [36]. For the initial experiment only identifying simple terms (those with one single word, such as *aspirina* or *ADN*) or words as part of a compound (*ascórbico* in *ácido ascórbico*, or *Down* in *síndrome de Down*) was considered. The objective was to evaluate which of the previous strategies would provide the best results. The process followed three steps (see Fig. 1):

1. Preselect candidates by means of one of the three methods
2. Filtering of term candidates by means of a list of biomedical lemmas and affixes
3. Manual check of each candidate term by consulting bibliography or other resources

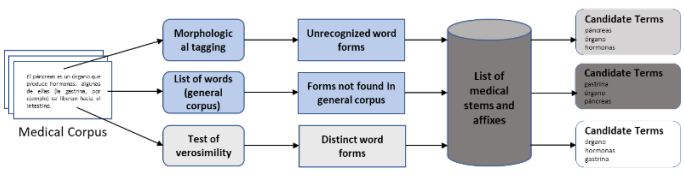

Fig. 1. Phases of the term extractor [36].

1. Preselect Terms Following each Method

Each method for term candidate extraction is not based on a similar strategy, and consequently the list obtained from each has a different size, although it is applied to the same data set. However, obtaining more candidates does not mean that the rate of success increases.

The first method uses a morphological tagger. It is an example of

the rule-based type: the analyzer contains a set of recognition rules and analysis of words in Spanish. Here only words with the tag "unknown" (*desconocido*) are of interest, because medical terms are assumed to have a morphological structure not included in the analyzer used: GRAMPAL [37] covers a lexicon with more than 50,000 lemmas of general use and is capable of analyzing more than 500,000 inflection forms. Obviously, GRAMPAL contains a large number of medical terms that have found their way into the common lexicon, as would be collected in any reference dictionary (DRAE or Maria Moliner being the most typical ones). But similarly, most of the specific and technical terms of the domain are not included (i.e, *ADN* or *distal*). After an initial run over the corpus with 4 million words, a total of 22,413 "unknowns" were produced, which then were listed as term candidates.

The second method uses a corpus-based strategy: words in MultiMedica are compared with those in the Spanish general corpus (CREA). Given that it is a large and balanced corpus, it can be considered as a reliable reference of general use of words in Spanish. CREA contains no less than 150 million words and around 700,000 different forms. However, this list presents around 50% of *noisy* words for the experiment: foreign words, orthographic and typographic mistakes as well as proper nouns. A task for cleaning up the list reduced the total number to 350,000 distinct forms. A lot of medical terms of general use (as opposed to technical or professional use) appear on this list, and, additionally, proper nouns such as Down or Alzheimer, that are part of compound terms, were removted. However, when reviewing the number of proper nouns that are not relevant, we chose to eliminate all of them. After this process, only a total of 23,239 candidate terms were included in the list, which are words that are not in the reviewed list in CREA. To provide additional context to the relative size that has been handled, a lexicon like GRAMPAL with 50,000 lemmas generates around 150,000 different forms more than those in a corpus like CREA with more than 150 million words.

The third method uses a purely statistical technique: the Log-Likelihood (LLH) is applied to identify distinct words in the medical corpus [38]. This test is always used in programs checking agreement (such as, Wordsmith or AntConc) to extract keywords in a text. The process performs a comparison of the occurrence frequency between the words in a given corpus with those in a reference corpus. In this case, MultiMedica was compared with the CREA version already pre-processed (see above). To achieve 99.9% of confidence rate, we applied a threshold of significance in 10.83. As a result, the list of candidate terms contains only words with a test value above 10, which renders a list of just 8,667 candidate terms.

Several natural language processing (NLP) techniques were utilized. First, each collection was processed and tags for part-of-speech were included. The Spanish subcorpus was tagged by using GRAMPAL [39], already mentioned. The tagging process is semisupervised, as it requires manual revision to ensure annotation quality. A random sample representing 5% of the popularized texts in Spanish was revised twice to compute the inter-annotator agreement (IAA) value. This was assessed by computing the F-measure, as exposed in Hripcsak and Rothschild (2005) [40], and it was found that both annotators agreed in about 98 per cent of the texts.

Herrero et al. (2014) [41] explain the methodology followed in the creation of the morphological tagging for the Japanese corpus. After considering three different taggers (ChaSen, Mecab and Juman), Juman was chosen, because it provides good segmentation and a wider range of morphological information. Similarly, the Arabic corpus was automatically annotated using the PoS tagger MADA+Tokan [42]. Finally, the tagged texts were indexed for all languages to enhance online queries.

2. Filtering with a List of Affixes and Lemmas

The next step was to create lists of medical terms for each language. The Spanish list was compiled semi-automatically, combining rule-based, tagger-based and statistical approaches [43], as already described in the section above. A gold standard list included terms that appeared in leading medical dictionaries (e.g., RANM 2011, Dorland 2005). A silver-standard list gathered terms that were found only in biomedical books and journals.

