LREC 2018 Workshop

MultilingualBIO:
Multilingual Biomedical Text Processing

PROCEEDINGS

Edited by

Maite Melero, Martin Krallinger and
Aitor Gonzalez-Agirre

ISBN: 979-10-95546-03-0
EAN: 9791095546030

8 May 2018

Proceedings of the LREC 2018 Workshop
"MultilingualBIO: Multilingual Biomedical Text Processing"

8 May 2018 – Miyazaki, Japan

Edited by Maite Melero, Martin Krallinger and Aitor Gonzalez-Agirre

https://multilingualbio.bsc.es/

Acknowledgments: This work has received funding from Spain's Plan for the Advancement of Language Technology of the Secretary of State for Information Society and Digital Agenda.

Organising Committee

- Maite Melero, OTG Plan Tecnologías del Lenguaje, Spain
- Aitor Gonzalez-Agirre, Centro Nacional de Investigaciones Oncológicas, Spain
- Martin Krallinger, Centro Nacional de Investigaciones Oncológicas, Spain
- Marta Villegas, Centro Nacional de Investigaciones Oncológicas, Spain
- Ander Intxaurrondo, Centro Nacional de Investigaciones Oncológicas, Spain
- Khalid Choukri, ELDA, France
- Núria Bel, Universitat Pompeu Fabra, Spain
- Monica Monachini, ILC-CNR. Italy
- David Pérez, Secretaria de Estado para la Sociedad de la Información y la Agenda Digital, Spain

Programme Committee

- Nigel Collier, EMBL-EBI, UK
- Hercules Dalianis, Stockholm University, Sweden
- Cristina España-Bonet, DFKI, Germany
- Aitor Gonzalez-Agirre, Centro Nacional de Investigaciones Oncológicas, Spain
- Jin-Dong Kim, DBCLS / ROIS, Japan
- Martin Krallinger, Centro Nacional de Investigaciones Oncológicas, Spain
- Anália Lourenço, Universidad de Vigo, Spain
- Paloma Martínez, Universidad Carlos III de Madrid, Spain
- Maite Melero, OTG Plan Tecnologías del Lenguaje, Spain
- Roser Morante, Vrije Universiteit Amsterdam, Holland
- Mariana Neves, German Federal Institute for Risk Assessment, Germany
- Alfonso Valencia, Barcelona Supercomputing Center, Spain
- Hua Xu, UTHealth School of Biomedical Informatics, US
- Pierre Zweigenbaum, CNRS, France

Preface

Clinical and biomedical text mining are popular tasks in NLP research that have reached considerable progress over the past years. However, the main achievements in processing of biomedical text are almost restricted to English, with most other languages lagging behind in this respect. In particular, one aspect that has not received enough attention so far is the multilingual aspect of Biomedical text processing. On the other hand, automatic translation strategies of English medical terms have resulted in promising attempts to increase the coverage of non-English terminologies (Afzal et al. 2015, Neveol et al. 2016, van Mulligen et al. 2016); machine translation and multilingual approaches have also been explored for entity recognition efforts in the biomedical domain (Rebholz-Schuhmann et al. 2013); finally, multilingual ontologies are valuable resources for disease surveillance systems (Collier et al 2006).

The need to translate biomedical texts occurs in many situations. To put an example, cross-border mobility of people may require specific translation of medical records and discharge reports. In addition, internationalization of the pharmaceutical industry demands that technical specifications and package leaflets of medicines be translated to the language of the customer in several countries; not to mention medical patent translation, which is a specific area by itself. Other common examples of translation of biomedical text are laboratory reports, clinical trials or scientific publications. The use of Machine Translation in scenarios such as the above is desirable, both to reduce costs and to ensure terminological consistency across languages. Machine Translation, one of the most challenging tasks in Natural Language Processing, has experienced a big quality leap in recent years, due to the application of deep learning techniques and the growing availability of bilingual in-domain corpora. Yet, high quality translation of specialized domains remains a challenge and is very much dependent on the availability of good quality parallel corpus in that domain.

