Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Data and text mining

Cross-lingual Semantic Annotation of Biomedical Literature: Experiments in Spanish and English

Naiara Perez [1,*,†], Pablo Accuosto [2,†], Àlex Bravo [2,†], Montse Cuadros [1], Eva Martínez-Garcia [1,‡], Horacio Saggion [2], German Rigau [3]

[1] SNLT, Vicomtech, Donostia, 20009, Spain
[2] TALN/DTIC, Universitat Pompeu Fabra, Barcelona, 08018, Spain
[3] IXA Group, HiTZ Centre, University of the Basque Country UPV/EHU, Donostia, 20018, Spain

* To whom correspondence should be addressed.
† Contributed equally.
‡ Present address: CEIEC Research Institute, Francisco de Vitoria University, Madrid, 28223, Spain.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Biomedical literature is one of the most relevant sources of information for knowledge mining in the field of Bioinformatics. In spite of English being the most widely addressed language in the field, in recent years there has been a growing interest from the natural language processing community in dealing with languages other than English. However, the availability of language resources and tools for appropriate treatment of non-English texts is lacking behind. Our research is concerned with the semantic annotation of biomedical texts in the Spanish language, which can be considered an under-resourced language where biomedical text processing is concerned.

Results: We have carried out experiments to assess the effectiveness of several methods for the automatic annotation of biomedical texts in Spanish. One approach is based on the linguistic analysis of Spanish texts and their annotation using an information retrieval and concept disambiguation approach. A second method takes advantage of a Spanish-English machine translation process to annotate English documents and transfer annotations back to Spanish. A third method takes advantage of the combination of both procedures. Our evaluation shows that a combined system has competitive advantages over the two individual procedures.

Availability: UMLSmapper (https://snlt.vicomtech.org/umlsmapper) and the annotation transfer tool (http://scientmin.taln.upf.edu/anntransfer/) are freely available for research purposes as web services and/or demos.

Contact: nperez@vicomtech.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Biomedical literature is one of the most relevant sources of information for the advancement of research in the life sciences. However, the current publication rate (Ware and Mabe, 2015) makes it very difficult for researchers to keep up with the most relevant results across this broad field of knowledge. Natural language processing (NLP) tools –including automatic summarisation and semantic search tools– can contribute to partially avoid this obstacle by facilitating the finding and use of biomedical knowledge. The unambiguous identification of biomedical terminology, also known as *term normalisation*, is an essential first step in the automatic extraction of this valuable knowledge. A *term* is a textual representation (usually composed of one or more words) that describes a particular concept and *normalisation* is the process of linking it to a unique identifier (in general, an entry in a manually curated database or ontology). As a consequence of the advancement of the biomedical field, the terms and concepts included in controlled vocabularies are in continuous evolution,

making term normalisation a very challenging research area in biomedical informatics.

In this paper we are concerned with the semantic annotation (e.g. term identification and normalisation) of Spanish biomedical text sources. In the last 5 years, Medline has indexed more than 9,500 publications per year in Spanish[1]. Furthermore, several bibliographical databases collect publications about Biomedicine and Health Sciences published in Spain and other Spanish-speaking countries. For instance, IBECS collects scientific journals published in Spain since 2000, containing more than 170,000 records[2]. While these numbers are small compared to the English literature publication rate, non-English biomedical text sources also contain valuable information that must be retrieved and processed to make it accessible to the international scientific community.

In this light, we have developed two different methods: a first approach is a native, lexical-motivated procedure which takes advantage of lexical resources and an information retrieval mechanism to identify terms based on a linguistic analysis and a disambiguation procedure; a second approach is cross-lingual and takes advantage of a machine translation procedure to automatically translate the Spanish input into English so that the text can be annotated with English tools and the annotations transferred back to Spanish. Finally a third system combines the annotations produced by both approaches. As it will be shown in this paper, the combined system has competitive advantages over both individual approaches. The paper provides a detailed description and a thorough evaluation – using appropriate evaluation metrics and competitive baselines and upper bounds– of the three approaches and components (including the machine translation approaches). It also includes a careful error analysis of the results and discusses avenues for further improvement.

The rest of the paper is structured as follows: in the next section we present related work; Section 3 describes the three approaches to semantic annotation and the evaluation framework; the results of the evaluation are reported in Section 4 and discussed in Section 5. Finally, we present the conclusions reached and the avenues for improvement in Section 6.

2 Related Work

The automatic identification of biomedical terminology in scientific texts is an active research area but most of the recent works are targeted at the English language. This is due, in part, to the greater availability of biomedical resources –such as scientific articles, vocabularies and ontologies– in English. In this scenario, MetaMap (Aronson, 2001, 2006), cTakes (Savova *et al.*, 2010) and NCBO Annotator (Dai *et al.*, 2008) are well-known tools for the semantic annotation of biomedical text. Metamap is probably the better-known tool. It is "knowledge intensive" as it relies heavily on the SPECIALIST Lexicon, a large syntactic lexicon of biomedical and general English. cTakes recognises biomedical concepts in texts and relates them to their UMLS concept. And the NCBO Annotator, developed by the National Center for Biomedical Ontology (NCBO), is a web service that provides links between the text of biomedical literature and the knowledge embedded in the BioPortal ontologies and the UMLS Metathesaurus.

In the last years, new works have emerged to face this challenging task, allowing the advance of the state-of-the-art. Nunes *et al.* (2013) developed BeCAS, a biomedical concept annotation system, which uses dictionary-matching techniques to recognise diverse types of concepts (including species, anatomical concepts, microRNAs, enzymes, chemicals, drugs, diseases, metabolic pathways, cellular components, biological processes and molecular functions) from multiple sources, including UMLS, NCBI BioSystems (Geer *et al.*, 2009), LexEBI (Rebholz-Schuhmann *et al.*, 2013b), ChEBI (Hastings *et al.*, 2016), miRBase (GriffithsâŁŁJones, 2004) and the Gene Ontology (Consortium, 2004). It provides an API and a web-based tool for biomedical concept identification.

NOBLE Coder (Tseytlin *et al.*, 2016) is another open-source system for biomedical text annotation in English. It can be configured through a graphical interface to work with different vocabularies, even with customised terminologies, allowing to select one or more branches of a set of vocabularies and/or filtering vocabularies by semantic types.