Regarding Japanese, a single list was compiled with terms from several medical dictionaries: *Online Life Science Dictionary* [44] and *Japanese-English-Chinese Dictionary* (1994). As for Arabic, the final list is a combination of full terms translated from English resources (SNOMED and UMLS) and a list of Arabic words equivalent to Spanish prefixes and suffixes, such as *-itis*, *cardio-*, etc. [45].

An initial review of the candidate terms shows that some kind of filter must be applied to the list since it contains words not included in the lexicon of the morphological analyzer nor in the CREA list, but that are words of common usage (i.e. *tabúes* or *vinculador*). To further enhance the precision of the selected terms a program was applied for identifying affixes and lemmas of medical terms. The program contains 2,128 items, including orthographic variations such as *aden-* or *adeno-*:

- Greek and Latin affixes in the medical knowledge area (i.e. *cardio-*, *-itis*) and frequent medical lemmas (i.e. *pancrea-*), collected from several sources of medical terms [46]. To avoid false positives, highly frequent affixes were removed from the list, because they are not restricted to the biomedical domain (such as *pre-* or *-able*).
- Lemmas and affixes for identifying pharmacological compounds (*-cavir*) and biochemical substances (*but-* or *-sterol*). All of them have been compiled from lists proposed and approved by the *World Health Organization* (WHO) [47], as well as lists approved by the *American Medical Association* (AMA) [48] for clinical compounds official denominations. As most of scientific English affixes have a unique correspondence with equivalent Spanish affixes, the adaptation was direct with a minimal effort, especially for those ending in vowels such as *-ine* > *-ina* (*creatine* > *creatina*).

In order to obtain the final list, all possible variations of each affix and lemma have been generated. On one side, graphic variations due to diacritics (i.e. tilde), such as *prόst-* (as in *próstata*) and *prost-* (as in *prostático*). On the other hand, variations due to an epenthetic vowel: *escoli- scoli-*. And finally, variations due to gender and number inflection, such as the suffix *-génico* can have four different forms: *-génico*, *-génica*, *-génicos* and *-génicas*.

The program that compares affixes with the candidate terms first compares each candidate with all affixes appearing in two different lists (prefixes and suffixes). When a candidate term contains a biomedical affix or lemma, it is considered a potential term. Fig. 1 above displays the whole process.

3. Manual Verification of each Proposed Term

The last phase performs a manual review of all the candidate terms, by confirming or rejecting each term. The final result can be called a *gold standard* or set of reference terms with all validated forms. For a term to be validated, it must appear in a well-known and accepted medical source. In order to avoid subjectivity, the decision is based on consulting the following reference works, and in this order:

- Diccionario de Términos Médicos [49]: with almost 52,000 terms
- Diccionario Médico Enciclopédico Dorland [50]: more than 112,000 terms
- Diccionario Espasa Medicina [51]: 18,000 terms (collected by medical professionals in the Universidad de Navarra)
- Dicciomed [52]: around 7,000 terms (with a historic and

etymological approach).

Similarly, terms found regularly in journals and books of biomedical research have been validated and included in the list. Table II is a summary of the classification criteria followed in order to accept or reject a term.

TABLE II. Four Types of Terms

	Term classification	Examples
Accept	List 1 – terms with an entry in a medical reference dictionary List 2 – terms without an entry in a medical reference dictionary, but found in books and scientific articles	*páncreas, ADN …* *RAS, cisteínico …*
Reject	List 3 – terms rejected by specialists, due to orthographic or typographic errors or poor adaptation into Spanish List 4 – non-biomedical terms	**perirenal, *croup…* *Aragón, Pfizer …*

Biomedicine is an extremely wide area for research, and establishing clear-cut boundaries to the domain is almost impossible. The terms of the golden standard come in such fields as Anatomy (*hígado* > *liver*, *nefrona* > *nephron*), Microbiology (*cilio*, "*Escherichia*"), Genetics (*transcripción*, *ARN*), Oncology (*oncogén*, *leukemia*), Biochemistry (*fosforilación*, *amina*), Pharmacology (*aspirina*, *prozac*), History of Medicine (*frenología*, *miasma*), or Surgery and other medical techniques or procedures (*tomografía*, *maniobra*), among others. Terms from other knowledge areas not strictly related to biomedicine, but common in medical texts were also accepted. For instance, concepts referring to statistical metrics (*variable*, *significance*), agents involved in a disease, like poisonous animals or environmental conditions (*anopheles*, *vipéridos*, *contaminación*) or plants producing pharmacological substances (*Vinca*, *cornezuelo*). In total, the list contains 24,639 terms.