May 2018

Programme

14:00 – 14:10 – Welcome and presentation
Maite Melero

14:10 – 14:40 – Introductory talk
Aurélie Névéol, *A scoping review of the literature on clinical Natural Language Processing in languages other than English*

Natural language processing applied to clinical text or aimed at a clinical outcome has been thriving in recent years. This presentation offers a broad overview of clinical Natural Language Processing (NLP) for languages other than English. It aims to provide insights to NLP researchers faced with establishing clinical text processing in a language other than English, as well as clinical informatics researchers and practitioners looking for resources in their languages in order to apply NLP techniques and tools to clinical practice or investigation. Recent studies are summarized to offer insights and outline opportunities in this area. Studies are classified into three groups: (i) studies describing the development of new NLP systems or components de novo, (ii) studies describing the adaptation of NLP architectures developed for English to another language, and (iii) studies focusing on a particular clinical application. The advantages and drawbacks of each method are discussed in light of the application context. Finally, the presentation will highlight major challenges and opportunities that will affect the impact of NLP on clinical practice and public health studies in a context that encompasses English as well as other languages.

Session 1
14:40 – 15:00
Salvador Medina and Jordi Turmo, *Building a Spanish/Catalan Health Records Corpus with Very Sparse Protected Information Labelled*

15:00 – 15:20
Lingyi Tang, Jun Xu, Xinyue Hu, Qiang Wei and Hua Xu, *Building a Biomedical Chinese-English Parallel Corpus from MEDLINE*

15:20 – 15:40
Gökhan Doğru, Adrià Martín-Mor and Anna Aguilar-Amat, *Parallel Corpora Preparation for Machine Translation of Low-Resource Languages: Turkish to English Cardiology Corpora*

15:40 – 16:00
Roland Roller, Madeleine Kittner, Dirk Weissenborn and Ulf Leser, *Cross-lingual Candidate Search for Biomedical Concept Normalization*

Programme

16:00 – 16:30 Coffee break

Session 2
16:30 – 16:50
Marta R. Costa-jussà, Noe Casas and Maite Melero, *English-Catalan Neural Machine Translation in the Biomedical Domain through the cascade approach*

16:50 – 17:10
Olatz Perez-de-Viñaspre, Maite Oronoz and Natalia Elvira, *KabiTermICD: Nested Term Based Translation of the ICD-10-CM into a Minor Language*

17:10 – 17:30
Marta Villegas, Ander Intxaurrondo, Aitor Gonzalez-Agirre, Montserrat Marimon and Martin Krallinger, *The MeSpEN Resource for English-Spanish Medical Machine Translation and Terminologies: Census of Parallel Corpora, Glossaries and Term Translations*

17:30 – 18:00 – Discussion and Closing

Table of contents

Building a Spanish/Catalan Health Records Corpus with Very Sparse Protected Information Labelled

Salvador Medina and Jordi Turmo
TALP Research Center - Universitat Politècnica de Catalunya
Carrer de Jordi Girona, 1-3, 08034 Barcelona
{smedina, turmo}@cs.upc.edu

Abstract

Electronic Health Records (EHR) are an important resource for the research and study of diseases, treatments and symptoms. However, due to data protection laws, information that could potentially compromise privacy must be anonymized before making use of them. Thus, the identification of these pieces of information is mandatory. This identification is usually performed by linguistic models built from EHRs corpora in which Protected Health Information (PHI) has been previously annotated. Nevertheless, two main drawbacks can occur. First, the annotated corpora required to build the models for a particular language may not exist. Second, unannotated corpora might exist for that language, containing very few words related to PHI mentions (i.e., very sparse population). In this situation, the process of manually annotating EHRs results extremely hard and costly, as PHI occurs in very few EHRs. This paper proposes an iterative method for building corpus with labelled PHI from a large unlabelled corpus with a very sparse population of target PHI. The method makes use of manually defined rules specified in the form of Augmented Transition Networks, and tries to minimize the seek of EHRs containing PHI, thus minimizing the cost of manually annotating very sparse EHRs corpora. We use the method with primary care EHRs written in Spanish and Catalan, although it is language-independent and could be applied to EHRs written in other languages. Direct and indirect evaluations performed to the resulting labelled corpus show the appropriateness of our method.