Recently, Soysal *et al.* (2017) implemented CLAMP, a pipeline composed of multiple modules for the analysis and the extraction of information contained in clinical text. It includes a named entity recogniser to detect biomedical terminology. Then, an UMLS encoder links each term with the corresponding concept in the UMLS Metathesaurus.

In the case of non-English biomedical text, term normalisation becomes even more difficult mainly by a shortage of biomedical resources. In this scenario, we present the most relevant works for term normalisation in Spanish. Carrero *et al.* (2008) presented one of the first works in using a combination of automatic translation and an English NER (MetaMap) in order to annotate biomedical entities in Spanish texts with their corresponding UMLS concepts.

Later, Castro *et al.* (2010) developed an automatic system for the recognition of SNOMED CT concepts by computing a similarity function between sentences in clinical notes and then term normalisation is based on the results obtained by querying an Apache Lucene[3] index of SNOMED CT and re-ranking the candidates with a function of their own. They obtained an average F1 score of 0.11 on their own corpus of 100 manually annotated documents. Furthermore, Berlanga *et al.* (2010) introduced the notion of *concept retrieval*, which was based on applying information retrieval methods in order to obtain UMLS concepts relevant to a text and later use them to properly annotate matching text spans.

The systems developed in the context of the 2013 CLEF-ER challenge for biomedical entity recognition in parallel multilingual corpora (Rebholz-Schuhmann *et al.*, 2013a) provide some of the first prototypes for the annotation of biomedical texts in languages other than English. Among the participating systems there were some targeted at Spanish including the ones proposed by Attardi *et al.* (2013) and Bodnari *et al.* (2013), which exploited word alignment information by statistical translation and parallel corpus, respectively, in order to transfer annotations from English to Spanish. Specifically, Attardi *et al.* (2013) translated a English corpus with biomedical entity annotations to Spanish, including the transfer of annotations. Then, a Named Entity Recognition (NER) was trained in the translated Spanish corpus in order to recognise biomedical entities in unseen Spanish text. Otherwise, Bodnari *et al.* (2013) manually annotated biomedical entities in English text from a parallel corpus and were transferred to Spanish (and French) text in order to train a NER for each language. These works were not evaluated against a golden corpus.

In the same year, Oronoz *et al.* (2013) presented FreelingMed, an extension of the Freeling Spanish analyser[4] to recognise biomedical entities extracted from available knowledge resources (lists of medical abbreviations and drug names, as well as the SNOMED CT thesaurus). Oronoz *et al.* (2013) evaluated their proposal with their own corpus of medical reports annotated by health professionals with diseases, medications and other substances, obtaining 0.90 F1 score with approximate boundary matching for the term recognition task.

Most recently, Roller *et al.* (2018) presented a sequential cross-lingual candidate search for biomedical term normalisation. The main component

[1] https://analytics.scielo.org/w/publication/article?la_scope=en&la_scope=es

[2] http://bvsalud.isciii.es/productos-y-servicios/

[3] https://lucene.apache.org/

[4] http://nlp.lsi.upc.edu/freeling/

of their approach is a character-based neural translation model trained on UMLS for multiple languages, such as Spanish, French, Dutch and German. To the best of our knowledge, this is the only work evaluated on a public gold standard corpus for semantic annotation: the Mantra Gold Standard Corpus (GSC) (Kors *et al.*, 2015). Roller *et al.* (2018) achieve 0.69 F1 score on the task of normalisation of golden terms in the Spanish Medline sub-corpus.

To this day, biomedical semantic annotation in non-English text is still one of the most challenging research topics in biomedical NLP. In this work, we contribute two novel approaches for biomedical term normalisation in non-English texts and thoroughly compare their performance in an existing parallel gold standard corpus.

3 System and methods

In this paper we compare two approaches to identify biomedical terminology from biomedical text in Spanish and English. In both approaches we use the UMLS Metathesaurus for the term normalisation step. The UMLS Metathesaurus is by far the largest thesaurus in the biomedical domain, containing medical terms from many sources and in several languages. It arranges terms by meaning, assigning the same Concept Unique Identifier (CUI) to the terms that denote the same concept (e.g. *'High blood pressure'*, *'Systemic arterial hypertension'* and *'Hypertensive vascular disease'* are terms related to the concept C0020538). Furthermore, each concept is categorised into one or more of the 133 *semantic types* of the UMLS Semantic Network (McCray and Nelson, 1995). For instance, the previous concept C0020538 belongs to semantic type 'Disease or Syndrome'. The UMLS also presents unbalanced data between languages being English the major language of the UMLS Metathesaurus with 7,460,514 distinct terms of 3,250,158 concepts[5]. The second largest subset by language of the Metathesaurus is Spanish, which contains 1,255,376 distinct terms related to 451,296 concepts.

The rest of this section describes in detail the approaches proposed to biomedical terminology identification with the UMLS, as well as the evaluation framework.

3.1 Baseline: MetaMap adapted to Spanish

MetaMap was selected as baseline because it is one of the most used tools for the mapping of biomedical terminology to UMLS concepts. MetaMap is a highly configurable program and by means of MetaMap Data File Builder[6], we have created a knowledge database of the Spanish UMLS subset for MetaMap 2016v2[7]. In order to minimise any disadvantages stemming from the knowledge base, we will use the same terms and concepts as those indexed for UMLSmapper (see Section 3.4.2). However, MetaMap relies heavily on lexical resources and morphological analyses of English text, we therefore expect worse results on languages other than English. Furthermore, it must be noted that MetaMap can only read ASCII encoded files; thus, both the terms indexed and the test input texts had to be converted to ASCII. This was done with the Linux command `iconv -f utf-8 -t ascii//TRANSLIT`, which replaces non-ASCII characters with their transliterations (e.g., it converts "publicaciones científicas en español" to "publicaciones cientificas en espanol").

[5] Version UMLS 2016AA Metathesaurus.