4. Developing a Term Extractor for Each Language

Each language required a different approach in order to build the term extractor. The Spanish extractor uses lists of terms, medical roots and affixes, the GRAMPAL tagger, and rules for multi-words and context patterns. The processing of the input text to detect candidate terms is as follows. First, a dictionary-based method that relies on pattern matching is applied. Each item found in the gold standard list is marked as a highly reliable candidate term (e.g., *pulmón*, 'lung'). Likewise, each term found in the silver standard list is selected as a medium reliable candidate term (e.g., *secundario*, 'secondary'). In the third stage, those words that were not found in any list are POS-tagged through the GRAMPAL tagger. Unrecognized items (i.e., words not included in the lexicon of the tagger, which was designed for the general language) are then filtered using a list of biomedical roots and affixes (e.g.,*hemat(o)-*, an affix related to blood). In this way, for example, an adverb such as *hematológicamente* ('hematologically') may be recognized as a term and highlighted with medium reliability. The last stage involves applying multi-word formation rules to the previous list of candidate terms. If any element of the multi-word candidate term has medium reliability, the whole unit is highlighted as such. For example, if the term *complejo* ('complex,' medium reliability) and *amigdalino* ('tonsillar,' high reliability) are recognized, a multi-word rule will join both terms in *complejo amigdalino* ('tonsillar complex') and mark it as a medium reliability candidate term. Fig. 2 outlines the architecture of the system.

The extractors for Japanese and Arabic follow a simpler procedure. The Japanese extractor performs an initial pattern matching throughout the dictionary, identifying those terms as highly reliable. Secondly, a series of rules are applied bearing in mind the agglutinative nature of the language. For example, if two dictionary terms are joined with a connective particle, it will be considered as a single multi-word term; also, if additional *kanji* characters are added to the initial or final part of a dictionary term, the extractor recognizes the whole string of characters as a single term. The terms detected using this rule-based procedure are classified as medium reliable ones. The Arabic language is mainly a dictionary-based extractor that recovers terms from the medical list created for this purpose.

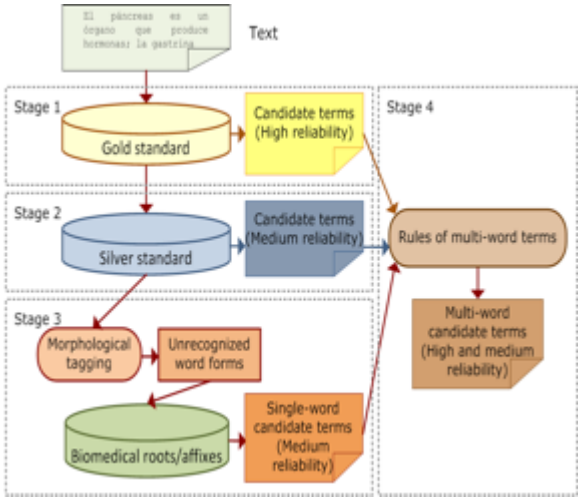

Fig. 2. Phased architecture of the Spanish term extractor [53].

Improvement in the term extraction in the future includes adding more medical terms, or codes from the *International Classification of Diseases version 10* (ICD-10) [54], the *Unified Medical Language System* (UMLS) and the *Systematized Nomenclature of Medicine – Clinical Terms* (SNOMED-CT) [55].

5. Interaction with the MultiMedica Corpus

Users can perform queries in the corpus in two ways: *simple word search* ("Search" tab, "Consulta" in the Spanish version) and *medical term search* ("Medical Term Search" tab, "Consulta de Términos Médicos" in Spanish). In addition, users can input a free text to detect and extract candidate terms in the domain ("Medical Term Extractor," "Extractor de Términos Médicos").

a) Word Search

Any word in the corpus can be searched according to form, lemma or part-of-speech (POS). For example, if the user inputs the lemma *cáncer*, the results may be *cáncer* or *cánceres* (respectively, 'cancer' or 'cancers'). The user has the option of looking up the collocations of the word as well as its frequency and log-likelihood value.

In the search results, frequency values are normalized per million words (hereafter, *pmw*). Counts are also compared to the frequencies in the *Corpus de la Real Academia Española* (CREA) corpus. This makes it possible to know the *distinctiveness* of the searched word in a specialized corpus and in relation to a general language corpus. For example, when the word *hepatitis* is searched, the normalized frequency in the MultiMedica corpus is 385.8 pmw, and 6.1 pmw in the CREA corpus. This shows that this token is highly related to this specialized genre. In contrast, if *corazón* ('heart') is searched, the normalized frequency in the MultiMedica corpus drops to 140.8 pmw, which is close to the normal frequency in the CREA corpus (125.3 pmw). This indicates that *corazón* appears with a similar frequency in a health and a general corpus. Since this is a polysemous word, other senses beyond the anatomical context are used in the general language (e.g., related to feelings, or as a synonym of 'nucleus' or 'core').

The word search for Spanish, Arabic, and Japanese are shown in Fig. 3, 4 and 5, respectively.

Fig. 3. Search medical terms in Spanish [53].

Fig. 4. Search medical terms in Arabic [53].