Keywords: sparse, anonymization, iterative method, Spanish, Catalan, health records

1. Introduction

The interest on identification Protected Health Information (PHI) in Electronic Health Records (EHR) with the objective of automatically de-identifying them has seen an important increment in recent years. For this reason, multiple PHI de-identification challenges have been issued, such as the Informatics for Integrating Biology to Bedside (i2b2) 2006 (Uzuner et al., 2007) and 2014 (Stubbs and Uzuner, 2015) or the 2016 CEGS N-GRID shared tasks (Stubbs et al., 2017). All of them focusing on the de-identification of English-written EHR following the guidelines by the Health Information Portability and Accountability Act (HIPAA).

The de-identification of PHI in health notes is indeed a very challenging task. To begin with, EHR are hard to parse, since they are often composed of unconnected observations in the form of short phrases containing severe syntactical and morphological errors. Moreover, PHI are usually uncommon, and can be easily confounded by procedures and drugs that are named after their developer. Because of this, general-purpose NER tools such as the dictionary-based *FreeLing* NER module are not appropriate for this task, and context-specific NER systems are required.

Thanks to the aforementioned challenges, multiple NER tools specifically crafted for PHI have been proposed. Supervised learning models such as the Bilinear Long Short-Time Memory (BiLSTM) network described by Dernoncourt (Dernoncourt et al., 2017) have managed to achieve remarkable results for PHI de-identification. However, with independency of the supervised model used, manually tagged corpora is mandatory for both training and evaluation. Moreover, this corpus should be big enough and representative of the diversity found in the unlabelled documents.

Several training corpora consisting of health records are available for English, most of them released for de-identification challenges, but others repurposed from de-identified health research datasets such as the MIMIC-II dataset (Saeed et al., 2002). Sadly, this is not the case for health records in languages other than English, for which labelled datasets are very limited. As a result, those state-of-the-art PHI de-identification supervised learning models cannot be adapted to health notes in other languages, due to the lack of annotated corpora.

Moreover, personal data protection laws, such as the Spanish *Ley de Protección de Datos*, do not allow researchers outside the health institutions to access identified health records. Consequently, the corpus must be manually labelled by the institution's personnel. However, due to the huge sparsity of PHI mentions in health notes (e.g. just less than 4 over 1000 words are names of person), the human annotators would be forced to check tens of thousands of documents to build a representative corpus. As a result, generating the needed training corpus can be prohibitively expensive for local health institutions.

In order to circumvent this drawback and make it cheaper for health institutions, we present an iterative method to build from scratch a corpus labelled with the occurrences of PHI. Inspired in active learning, the method selects relevant examples from unlabelled corpus using a set of manually defined search rules that is also enriched at each iteration. We used this method to build a bilingual Spanish/Catalan corpus with very sparse PHI labelled. Direct and indirect evaluations of the resulting labelled corpus are also reported.

The rest of the paper is structured as follows. Section 2. describes the iterative method used to select new relevant examples of PHI occurring in the unlabelled corpus. Sec-

tions 3. and 4. present the corpus to be labelled using our method, and both direct and indirect evaluations of the resulting labelled corpus, respectively. The results of these evaluations are described in Section 5.. Finally, Section 6. concludes.