[6] `https://metamap.nlm.nih.gov/DataFileBuilder.shtml`

[7] `https://metamap.nlm.nih.gov/MetaMap.shtml`

3.2 Transfer pipeline: Transferring English entities to Spanish

This pipeline exploits an automatic annotation tool of UMLS concepts for English, an automatic English to Spanish machine translation tool and an annotation transferring process, for the annotation of Spanish biomedical texts. This pipeline therefore consists of the following steps: *i)* translation of Spanish texts into English, *ii)* automatic annotation of the English text, and *iii)* transfer of the obtained annotations to the original Spanish texts.

3.2.1 Machine Translation

Two automatic translation systems have been tested: Google Translate[8] and a Neural Machine Translation (NMT) model adapted to the biomedical domain. The NMT model trained is a Recurrent Neural Network (RNN) Encoder-Decoder (Cho *et al.*, 2014) with attention (Bahdanau *et al.*, 2015).

In detail, the encoder consists of a 4-layered bi-directional RNN. The decoder is a forward RNN, with 4 layers as well. Each hidden layer of both the encoder and decoder RNNs has 800 LSTM units. The embeddings are of dimension 500, both for the source –Spanish– and target –English– languages. In order to control the system vocabulary, Byte Pair Encoding (BPE) segmentation (Sennrich *et al.*, 2016) has been applied to the data. The BPE model has been trained jointly on both English and Spanish data to extract 30,000 merge operations resulting in dictionaries of 23,272 and 27,707 subwords for English and Spanish, respectively.

The training dataset is composed of two parallel corpora from UFAL[9] – EMEA and Medical Web Crawl (MWC)–, and a parallel corpus of Medline titles released for the WMT2016 translation shared task[10]. Additionally, a parallel corpus has been generated from SNOMED-CT by extracting all the possible bilingual term-pairs for each CUI in order to introduce the terminology in the training process. This corpus is also added to the training dataset.

3.2.2 Transfer annotations back to Spanish texts

Previously, MetaMap 2016v2[11] was used to annotate the English texts with UMLS concepts. This version of MetaMap also used the same terms and concepts as those indexed for UMLSmapper (see Section 3.4.2).

Once the relevant UMLS concepts are identified in English by means of the MetaMap tool, the spans in the corresponding Spanish texts that best match each of them have to be determined in order to transfer the annotations. The transfer system is described in more detail in (Accuosto and Saggion, 2018). We summarise it here and describe the modifications introduced to the original system. We assume that the instances of the same concepts appear in the same order in the English and Spanish texts. Therefore, we process the Spanish text sequentially to find, for each identified concept instance, the text span in Spanish that best matches it.[12] In order to do this, we compute the similarity between each considered text span and all the lexicalisations of the concept available in UMLS.

3.2.3 Candidate terms generation

In order to identify spans of text in Spanish as candidates for being annotated as biomedical terms, we first split the text into sentences,

[8] `https://translate.google.com/`; June 2018.

[9] `https://ufal.mff.cuni.cz/ufal_medical_corpus`

[10] `http://www.statmt.org/wmt16`

[11] `https://metamap.nlm.nih.gov/MainDownload.shtml`

[12] The transferring process does not produce overlapping or nested annotations.

tokenise it and perform part-of-speech (POS) tagging[13] by means of the Stanford CoreNLP library (Manning *et al.*, 2014)[14]. We hypothesise that there is a correspondence between the syntax of the terms to be annotated and that of the strings that describe UMLS concepts. Based on the manual examination of 2,000 strings randomly selected from the Spanish UMLS, we observe that 1,929 (96.45%) begin with a token with NOUN or PROPN as POS tag and that 1,739 (86.95%) have a length less than or equal to eight tokens. We therefore define a heuristic rule for the generation of term candidates in which we consider sequences of up to eight tokens beginning with a token POS-tagged as NOUN or PROPN. The adequacy of our hypothesis is verified for the case of the Mantra corpus, used in the evaluation, where 97.22% of the annotated terms begin with a NOUN or PROPN and 99.74% have a length of up to eight tokens.

3.2.4 Similarity computation

The similarity score between candidate terms and UMLS concept to be transferred is computed as the maximum cosine similarity between dense vector representations of the candidate term and all the Spanish lexicalisations of the UMLS concept. The annotation is produced for the candidate term that gets the highest score, as long as it is greater than a pre-established threshold (see Section 3.4.2). In the implementation used in these experiments we also consider the length of the candidate terms when very similar scores[15] are obtained, preferring longer terms.

3.2.5 Word embeddings

As dense vector representation of the Spanish lexicalisation of UMLS concepts and candidate terms we use 300-dimensional fastText vectors (Bojanowski *et al.*, 2017)[16] tailored specifically for the particular task at hand: they were trained with texts in Spanish from the SciELO corpus (Neves *et al.*, 2016) (407.5 million words) and the Spanish lexicalisation of the UMLS concepts available in the Metathesaurus (5.7 million words), obtaining a final vocabulary of 603,195 tokens[17]. Both for the candidate terms and the UMLS concepts, their corresponding dense vectors were computed as the average of the normalised embeddings of the words included in them.

Since word representations in fastText are computed as the sum of their character n-gram vectors, embeddings for out-of-vocabulary words can be generated on the fly. This is particularly useful when dealing with biomedical terms, as subword vectors can capture relevant meaning conveyed by word roots, prefixes and suffixes.

3.3 UMLSmapper: Biomedical term normalisation in Spanish

A detailed description of the initial version of this pipeline is given in (Perez *et al.*, 2018). Here, we provide an overview of the pipeline and describe several minor changes. The system proposed consists of three main components: *i)* a NLP pipeline, *ii)* an information-retrieval component, and *iii)* a word-sense disambiguation component. Briefly, UMLSmapper proceeds as follows: first, the NLP pipeline calculates phrases from the input text; second, the information retrieval system, which contains an index of the knowledge base to be projected in the input texts –in our case, a subset of UMLS–, generates candidate phrase-to-concept mappings by lexical matching; the candidates are ranked based on heuristics and a threshold is applied to discard those with too low scores; finally, the candidate with highest score is chosen as final concept for a phrase; ambiguities (i.e., phrases with more than one top-ranked candidate mapping) are resolved by means of a word-sense disambiguation library.