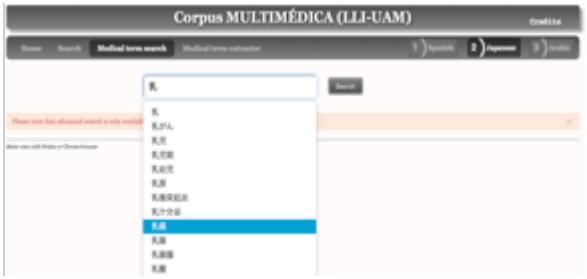

Fig. 5. Search medical terms in Japanese [53].

The search tool for the Spanish corpus also provides information about word distribution (i.e., its frequency in each type of text). This feature makes it possible to compare different text genres (popular vs. technical documents). If we search for *dolor de espalda* ('upper back pain'), the results show that this term is more frequent in popularized texts than in technical texts. However, when we search for *dorsalgia* (the technical synonym of 'dolor de espalda'), the results reveal that this term is restricted to academic documents.

b) Medical Term Search

The medical term search allows users to look up the most frequent medical terms in the corpus. An autocomplete function provides a list of all the possible terms that contain the typed letters introduced by the user. The list is based on the 5,000 more frequent terms in the corpus.

c) Medical Term Extractor

The medical term extractor detects candidate terms from an input text (Fig. 6 and 7). The tool highlights medical terms according to their level of reliability: high (terms included in the gold standard list) and medium (terms in the silver list). The user may also download the term list in text format for further use. In addition, terms that are found in the *BabelNet* dictionary [56] contain a hyperlink to this resource, which provides their translation in many languages.

Fig. 6. The medical term extractor for Spanish texts [53].

Fig. 7. A screenshot of the Japanese term extractor [53].

IV. Future Work

Biomedical Natural Language Processing (BioNLP) is receiving a growing interest from both academia and industrial specialized applications. The specific field of biomedical text mining is one of the most mature domains. Biomedical text mining, of which term extraction is just one area, is providing great advances in terms of widespread availability of expert-annotated text resources, biomedical term banks, and a great number of information extraction components. Biomedical text processing components have been published, covering various aspects, from tokenization approaches [57] to the creation of specialized tokenizers for biomedical texts [58]. Equally important are special linguistic and NLP tools for biomedical texts, such as POS taggers [59] or dependency-based parsers [60] for pure syntactic analysis (Enju/Mogura [61], GDep), which present biomedical domain models to create graphic representations of syntactic dependency relations. These syntactic relations are used to express bioentity relationships present in the text (such as protein-protein interactions [62]) in combination with recent machine learning techniques.

Current and future promising trends biomedical natural language processing include the following: to rank a classification of topics of relevance in a text after term identification [63]; detection of different types of bioterms applying semantic roles; indexing of documents to terms and concepts from controlled vocabularies and corpora, as in the case of Multimedica, which may build bioontologies [64] to be applied in other domains, and extracting relationships between biomedical terms (protein or gene relations [65]). Another area of biomedical term extraction research field is the detection of associations between

disease concepts and actual disease areas [66], like in the bioontologies mentioned above.

As already covered in the present paper, the first step or phase in most biomedical term identification is to locate mentions of biological entities of interest or terms, in the sense used here. Work in biomedical natural language processing is very much dependent on research in the biomedical sciences, which have recently focused on the study of a set of concepts, like genes, proteins, chemicals, drugs or certain diseases. Tools, like the term extractor and search engine presented here, can be a great help for a more efficient way of finding information in documents, that build up the corpora, and then characterize those concepts so researchers can reach deeper insights into their own domains.

One example of the importance given to this topic are initiatives like BioASQ [67]. This is a European Commission-funded project under the FP7 programme, whose goal is to organize challenges on biomedical semantic indexing and question answering (QA). The challenges include tasks relevant to hierarchical text classification, machine learning, information retrieval, QA from texts and structured data, multi-document summarization and many other areas.

In the last couple of years, the work in biomedical NLP was dominated by applications of deep learning to: punctuation restoration [68], text classification [69], relation extraction [70] [71] [72] [73], information retrieval [74], and similarity judgments [75], among other exciting progress in biomedical language processing. For a more detailed exploration of recent topics, the BioNLP Annual Workshop [76] covers the most researched and debatable areas.

Term extraction has other applications beyond BioNLP, as is the case with chemical terminology, legal texts, the engineering documentation for the oil & gas industries, or research of new drugs in the pharma industries, just to name but a few.

V. Conclusion

This paper has covered a use case of term extraction in the BioNLP domain, starting form a description of the basic techniques used to the methodology followed in the creation of a multilingual corpus of medical texts for medical term extraction, their morphological annotation and further indexation, the actual term list extraction and the development of an online tool so a user can reach the information and use it for consultation or clarification of the medical term. Three languages were selected: Spanish, Arabic and Japanese, languages so different genetically and typologically, that specific approaches and tools had to be chosen for each of them. This led to identifying several problems for the computational treatment of medical terms in these languages, for example, the lack of language resources in medical NLP for Arabic (either professional texts or electronic dictionaries). In this sense, MultiMedica is a pioneering effort in this Biomedicine domain and for this combination of languages. It has also provided an interesting typological insight into how languages behave within the medical domain. Each of the three languages presented different challenges when developing the extractor: the variation in inflection of Spanish terms, variation in the Arabic writing system or word segmentation in Japanese due to the lack of white spaces between words. Even though the initial steps of creating the corpus, tagging, and development of a medical term list was approximately equal in the three languages, the processing of the texts and the creation of the extractor had to be adapted to the specificities of each language.