2. The rule-based method

An strategy that is often used for building models when having a small labelled corpus but big sets of unlabelled data is active learning (Settles and Craven, 2008). Nevertheless, an initial training corpus is still required for building the initial supervised model, specially when applying state of the art models such as *Bilinear LSTM*, which may over-fit if the initial corpus is not large enough and active learning may not be successful as a result. Building such training corpus can be extremely time-consuming in the context of health records, since the density of some PHI categories such as names of people is significantly low (e.g., less than 0.28% of tokens in our corpus).

A possible alternative is to begin with a simpler manually defined rule-based system until the training corpus is big enough for a supervised model to be able to generalize from it. Regular expressions or gazetteer-based manual rules are commonly used for building simple initial models when not enough training data is available (Kozareva, 2006). The main issue with this approach, however, concerns diversity, as the examples obtained with a handcrafted rule-based system are usually similar among them and biased. This could be circumvented by defining several rules with minimal correlation, the main challenges being the ability to come up with diverse rules and knowing how many of them are needed.

Our approach revolves over the idea of starting from a diverse set of manually defined rules and defines an iterative methodology, inspired by active learning, for adding new rules to such set. New rules are defined and previous ones are refined with each iteration of our method so that the set keeps growing in complexity and diversity. We take profit of the expressiveness of Augmented Transition Networks (ATN) to be able to define and update such complex rules with ease.

2.1. The iterative method

The iterative method begins from an empty corpus. The first step is then to build a basic one by using text queries based on domain knowledge and gazetteers; and manually correcting a random sample of the retrieved documents, covering all possible categories. In our experiments, we began with 100 documents, but this would depend on the characteristics of the corpus and PHI categories.

With this basic corpus as a base, repeat the scheme below until the user is unable to come up with new rules given the requirements imposed to the F_1 score achieved by the rule set in the iterative training and validation corpora.

1. Run the set of rules against the training set and list the errors.
2. If a rule can be defined that covers more than one incorrect example in the training set without decreasing the F_1 score, add that rule. If no new rule can be defined, go to 4.
3. Evaluate using the validation set.
4. If recall is not increased and F_1 decreases, discard all the new rules and repeat from 2. If both precision and F_1 are decreased, update an existing rule so that precision increases in the training set and repeat from 3. Otherwise, repeat from 2.
5. Once no new rule can be added with the defined conditions, run the new set of rules against a subset of the unlabelled data.
6. Rank the new set of documents using the score function described in Section 2.2. and select those that are over the threshold. Repeat from 1.

2.2. Ranking and selection of new examples

The new set of documents that is added to the training set in each iteration is selected from the pool of documents by ranking them using the score function defined in equation 1 and discarding those below a given threshold score. We have designed this score function so that documents with multiple relevant instances are prioritized, specially those that belong to classes that are infrequent and hard to define manual rules for.

We determine the threshold score as the score that corresponds to the *elbow* point of the curve defined by the document's scores sorted in decreasing order. The *elbow criterion* is often used in cluster analysis to determine the optimal number of clusters so that adding another cluster does not give much better modeling of the data (Madhulatha, 2012). Similarly, in our case, we determine the number of selected documents so that including more does not add much more relevant examples.

$$f(d) = \sum_{i \epsilon K} N_i(d) * (1 - F_1(i)) * (1 - p_i)$$

$$p_i = \frac{\sum_{t \epsilon T} N_i(t)}{\sum_{i \epsilon K} \sum_{t \epsilon T} N_i(t)} \quad (1)$$

Where K is the set of classes, T is the set of documents in the training set, $N_i(d)$ is the number of examples of class i in document d and $F_1(i)$ is the F_1 score of class i in the validation set.

This score function prioritizes documents with multiple examples. What is more, the weight associated to each label decreases with increasing F_1 score for it - reaching 0 if $F_1 = 1.0$ - and a higher weight is assigned to labels that are uncommon in the training set.

2.3. The augmented transition networks

In our ATN implementation, sentences are parsed at token level by default, however we allow partial consumption of tokens via custom arc actions so that words with no spacing between them can be successfully identified.