3.3.1 The NLP pipeline

The pipeline starts with an optional step: abbreviation and acronym recognition and resolution. The recognition module consists of a simple Random Forest classifier, described in (Cuadros *et al.*, 2018). It has been learned from the training and development sets provided at the 2nd Edition of the Biomedical Abbreviation Recognition and Resolution Workshop (Intxaurrondo *et al.*, 2018). Resolution is performed by dictionary look-up. Then, the input text is segmented and tokenised with IXA-pipes (Agerri *et al.*, 2014). Finally, phrases are extracted by calculating n-grams up to size 5 that do not start or end with a stopword.

3.3.2 The concept index

As a pre-processing step, an Apache Lucene™ index is calculated for each of the concepts in the relevant UMLS Metathesaurus subset (see the Section 3.4.2 for more details). Each entry in the index associates a term to its concept, its vocabulary source, and its semantic type. A normalised version of the original term is also indexed. Normalisation consists in undoing transpositions, erasing spurious parenthetical material, and erasing stopwords. Thus, the term "en blanco, cara que mira fijo durante sonambulismo (hallazgo)" (*Blank, staring face whilst sleep walking (finding)*) would become "cara mira fijo sonambulismo blanco" (*staring face sleep walking blank*). At run time, the index is queried with the normalised versions of the phrases extracted from the input text and returns entries with terms similar to those phrases. A small blacklist of terms and concepts avoids generating frequent erroneous candidate mappings. The blacklist consists of the terms "ii" and "hace" (an ambiguous word that can be translated as the adverb *ago* or the verbal form *does*), and the concepts C0032863 (according to MSH, the exertion of a strong influence or control over others) and C0557651 (a room prepared for studying).

3.3.3 Ranking and thresholding

New scores are assigned to the candidates by applying the function by Castro *et al.* (2010):

$$score = \frac{\gamma^2}{\text{length}(Q) \cdot \text{length}(R)} \quad (1)$$

where γ is the number of matched tokens between the query (Q) and the retrieved term (R). The threshold is set at 0.7.

3.3.4 Word-sense disambiguation

Remaining ambiguous terms are further processed with UKB (Agirre and Soroa, 2009), a library that performs knowledge-based word-sense disambiguation. The algorithm behind UKB is Personalized PageRank (Haveliwala, 2002). We apply UKB to the graph created from the same concepts indexed with Apache Lucene™ (see the Section 3.4.2) and all their relations, giving the same weight to all of them). At run time, the graph is initialised with the tokens in the input text and, when a disambiguation is required, the system just selects the candidate concept with highest activation in the resulting Personalized PageRank Vector.

[13] Using a slightly modified version of the universal POS tags: `http://universaldependencies.org/`

[14] `https://stanfordnlp.github.io/CoreNLP/`

[15] Differences in cosine similarities of less than 0.015.

[16] The embeddings were generated with fastText's skipgram mode with negative sampling and a context window of 3 words. The vectors are made available to download at: `http://scientmin.taln.upf.edu/anntransfer/scielo_umls_embeddings.tgz`

[17] Including UMLS lexicalisations in the process of training the embeddings contributes to obtain representations for the Mantra terms closer to the representations of the UMLS terms in Spanish.

3.4 Evaluation

In what follows, we present the corpora used to evaluate the proposed systems –the Mantra GSC–, as well as the evaluation setup, system configurations, and the metrics applied.

3.4.1 The Mantra GSC

The Mantra Gold Standard Corpus (GSC) (Kors *et al.*, 2015) is a collection of several parallel corpora in the biomedical domain annotated with UMLS concepts to test concept recognition systems. It contains documents in English, German, French, Spanish, and Dutch.

The annotations are limited to following subset of UMLS: concepts that *a)* belong to the terminologies Medical Subject Headings (MeSH), SNOMED-CT, and/or the Medical Dictionary for Regulatory Activities (MedDRA); and that *b)* belong to one or more of these semantic groups: Anatomy, Chemicals and drugs, Devices, Disorders, Geographic areas, Living beings, Objects, Phenomena, Physiology, and Procedures.

It is relevant to note that annotators were allowed to assign more than one concept identifier to the same text span, that is, to make *ambiguous* annotations, in cases where they could not discern the intended meaning of the suggested concepts. Furthermore, they were allowed to make *discontinuous* annotations, that is, annotations with disjoint textual spans.

Henceforth, we focus on the English/Spanish dataset, which is the one used in our evaluation. It consists of two sub-corpora of different genres: titles of abstracts from Medline, and drug labels from the European Medicines Agency (EMEA). Table 1 shows the size of this dataset. The Spanish documents contain a total number of 101 ambiguous annotations ($\sim$16% of the all the annotations) and 17 discontinuous annotations.

Table 1. Size of the Mantra GSC Spanish (es) and English (en) datasets

	Medline		**EMEA**	
	es	**en**	**es**	**en**
documents	100	100	100	100
words	1,087	989	1,984	1,738
annotations	278	285	361	363
unique concepts	285	288	295	301

3.4.2 Experimental design

The first set of experiments (see results in Section 4.1) aims at evaluating the performance of the two approaches presented for term identification in Spanish biomedical text: transfer and UMLSmapper. Two variants of transfer are tested –one uses Google Translate to translate the input texts from Spanish to English and is henceforth referred to as transfer$_G$; the second variant, transfer$_{NMT}$, does it with a domain-specific NMT component. Furthermore, three combinations of the transfer$_{NMT}$ and UMLSmapper are also evaluated, which differ in the way that overlapping predictions are handled:

- combi$_{union}$: Annotates the union of spans with the union of the CUIs.
- combi$_{transf}$: Takes as valid the prediction made by transfer$_{NMT}$.
- combi$_{UMLSm}$: Takes as valid the prediction made by UMLSmapper.

Table 2 illustrates these combinations with a real example.

A baseline is set for these experiments by the adaptation of MetaMap to Spanish, making a total of 7 approaches evaluated. Their performance is measured against the Mantra GSC in terms of precision, recall, and F1-score for exact text spans (same boundaries required in the gold standard and predictions) as well as for overlapping spans. In order to assess the loss of accuracy when non-exact matching spans are considered, an "overlapping percentage" (OP) has been calculated. Section 3.4.3 provides an explanation of all these metrics.