Looking into the future it is reasonable to expect that the corpus and online tools may provide the users with a good amount of data for future linguistic research into biomedical discourse and may be used for many other use cases. The term extractor may fulfil terminologists' and translators' needs by helping them identify term candidates and finding their equivalents in other languages. In addition, health professionals, in the broad sense, including clinical, pharma or chemical professionals, and medical students could make use of this interface to seek and translate biomedical information online.

References

[1] Z. Zhang, J. Iria, C. Brewster, and F. Ciravegna. "A comparative evaluation of term recognition algorithms", in *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08*), Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrecconf.org/proceedings/lrec2008/. 2008

[2] K. Kageura and B. Umino, "Methods of automatic term recognition: A review", Terminology, 3(2) (Ámsterdam, 1996), págs. 259-289; and M. Krauthammer y G. Nenadic, "Term identification in the biomedical literature", *Journal of Biomedical Informatics, 37*, pp. 512-526, Ámsterdam, 2004

[3] S. Ananiadou and J. Mcnaught (eds.), *Text Mining for Biology and Biomedicine*, Artech House, Boston, MA, 2006.

[4] K. B. Cohen, "Biomedical Text Mining", in N. Indurkhya and F. J. Damerau (eds.), *Handbook of natural language processing, 2ª ed*., Chapman and Hall, Boca Raton, pp. 605-625, 2010.

[5] A. Moreno Sandoval and T. Redondo. "Text Analytics: the convergence of Big Data and Artificial Intelligence". *International Journal of Artificial Intelligence and Interactive Multimedia*, Vol. 3-6. 2016.

[6] J. Vivaldi, *Extracción de candidatos a término mediante la combinación de estrategias heterogéneas*, PhD Thesis, Universidad Politécnica de Cataluña, 2001.

[7] M. T Cabré, *Terminology: Theory, methods and applications*, John Benjamins, Ámsterdam, 1999.

[8] K. Kageura and B. Umino, op. cit.

[9] S. Ananiadou and G. Nenadic, "Automatic terminology management in biomedicine", in S. Ananiadou and J. McNaught (eds.), op. cit, 2006.

[10] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology", *Nucleic Acids Research, 32* (Database issue), Oxford, 2004.

[11] MeSH (Medical Subject Headings) is the National Library of Medicine controlled vocabulary thesaurus used for indexing articles for PubMed (https://www.ncbi.nlm.nih.gov/mesh).

[12] SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) is the most comprehensive and precise clinical health terminology product in the world (https://www.snomed.org/snomed-ct/).

[13] ICD10Data.com is a free reference website designed for the fast lookup of all current American ICD-10-CM (diagnosis) and ICD-10-PCS (procedure) medical billing codes (http://www.icd10data.com/).

[14] A. Ballester, Á. Martín Municio, F. Pardos, J. Porta, R. J. Ruiz and F. Sánchez, "Combining statistics on n-grams for automatic term recognition", in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. Universidad de Las Palmas de Gran Canaria, 2002.

[15] K. Kageura and B. Umino, op. cit.

[16] M. Krauthammer and G. Nenadic, 2004, op. cit.; in S. Ananiadou and G. Nenadic, op. cit.

[17] I. Segura-Bedmar, P. Martínez Fernández and D. Samy, "Detección de fármacos genéricos en textos biomédicos", *Procesamiento del Lenguaje Natural, 40*, Jaén, 2008.

[18] S. Ananiadou, "A methodology for Automatic Term Recognition", COLING'94 – *Proceedings of the 15th Int. Conf. on Computational Linguistics*, pp. 1034-1038, 1994.

[19] I. Dagan and K. Church, "TERMIGHT: Identifying and Translating Technical Terminology", in *4th Conference on Applied Natural Language Processing*, 1994; and J. S. Justeson and S. M. Katz, "Technical terminology: some linguistic properties and an algorithm for identification in text", *Natural Language Engineering, 1(1)* Cambridge, 1995.

[20] W. Koza, Z. Solana, M. DA S. Conrado, S. O. Rezende, T. A. Pardo, J. Díaz-Labrador and J. Abaitua, "Extracción terminológica en el dominio médico a partir del reconocimiento de sintagmas nominales", *INFOSUR, 5*, Rosario, Argentina, 2011.