Sentences are tokenized using the open-source *FreeLing 4.0* natural language processing suite (Padró and Stanilovsky, 2012), as it is the most feature-complete NLP

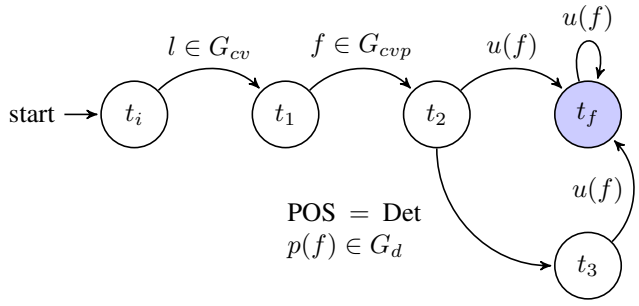


Figure 1: Example of an ATN rule. l, f and POS stand for lemma, form and Part of Speech respectively. $p(f)$ means to partially consume form f and $u(f)$ stands for uppercase. G_{cv}, G_{cvp} and G_d are gazetteers for communication verbs, communication verb pronouns and determinants. This rule can handle Examples **E.1** and **E.2**.

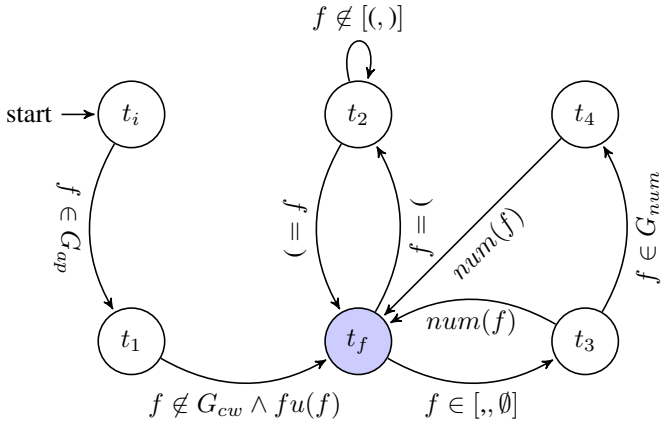


Figure 2: Fragment of an ATN rule. f stands for form. G_{ap} and G_{num} are gazetteers for addresses' prefixes and addresses' numbers prefixes respectively. $num(f)$ is a regular expression that determines whether f is a number or not. Maximum length of t_2 is limited to 5 tokens using edge actions and edge conditions. This rule can handle Examples **E.3** and **E.4**.

library for both Catalan and Spanish documents. However, due to the fact that *FreeLing 4.0* is optimized for texts written in standard language, the Named Entity Recognition and multi-word detection modules can lead to severe tokenization errors and we opted for disabling them.

Our ATNs can consume tokens based on their morphology, lemma or Part-of-Speech (POS) tag. Environmental variables can also be set based on the appearance of a certain token. In most of the cases, arcs check whether or not the token or sequence of tokens are included in a certain list of gazetteers, optionally adding restrictions relative to capitalization or POS tag.

The list below shows some examples of sentences that can be successfully parsed by some the of the rules that we have defined. A simplification of those rules are shown in Figures 1 and 2.

E.1 *Los derivo a bienestar social para hablar con Oliach.* (I derive them to social wellness to talk with Oliach).

E.2 *Parlo amb lAnna de la pauta a seguir.* (I talk to Anna about the guideline to follow).

E.3 *AVINGUDA MONTILIVI Nº 5 (al costat Suca-Mulla), tercer pis, porta D.* (5 MONTILIVI AVENUE (next to Suca-Mulla), third floor, door D.).

E.4 *AVENIDA DRASSENAS 17-21 TLF. 934416126.* (17-21 DRASSENAS AVENUE TEL. 934416126.). Note that *'DRASSENAS'* is misspelled, the correct form is *'Drassanes'*.