In this scenario, we have considered some specific configurations related to the approaches described:

- Regarding the dataset used to train the NMT model, bilingual duplicates and all the sentences that overlap with the Mantra GSC have been removed. We have selected 2,000 sentence pairs as development set during training to optimise the model against it. After applying BPE segmentation, we have also filtered out sentences with more than 50 tokens to avoid hindering the NMT system's performance, eventually obtaining 1,875,961 parallel sentences to train the model.
- The default configuration of the MetaMap used in the Transfer approach was kept, except the following functionalities: *i)* MetaMap is forced to perform disambiguation, that is, to produce one single CUI per annotation; *ii)* MetaMap is constrained to use only the terminologies included in the Mantra GSC, i.e., SNOMED-CT, MedDRA, and MeSH; and finally *iii)* it is also constrained to use only concepts of the semantic groups allowed in the Mantra GSC. After preliminary experiments, the threshold in the similarity computation step was established as 0.825.
- Regarding the UMLSmapper approach, the relevant UMLS concepts used to compute the concept index and UKB graph comprise a total of 675,175 concepts –all the Spanish terms in the Mantra GSC terminology, plus all the English terms belonging to the Chemicals and drugs semantic group in the Mantra GSC terminology– and 4,669,477 relations.

A second series of experiments provide a detail evaluation of the three steps involved in the transfer pipeline. First, we start by assessing the

Table 2. Individual and combined pipeline predictions on EMEA d320.u172

Gold annotations	
	Con la **inmunoglobulina**$_A$ humana normal pueden producirse **reacciones adversas**$_D$ como **escalofríos**$_E$, **cefalea**$_F$ [...]
Individual pipepeline predictions	
transfer$_{NMT}$	Con la **inmunoglobulina humana normal**$_B$ pueden producirse **reacciones adversas**$_D$ como **escalofríos**$_E$, cefalea [...]
UMLSmapper	Con la **inmunoglobulina humana**$_C$ normal pueden producirse **reacciones adversas**$_D$ como **escalofríos**$_E$, **cefalea**$_F$ [...]
Combined pipeline predictions	
combi$_{union}$	Con la **inmunoglobulina humana normal**$_{(B,C)}$ pueden producirse **reacciones adversas**$_D$ como **escalofríos**$_E$, **cefalea**$_F$ [...]
combi$_{transf}$	Con la **inmunoglobulina humana normal**$_B$ pueden producirse **reacciones adversas**$_D$ como **escalofríos**$_E$, **cefalea**$_F$ [...]
combi$_{UMLSm}$	Con la **inmunoglobulina humana**$_C$ normal pueden producirse **reacciones adversas**$_D$ como **escalofríos**$_E$, **cefalea**$_F$ [...]

Note: boldface spans illustrate annotated terms; subscript letters represent the concepts assigned to each term.
Translation: *'With normal human immunoglobulin, adverse reactions such as chills, headache [...] may occur.'*

two translation systems on the Mantra GSC English/Spanish documents. Results are presented in terms of Bilingual Evaluation Study (BLEU) and Brevity Penalty (BP) (Papineni *et al.*, 2002). Then, an upper bound for MetaMap is estimated by evaluating it on the English Mantra GSC in terms of precision, recall, and F1-score. Additionally, we also analyse the performance of MetaMap on the translated documents. In this case, because the texts in which the annotations have been made are different from the English Mantra GSC texts, we can no longer measure precision, recall, and F1 score. Instead, we report the Jaccard coefficient (Jaccard, 1912) and Cohen's kappa coefficient (Cohen, 1960) of the predicted and gold CUIs. Finally, we evaluate the transfer of annotations. An upper bound for this step is set by transferring the English Mantra GSC annotations to the Spanish Mantra GSC texts and measuring precision, recall, and F1-score.

3.4.3 Metrics

Precision, recall and F1 score. The cornerstone of these metrics is the notion of *true positive* (TP) predictions. In the context of this work, a TP prediction meets two criteria: *a)* it matches in text span with a gold annotation; and, *b)* it has the same CUI as the gold annotation it matches with. This applies to discontinuous gold annotations as well, even if none of the systems assessed, except MetaMap, is able to produce discontinuous predictions. In the less restrictive scenario, a TP prediction is not required to match in exact boundaries with a gold annotation, but it must just overlap with one. As for ambiguous gold annotations, a prediction is only required to guess one of the gold CUIs in order to be counted as a TP, on account of the suggested gold CUIs being interchangeable rather than complementary, as explained in Section 3.4.1. Given this definition of TP, precision (P) is the ratio of TPs produced by a system to its total number of predictions; recall (R) is the ratio of TPs produced by a system to the total number of gold annotations; and, F1-score (F1) is the harmonic mean of P and R. We report micro-average P, R and F1.

Overlapping percentage. The overlapping percentage, OP, of two annotations is calculated as the relation between the length of the overlapping span and the length of the longest annotation. We report macro-average OP.

Jaccard coefficient. This metric (J) measures the overlap between gold annotations and predictions. Specifically, it is the ratio of the intersection to the union of the set of gold CUIs and the set of predicted CUIs. In order to account for concept mention repetitions, we add a counter to each CUI, so that the size of each set is equal to the number of annotations from which it is calculated. We report the micro-average ratio.

Cohen's kappa coefficient, κ. κ measures agreement for nominal items between two systems, "*after* chance agreement is removed from consideration" (Cohen, 1960, p. 40). We use this metric as an additional indicator of how similar the annotations made by two systems are, but without taking repetitions into account. In our experiments, two systems only agree when both say that a given concept is present in the input document, (regardless of the frequency with which it is mentioned) or both say that it is not. κ ranges between -1 and 1, where negative values indicate agreement is worse than random and 1 indicates perfect agreement. We report micro-average κ.

BLEU and Brevity Penalty. BiLingual Evaluation Understudy (BLEU) (Papineni *et al.*, 2002) measures the quality of MT output. It computes a weighted average of the number of n-grams that overlap in an MT system's output with reference human translations. BLEU ranges between 0 and 1, 0 indicating a perfect mismatch, and 1 a perfect translation. The Brevity Penalty (BP) is 1 if the translations are longer or equal to the original sentence, and gets closer to 0 as the translations become shorter.