[21] For a comparison between Greek-Latin suffixes in English and Japanese, two languages belonging to two very distinct language families, please review C. Herrero Zorita, C. Molina and A. Moreno Sandoval, "Medical term formation in English and Japanese: A study of the suffixes -gram, -graph and -graphy", *Review of Cognitive Linguistics, 13(1),* Amsterdam, 2015.

[22] A. Moreno Sandoval and J. M. Guirao, "Frecuencia y distintividad en el uso lingüístico: casos tomados de la lematización verbal de corpus de distintos registros", in *Actas del I Congreso Intl. de Lingüística de Corpus*, Universidad de Murcia, Murcia, 2009.

[23] T. Dunning, "Accurate methods for the statistics of surprise and coincidence", *Computational Linguistics, 19(1),* Cambridge, MA, 1993.

[24] A. Kilgarriff, P. Rychly, P. Smrz and D. Tugwell, "The Sketch Engine", in *Proceedings of EURALEX 2004*, Lorient, France, 2004.

[25] H. Nakagawa and T. Mori, "Automatic term recognition based on statistics of compound nouns and their components", *Terminology, 9(2*), Amsterdam, 2003.

[26] R. Nazar, J. Vivaldi and L. Wanner, "Automatic taxonomy extraction for specialized domains using distributional semantics", *Terminology, 18(1)*, Amsterdam, 2012.

[27] A. Ballester Á. Martín Municio, F. Pardos, J. Porta, R. J. Ruiz and F. Sánchez, "Combining statistics on n-grams for automatic term recognition", in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. Universidad de Las Palmas de Gran Canaria, 2002.

[28] R. Nazar and M. T. Cabré, "Un experimento de extracción de terminología utilizando algoritmos estadísticos supervisados", *Debate Terminológico, 7*, 2010.

[29] S. Hochreiter and J. Schmidhuber. "Long short-term memory". *Neural computation*, vol. 9, no 8, pp. 1735-1780, 1997.

[30] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. "Neural architectures for named entity recognition", in *Proceedings of NAACL 2016*, San Diego, CA, 2016.

[31] M. Rei. "Semi-supervised Multitask Learning for Sequence Labeling", *Proceedings of ACL 2017*, Vancouver, Canada, 2017.

[32] T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Efficient estimation of word representations in vector space". 2013. https://arxiv.org/abs/1301.3781.

[33] J. Pennington, R. Socher, and C. D. Manning. "Glove: Global vectors for word representation". in *EMNLP*. volume 14, pages 1532– 1543, 2014.

[34] A. Barrón Cedeño, G. Sierra, P. Drouin and S. Ananiadou, "An Improved Automatic Term Recognition Method for Spanish", in A. Gelbukh (ed.), *CICLing2009 LNCS 5449*. Springer, Berlín, 2009.

[35] A. Moreno Sandoval and L. Campillos Llanos, "Design and annotation of MultiMedica - a multilingual text corpus of the biomedical domain", in C. Vargas-Sierra (ed.), *Procedia, 95*, Elsevier, Berlín, 2013.

[36] A. Moreno Sandoval and L. Campillos Llanos, "Combined strategies for automatic term recognition and its application to a Spanish corpus of medicine". *Lingüística española actual*, 37(2), pp. 173-197. 2015.

[37] A. Moreno Sandoval and J. M. Guirao Miras, "Morpho-syntactic Tagging of the Spanish C-ORAL-ROM Corpus: Methodology, Tools and Evaluation", in Y. Kawaguchi, S. Zaima and T. Takagaki (eds.), *Spoken Language Corpus and Linguistic Informatics*, John Benjamins, Amsterdam, 2006.

[38] T. Dunning, "Accurate methods for the statistics of surprise and coincidence", *Computational Linguistics, 19(1),* Cambridge, MA, 1993.

[39] A. Moreno Sandoval and J. M. Guirao Miras, op. cit.

[40] G. Hripcsak, and A. S. Rothschild. "Agreement, the F-measure, and reliability in information retrieval." *Journal of the American Medical Association 12*: 296-298, 2005.

[41] C. Herrero Zorita, L. Campillos Llanos and A. Moreno Sandoval. "Collecting a POS-Tagging a Lexical Resource of Japanese Biomedical Terms from a Corpus." *Procesamiento del Lenguaje Natural 52*: 29-36. 2014.

[42] N. Habash, O. Rambow and R. Roth. "Mada+Tokan: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization." *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*. Cairo, Egypt: 242-245. 2009.

[43] A. Moreno Sandoval and L. Campillos Llanos, op. cit.