4 Results

4.1 First set of Experiments: System comparison

Table 3 shows the results of the evaluated systems on the Spanish Mantra GSC (more detailed results are given in the supplementary file supp-all-results.ods). Regarding Medline and considering non-combination systems, all systems improve the baseline by > 0.090 F1 score. The pipelines based in transfer are remarkably precise (0.720 and 0.767 on exact span match and span overlap, respectively) compared to UMLSmapper and the baseline, but they do not improve the baseline's recall at all. The difference between transfer$_G$ and transfer$_{NMT}$ seems to be practically negligible. Overall, UMLSmapper achieves the best F1 score (0.630 and 0.634). It exceeds the other systems in terms of recall particularly, while lifting precision as well with respect to the baseline.

As for EMEA, the same pattern as in Medline can be observed, although a clear difference between transfer$_G$ and transfer$_{NMT}$ arises: Google translations yield a better recall, while NMT translations allow for more precise annotations. Another evident difference with the previous table is that the best F1 score when span overlaps are allowed is achieved by transfer$_{NMT}$. This is due to the outstandingly high precision, which outdoes the better recall obtained by UMLSmapper.

On the other hand, combining the pipelines yields slightly better results than using them in isolation (the improvement is more pronounced in the case of EMEA). Specifically, recall does raise with respect to UMLSmapper –the best evaluated system in this regard–, but precision is almost always worse. Among the three combinations, combi$_{transf}$ seems to work best. While the difference is small in Medline, in EMEA this combination achieves 3 percentage points more than combi$_{UMLSm}$.

Table 3. Evaluation of term identification on the Spanish Mantra GSC.

	Medline				EMEA			
System	**P**	**R**	**F1**	**OP**	**P**	**R**	**F1**	**OP**
Exact span match								
baseline	0.472	0.486	0.479		0.401	0.449	0.424	
transfer$_G$	0.719	0.496	0.587		0.703	0.524	0.600	
transfer$_{NMT}$	**0.720**	0.489	0.582		**0.730**	0.501	0.594	
UMLSmapper	0.645	0.615	0.630		0.615	0.632	0.623	
combi$_{union}$	0.598	0.678	0.636		0.584	0.701	0.637	
combi$_{transf}$	0.600	**0.680**	**0.637**		0.598	**0.717**	**0.652**	
combi$_{UMLSm}$	0.597	0.676	0.634		0.570	0.684	0.622	
Span overlap								
baseline	0.486	0.500	0.493	98.52	0.418	0.468	0.442	**98.06**
transfer$_G$	0.755	0.522	0.617	98.09	0.770	0.573	0.657	96.64
transfer$_{NMT}$	**0.767**	0.522	0.621	97.66	**0.810**	0.557	0.660	96.52
UMLSmapper	0.649	0.619	0.634	**99.21**	0.636	0.654	0.645	97.98
combi$_{union}$	0.629	0.712	0.668	97.79	0.640	0.767	0.698	96.16
combi$_{transf}$	0.632	**0.716**	**0.671**	97.75	0.654	**0.784**	**0.713**	96.31
combi$_{UMLSm}$	0.625	0.709	0.664	97.91	0.626	0.751	0.683	96.10

4.2 Second set of Experiments: Transfer Pipeline

In this section, we evaluate the intermediate steps of the transfer pipelines: *a)* translations, *b)* MetaMap, and *c)* the transfer proper.

4.2.1 Translation

Table 4a shows the evaluation on the Mantra GSC of the two MT systems used in the transfer pipelines: Google Translate and the adapted NMT system described in Section 3.2. The NMT model outperforms Google

both in Medline and EMEA in terms of BLEU. At the same time, the NMT model yields translations shorter than expected.

4.2.2 MetaMap

Table 4b shows MetaMap's (MM) upper-bound in the transfer pipelines, that is, the results it could obtain given perfect translations. This is measured by processing the English Mantra GSC texts with MetaMap and evaluating the predictions against the English Mantra GSC annotations.

Results show that MetaMap is a major limitation of the pipelines. The scores obtained are actually comparable to those obtained by UMLSmapper on the Spanish text, although a critical step remains, namely transferring the annotations, that inevitably increases the eventual amount of errors. These results in combination with Table 3 suggests that the transferring step drops more FPs than TPs made by MetaMap, thus increasing precision notably at the expense of recall.

Next, we evaluate MetaMap's annotations on the translations. Table 4c shows that, per the Jaccard coefficient, the bag of annotations produced for Medline by the pipeline with Google is more similar to the gold annotations than that produced by the pipeline with NMT. In terms of Cohen's kappa coefficient, Google translations seem to be more helpful as well; however, the differences with respect to the adapted NMT are not statistically significant in this case. These results are surprising in light of the evaluation of the translation systems (Table 4a), where the NMT translator obtains better BLEU scores than Google. We look into this phenomenon in the error analysis.

4.2.3 Annotation transfer

Table 4d shows the upper-bound of the transferring step, that is, the results it could obtain given perfect annotations on perfect translations. This is measured by transferring the English Mantra GSC annotations to the Spanish Mantra GSC texts and evaluating the result against the Spanish Mantra GSC annotations. The results show that the transfer is done with great precision (notably better in Medline than in EMEA), although around a third of the annotations are lost in the process.

Table 4. Evaluation of the transfer pipeline's intermediate steps

(a) Evaluation of the translation systems on the Mantra GSC (Spanish to English)

	Medline		EMEA	
System	**BLEU**	**BP**	**BLEU**	**BP**
Google Translate	0.466	**0.999**	0.459	**1.000**
Adapted NMT	**0.516**	0.979	**0.541**	0.970

(b) Evaluation of term identification by MetaMap on the English Mantra GSC.

	Medline				EMEA			
Span	**P**	**R**	**F1**	**OP**	**P**	**R**	**F1**	**OP**
exact	0.628	0.628	0.628		0.600	0.653	0.625	
overlap	0.663	0.663	0.663	98.33	0.613	0.667	0.639	98.67

(c) Similarity between the set of predicted CUIs and the set of gold standard CUIs.

	Medline		EMEA	
System	**J**	**κ**	**J**	**κ**
Original English texts + MM	0.452	0.637	0.428	0.637
Google Translate + MM	**0.415**	0.607	**0.411**	0.609
Adapted NMT + MM	0.393	0.585	**0.411**	0.610

(d) Evaluation of transfer of gold annotations from English to Spanish Mantra GSC.

	Medline				EMEA			
Span	**P**	**R**	**F1**	**OP**	**P**	**R**	**F1**	**OP**
exact	0.915	0.662	0.768		0.823	0.579	0.680	
overlap	0.985	0.712	0.827	96.67	0.972	0.684	0.803	94.61

4.3 Error Analysis

In order to carry out an error analysis we have manually analysed predictions by UMLSmapper, transfer$_G$, and transfer$_{NMT}$ on 50 EMEA documents and 50 Medline documents (i.e., half the dataset).

4.3.1 Transfer pipeline error analysis

The origin of the errors is almost identical in both transfer pipelines. Around 78% of the false positives stem from errors made by MetaMap, while the remaining 21% are errors made at the transfer step. Regarding false negatives, the distribution is approximately 67%-33%, respectively, transfer$_{NMT}$ having made a few false positives due to translation errors as well. The main difference between the two pipelines seems to be that the NMT translator makes more alterations in word order than Google Translate, particularly of nouns, adjectives, and coordinated items, thus frequently producing more natural-sounding English sentences; as a consequence, either *i)* MetaMap does not predict anything, *ii)* MetaMap predicts annotations different to the gold standard, or *iii)* MetaMap predicts the annotations properly, but they cannot be transferred to the Spanish text due to the word order difference. This partly explains why, despite the NMT model achieving better BLEU scores than Google Translate (see Table 4a), the overall results of the pipelines are almost identical. On the other hand, the translation errors made by the NMT component also have a higher impact in the pipeline overall as compared to Google Translate, specially in producing false positives, as has been mentioned before.

4.3.2 UMLSmapper error analysis

Regarding false positives, 46% are a consequence of having missed a multi-word gold annotation, and having predicted shorter spans contained in the gold span. Another 41% stems from UMLSmapper's completely relying on pure lexical match with the knowledge base, while the knowledge base does not capture all the meanings of the terms it contains. Thus, UMLSmapper sometimes annotates concepts that are not denoted in the texts. While we acknowledge that we are not domain experts, the remaining false positives seem to be correct predictions. Upon close inspection, the reason seems to be the parallel nature of the Mantra GSC annotations. A compromise had to be reached to make the annotations parallel, in spite of the annotated phrases sometimes varying in syntax and wording. As a consequence, annotations might seem to be missing in some languages because the corresponding choice of words in the other languages does not allow for a sound annotation (e.g., when they are expressed as noun phrases in one language but as propositions in the others).

As for the false negatives, the causes are more varied. The vast majority (46%) mainly occur because the Metathesaurus does not capture all the existing lexical variability for each concept, and UMLSmapper does not treat this problem other than with lemmatisation and the expansion of abbreviated forms. 11% of the false negatives are due to having made multi-word predictions that span over gold annotations. Another 11% of the negative errors occurs because the gold annotations identify hyponyms of the actual words annotated. Less frequent false negatives stem from *a)* UMLSmapper's configuration (e.g., maximum prediction length, not allowing for discontinuous predictions); *b)* a faulty tokenisation

and/or lemmatisation; and *c)* incorrect sense disambiguation of correctly recognised terms. Finally, 9% false negatives are seemingly correct alternative CUIs, although we acknowledge once again that we are not healthcare experts.

5 Discussion

The presented experiments seem to indicate that while UMLSmapper achieves higher recall, the transfer approach is better at precision. At the same time, a considerable increase in recall is observed when both approaches are combined at the expense of precision, producing a more competitive system in terms of F1, and this in spite of the limitations (e.g. translation noise) imposed by the translation to English.

Where the translations are of concern, the shorter translations produced by the NMT approach could explain in part why, despite producing better translation in terms of BLEU, transfer$_{NMT}$ is more precise but obtains less recall than transfer$_G$. However, having manually compared the errors made by the two system variations in half of the dataset, the difference in performance seems to be better explained by the fact that the NMT translator makes more alterations in word order than Google Translate, which affects both the performance of MetaMap and of the transfer algorithm, ultimately yielding a better precision and worse recall than transfer$_G$. Considering combi$_{transf}$ being the best pipeline combination, we argue that these results make sense as the transfer pipeline identifies more correct CUIs while UMLSmapper provides more annotations.

Achieving 100% correct annotation in Spanish by transferring annotations from English is faced with an upper bound of 0.768 F1 for Medline and 0.680 F1 for EMEA which could be difficult to overcome. We argue that the combination of a cross-lingual approach with a native, lexically-based information retrieval procedure appears to be a reasonable option when dealing with low-resourced languages such as Spanish.

6 Conclusion

Facing the increasing availability of biomedical information in a multilingual setting requires advanced natural language processing tools to appropriately transform unstructured information into standardised, structured knowledge. This structured knowledge could better serve medical experts in tasks such as information retrieval by increasing the effectiveness of pure lexical-based systems. At the same time, the transformation of raw text into more structured representations could pave the way for information discovery and automated reasoning and inferencing in Bioinformatics. Although much has been done in terms of terminology identification and conceptual indexing in biomedicine, most approaches deal with the English language only for which a considerable body of lexical resources and tools exist. We were here concerned with the development of tools and resources for the appropriate recognition and conceptual indexing of terms of multilingual biomedical literature. In particular, we focus on experiments regarding parallel documents in English and Spanish. Our main contribution is a set of methods based on a combination of multilingual and cross-lingual approaches. Two independent methods –a native approach based on a multilingual lexicon and a cross-lingual transfer approach based on machine translation– are thoroughly evaluated in two datasets to assess their effectiveness. Interestingly, both transfer and native methods outperform MetaMap adapted to Spanish. Individually, the system exploiting curated medical terminology for Spanish (UMLSmapper) obtains better results in terms of F1 than the ones using state-of-the art machine translation (general or adapted to the medical domain). Moreover, taking advantage of both methods, which differ in terms of precision and recall, an improved method is obtained. Two different settings of the transfer approach are also tested so as to assess the influence of machine translation during transfer. Overall, our research provides a set of new methods for dealing with multilingual and cross-lingual biomedical information, which although tested in Spanish, could be adapted to other under resourced languages with presence in UMLS and off-the-shelf machine translation.