[44] Online Life Science Dictionary. Available at https://lsd-project.jp/cgi-bin/lsdproj/ejlookup04.pl [30/12/2017]

[45] D. Samy, A. Moreno Sandoval, C. Bueno-Díaz, M. Garrote-Salazar and J.M. Guirao. "Medical Term Extraction in an Arabic Medical Corpus." *Proceedings of the 8th Language Resources and Evaluation Conference,* pp. 640-645. Istanbul: LREC, 2012

[46] J. M. López Piñero and M. L. Terrada Ferrandis, *Introducción a la terminología médica*, Masson, Barcelona, 2005; and M. E. Jiménez, "Afijos grecolatinos y de otra procedencia en términos médicos", *MEDISAN*, 16(6) (Santiago de Cuba, 2012), pp. 1005-1021; M. A. Sánchez González, *Historia de la medicina y humanidades médicas*, 2 ed., Elsevier/Masson, Barcelona, 2012

[47] WHO, "The use of stems in the selection of International Nonproprietary Names (INN) for pharmaceutical substances", 2013, http://www.who.int/medicines/services/inn/StemBook_2013_Final.pdf and http://www.who.int/medicines/services/inn/Addendum_StemBook2013_201506.pdf [01/12/2017]; OMS, "International Nonproprietary Names (INN) for biological and biotechnological substances (a review)", 2016, < http://www.who.int/medicines/services/inn/BioReview2016.pdf> [01/12/2017].

[48] AMA: < https://www.ama-assn.org/sites/default/files/media-browser/public/usan/stem-list-cumulative_0.xlsx > [01/12/2017]. Michael Quinion's list of affixes was also used: < www.affixes.org > [01/12/2017]

[49] Real Academia Nacional de Medicina, *Diccionario de términos médicos*, Editorial Médica Panamericana, Madrid, 2011.

[50] Dorland, *Diccionario enciclopédico ilustrado de medicina* Dorland, Elsevier, Madrid, 30 edition, 2005; online: < https://www.dorlandsonline.com/dorland/home> [01/12/2017]

[51] L. M. Gonzalo Sanz (coord.), *Diccionario Espasa Medicina*, Espasa S.L., Madrid, 1999.

[52] F. Cortés Gabaudán (coord.), *Dicciomed*, 2007-2013. http://dicciomed.eusal.es [01/12/2017]

[53] A. Moreno SandovaL, L. Campillos Llanos, C. Herrero-Zorita, J. M. Guirao Miras, A. González Martínez, D. Samy y E. Takamori "An online tool for enhancing NLP of a biomedical corpus", *6th International Conference on Corpus Linguistics (CILC 2014)*, Las Palmas de Gran Canaria, 2014

[54] A Spanish version of the ICD-10 is accessible through the web of the Spanish Ministry of Health (http://eciemaps.msssi.gob.es/ecieMaps/browser/index_10_mc_old.html) [6/01/2018]

[55] Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT): https://www.snomed.org/snomed-ct/ [6/01/2018]

[56] BabelNet: http://babelnet.org/ [6/01/2018]

[57] Y. He, and M. Kayaalp. "A Comparison of 13 Tokenizers on MEDLINE"; *Technical Report LHNCBC-TR-2006-003*; The Lister Hill National Center for Biomedical Communications: Bethesda, MD, December 2006

[58] N. Barrett, and J. Weber-Jahnke, "Building a Biomedical Tokenizer Using the Token Lattice Design Pattern and the Adapted Viterbi Algorithm". *BMC Bioinf. 2011*, 12 (Suppl 3), S1, 2011

[59] Y. Tsuruoka,, Y. Tateishi, J. D. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii, "Developing a Robust Part-of-Speech Tagger for Biomedical Text" in *Advances in Informatics*; Bozanis, P., Houstis, E. N., Eds.; Vol. 3746 Springer Berlin Heidelberg: Berlin, Heidelberg, 2005.

[60] K. Sagae, and J. Tsujii, "Dependency Parsing and Domain Adaptation with LR Models and Parser Ensembles". *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL*; Prague, Czech Republic, June, 2007.

[61] Y. Miyao and J. Tsujii. "Feature Forest Models for Probabilistic HPSG Parsing". *Computational Linguistics*. 34 (1), 2008.

[62] M. Miwa, R. Sætre, Y. Miyaq, and J. Tsujii, "Protein-Protein Interaction Extraction by Leveraging Multiple Kernels and Parsers". *International Journal of Medical Informatics*, 78 (12), 2009.

[63] J-F. Fontaine, A. Barbosa-Silva, M. Schaefer, M.R. Huska, E. M. Muro, M.A. Andrade-Navarro, "MedlineRanker: Flexible Ranking of Biomedical Literature". *Nucleic Acids Res. 37*, 2009.

[64] I. Spasic, S. Ananiadou, J. Mcnaught, and A. Kumar, "Text Mining and Ontologies in Biomedicine: Making Sense of Raw Text", *Briefings in Bioinformatics.* 6 (3), 2005.

[65] F. Leitner, S.A. Mardis, M. Krallinger, G. Cesareni, L.A. Hirschman, A. Valencia, "An Overview of BioCreative II.5". *IEEE/ ACM Transactions in Computational Biological Bioinformatics*. 7 (3), 2010.

[66] M. vazquez, M. krallinger, F. leitner, and A. Valencia, "Text Mining for Drugs and Chemical Compounds: Methods, Tools and Applications". *Molecular Informatics*. 30 (6−7), 2011.