There are several interesting research lines worth to investigate. For instance, we would like to extend this research to alternative system configurations and ensembles (i.e. UMLSmapper+transfer, Statistical MT instead of Neural MT, etc.), larger datasets and languages other than English and Spanish. Furthermore, exploiting the soft-alignment vectors that the NMT system generates could help overcome the problem of word-order difference between the original texts and the translations produced, thus improving the recall of the transfer pipeline. Finally, exploiting cross-lingual word embeddings obtained with and without the use of parallel corpora also looks very promising.

Acknowledgements

We want to thank Thierry Etchegoyhen from Vicomtech for helping on adapting the NMT system to the medical domain.

Funding

Our work is partly supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502) and the projects CROSSTEXT (TIN2015-72646-EXP, MINECO/FEDER, UE) and DeepReading (RTI2018-096846-B-C21 MCIU/AEI/FEDER, UE).

References

Accuosto, P. and Saggion, H. (2018). Improving the accessibility of biomedical texts by semantic enrichment and definition expansion. *Revista de Procesamiento del Lenguaje Natural*, **61**, 57–64.

Agerri, R. *et al.* (2014). IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3823–3828.

Agirre, E. and Soroa, A. (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 33–41.

Aronson, A. R. (2001). Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. In *Proceedings of the AMIA Symposium*, pages 17–21.

Aronson, A. R. (2006). MetaMap: Mapping Text to the UMLS Metathesaurus. *Bethesda, MD: NLM, NIH, DHHS*, pages 1–26.

Attardi, G. *et al.* (2013). Machine Translation for Entity Recognition across Languages in Biomedical Documents. In *CLEF (Working Notes)*.

Bahdanau, D. *et al.* (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*.

Berlanga, R. *et al.* (2010). Semantic annotation of biomedical texts through concept retrieval. *Revista de Procesamiento del Lenguaje Natural*, **45**, 247–250.

Bodnari, A. *et al.* (2013). Multilingual Named-Entity Recognition from Parallel Corpora. In *CLEF (Working Notes)*.

Bojanowski, P. *et al.* (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146.

Carrero, F. *et al.* (2008). Building a Spanish MMTx by using automatic translation and biomedical ontologies. In *Proceedings of*

the International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2008), pages 346–353.

Castro, E. *et al.* (2010). Automatic Identification of Biomedical Concepts in Spanish Language Unstructured Clinical Texts. In *Proceedings of the 1st ACM International Health Informatics Symposium (IHI'10)*, pages 751–757.

Cho, K. *et al.* (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1724–1734.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

Consortium, G. O. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, **32**(Database issue), D258–D261.

Cuadros, M. *et al.* (2018). Vicomtech at BARR2: Detecting Biomedical Abbrebiations with ML Methods and Dictionary-based Heuristics. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, pages 322–328.

Dai, M. *et al.* (2008). An efficient solution for mapping free text to ontology terms. *AMIA Summit on Translational Bioinformatics*, **21**.

Geer, L. Y. *et al.* (2009). The NCBI biosystems database. *Nucleic Acids Research*, **38**(Database issue), D492–D496.

GriffithsâŁłJones, S. (2004). The microRNA Registry. *Nucleic Acids Research*, **32**(Database issue), D109–D111.

Hastings, J. *et al.* (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*, **44**(D1), D1214–D1219.

Haveliwala, T. H. (2002). Topic-sensitive PageRank. In *Proceedings of the 11th international conference on World Wide Web (WWW'02)*, pages 517–526.

Intxaurrondo, A. *et al.* (2018). Finding Mentions of Abbreviations and Their Definitions in Spanish Clinical Cases: The BARR2 Shared Task Evaluation Results. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, pages 280–289.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New phytologist*, **11**(2), 37–50.

Kors, J. A. *et al.* (2015). A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. *Journal of the American Medical Informatics Association*, **22**(5), 948–956.

Manning, C. *et al.* (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014): System Demonstrations*, pages 55–60.

McCray, A. T. and Nelson, S. J. (1995). The Representation of Meaning in the UMLS. *Methods of Information in Medicine*, **34**(01/02), 193–201.

Neves, M. *et al.* (2016). The Scielo Corpus: a Parallel Corpus of Scientific Publications for Biomedicine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2942–2948.

Nunes, T. *et al.* (2013). BeCAS: biomedical concept recognition services and visualization. *Bioinformatics*, **29**(15), 1915–1916.

Oronoz, M. *et al.* (2013). Automatic annotation of medical records in Spanish with disease, drug and substance names. In *Proceedings of the 17th Iberoamerican Congress on Pattern Recognition (CIARP 2012)*, pages 536–543.

Papineni, K. *et al.* (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311–318.

Perez, N. *et al.* (2018). Biomedical term normalization of EHRs with UMLS. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2045–2051.

Rebholz-Schuhmann, D. *et al.* (2013a). Entity recognition in parallel multi-lingual biomedical corpora: The CLEF-ER laboratory overview. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 353–367.

Rebholz-Schuhmann, D. *et al.* (2013b). Evaluation and cross-comparison of lexical entities of biological interest (LexEBI). *PloS One*, **8**(10), e75185.

Roller, R. *et al.* (2018). Cross-lingual Candidate Search for Biomedical Concept Normalization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 16–20.

Savova, G. K. *et al.* (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, **17**(5), 507–5013.

Sennrich, R. *et al.* (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 1715–1725.

Soysal, E. *et al.* (2017). CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, **25**(3), 331–336.

Tseytlin, E. *et al.* (2016). NOBLE–Flexible concept recognition for large-scale biomedical natural language processing. *BMC Bioinformatics*, **17**(1), 32.

Ware, M. and Mabe, M. (2015). The STM Report: An overview of scientific and scholarly journal publishing. Technical report, International Association of Scientific, Technical and Medical Publishers.