[67] Project BioASQ: http://bioasq.org/ [6/01/2018]
[68] W. Salloum, G. Finley, E. Edwards, M. Miller and D. Suendermann-Oeft *Deep Learning for Punctuation Restoration in Medical Reports*, 2017
[69] S. Baker and A. Korhonen, *Initializing neural networks for hierarchical multi-label text classification*, 2017
[70] C. Lin, T. Miller, D. Dligach, S. Bethard and G. Savova, *Representations of Time Expressions for Temporal Relation Extraction with Convolutional Neural Networks*, 2017
[71] M. Asada, M. Miwa and Y. Sasaki, *Extracting Drug-Drug Interactions with Attention CNNs*, 2017
[72] Y. Peng and Z. Lu, *Deep learning for extracting protein-protein interactions from biomedical literature*, 2017
[73] J. Tourille, O. Ferret, A. Neveol and X. Tannier. "Neural Architecture for Temporal Relation Extraction: A Bi-LSTM Approach for Detecting Narrative Containers.". *Proceedings of ACL 2017*, pp. 224-230. 2017.
[74] S. Mohan, N. Fiorini, S. Kim and Z. Lu, *Deep Learning for Biomedical Information Retrieval: Learning Textual Relevance from Click Logs*, 2017.
[75] B. McInnes and T. Pedersen, *Improving Correlation with Human Judgments by Integrating Semantic Similarity with Second–Order Vectors*, 2017
[76] K. Bretonnel Cohen, D. Demner-Fushman, S. Ananiadou, and J.-I. Tsujii. *Biomedical natural language processing in 2017: The view from computational linguistics* (http://aclweb.org/anthology/W17-23) [6/01/2018]

Antonio Moreno-Sandoval

Antonio Moreno-Sandoval (BArts 1986, MArts 1988, PhD 1991, Universidad Autónoma de Madrid, UAM) is Professor of Linguistics and Director of the Computational Linguistics Lab at UAM. He is a former Fulbright postdoc scholar at the Computer Science Dept., New York University (1991-1992) and a former DAAD scholar at Augsburg Universität (1998). His training in Computational Linguistics began as a research assistant in the Eurotra Machine Translation Project (EU FP-2) and then at IBM Scientific Center in Madrid (1989-1990). He was the principal researcher of the Spanish team in the C-ORAL-ROM Project (EU FP-5). He has managed over 15 projects (national, regional-funded) as well as industry contracts. Since 2010 he is Senior Researcher at the Instituto de Ingeniería del Conocimiento (IIC-UAM) in the Social Business Analytics group. Moreno-Sandoval has supervised 9 theses to completion. He is author or co-author of 4 books and over 80 scientific papers.

Julia Díaz

Julia Díaz received a MSci Degree in Mathematics, PhD in Computer Science (both from Universidad Autónoma de Madrid – UAM-Spain) and General Management Program (IESE-Universidad de Navarra – Spain). At present she is Senior Innovation Manager at the Instituto de Ingeniería del Conocimiento (IIC-UAM), a private R&D&i institution dedicated to extracting knowledge from high volumes of heterogeneous data (Big Data) and optimizing business processes in areas such as Healthcare and Energy. She also is Part Time PhD Professor in Computer Sciences in the UAM and Professor in the Big Data & Data Sciences Master in UAM and ESADE.

Leonardo Campillos-Llanos

Leonardo Campillos-Llanos (BArts 2004, Universidad Complutense de Madrid-UCM; MArts 2006, PhD 2012, Universidad Autónoma de Madrid-UAM) is a postdoctoral researcher at the Laboratoire d'Informatique pour la Mécanique et les Sciencies de l'Ingénieur (LIMSI-CNRS) at Orsay (France). He has been working in the area of BioNLP since 2011 in different projects (MULTIMEDICA, PatientGenesys). Currently, Dr. Campillos is in charge of the natural language interaction and terminology modules of a dialogue system simulating a virtual patient in a medical consultation (PVDial project).

Teófilo Redondo

Teófilo Redondo (BArts -1985, MArts - 1986; Universidad Complutense de Madrid - UCM) is Senior Consultant – ICT at Ayming España, with a special focus on Artificial Intelligence, Cognitive Robotics and all things NLP. Prior to this he was Project Portfolio Coordinator at Zed Worldwide, in the Deparment of Innovation. He was before Technology Architect & Director of Innovation Projects at Universidad Internacional de La Rioja (UNIR). Previously he developed a career at IBM covering several areas like Cloud Computing and Big Data Architectures, Enterprise Solutions Architect (SAP, Oracle Solutions, Dassault Systèmes), and as SOA Architect. He started in the IBM Research Division (IBM Scientific Center in Madrid) with several projects on Machine Translation, during which he produced a number of articles on this subject. He was Visiting Scholar at Stanford University (1987). He is affiliated with SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural) since almost the beginning.