2.4. Observations

Given the characteristics of the iterative method that we propose, the resulting set of rules and labelled corpus ensures that:

- Rules that increase coverage (recall) are prioritized over those that increase precision, as the latter are not added unless the overall F_1 decreases.
- F_1 score increases monotonically in both training and validation set, first by increasing recall and then increasing precision in the latter iterations.
- New examples can be added indefinitely so that labelled sets of arbitrary size can be generated. As such, an additional stopping criteria should be defined. In this case, we stop once our set of rules surpasses 0.8 in the validation set.
- The resulting corpus does not maintain the proportions of entities found in the unlabelled corpus, since instances with a very low frequency and not easily identifiable are preferred.
- Documents with no entities are ignored so that the curators spend minimum time validating instances that do not contain relevant information.

The fact that the resulting corpus is unbalanced could potentially lead to worse performance in supervised learning models. Nevertheless, multiple methods have been presented over the last years that can be applied to unbalanced datasets so that this limitation is halved, which include semi-supervised algorithms (Huang and Kecman, 2004), re-sampling strategies (Liu et al., 2003) or weight adjusting mechanisms (Saerens et al., 2002).

3. The Spanish/Catalan corpus of health records

We apply the iterative method to a bilingual corpus of Electronic Health Records containing admission, progress, operative and discharge notes taken by doctors in the Catalan primary health care system. The goal is to be able to identify the Protected Health Information included in these records.

The *Institut Català de la Salut* (ICS) Primary Care Service's corpus of 2013 is composed by 12 files, each containing the short comments attached to the medical reports issued during the corresponding month of 2011. The notes are written in Spanish and Catalan, often combining words

	Full Corpus	Test	Validation
Notes	32882336	5000	311
Word occurrences	631020021	112281	15430
Tokens/Note	19.19	22.46	49.61
Words	3430167	33007	6346
Words ($F > 5$)	582047	2717	219
Spanish/Catalan	1.268:1	1.390:1	1.022:1

Table 1: Statistics of the Electronic Health Record corpus by the *Institut Català de la Salut* of 2013.

from both languages, and cover multiple fields: from common illnesses to psychology, dependency, drug use and so forth. Each record entry is identified by three numbers divided by vertical vars, but no additional structured information is provided. The first column in Table 1 summarizes some figures about the full unlabelled corpus. First row stands for the number of notes in the corpus; second and third rows refer to the number of word occurrences and the ratio of word ocurrences per note; fourth and fifth rows stand for the number of words without repetition with frequency greater than 1 and greater than 5, respectively; finally, the last row shows the ratio between Spanish and Catalan records within the corpus according to *FreeLing*'s language detection module.

3.1. Textual characteristics of the corpus

The documents in this corpus are written in natural language, usually composed of short sentences lacking verbal phrases and having severe non-gramatical morphological and syntactical phenomena. In addition to those, the list of phenomena listed below is recurrent in the corpus:

- Incoherent use of capitalization. For instance, *"realitzarem innmovilitzaació, recomanen e insisteim anar aH DE CALELLA PER CONFIRMAR FISURA I FRACTURA, DIU QUE NO HI ANIRÀ QUE NO VOL ESPERAR-SE 4 H.P:Realitzem inmovilització i control en una seetmana."* combines fully lowercased phases with fully uppercased ones.
- Use of contractions. An example of this can be found in the sentence *"Pac que finaliza tto"*, where the words *Pac* and *tto* are used instead of *Paciente* (patient) and *tratamiento* (treatment).
- Use of punctuation marks instead of spaces or lack of them. For example, in the sentence *"Algun subcrepitante en bases...Normas.Pulmicort-100 2-1(15 dias)."*, the words *bases*, *Normas* and *Pulmicort-100* are not spaced. What is more, in sentence *"Controlada HVhebron anualment."*, *HVhebron* should be *H. V. Hebron*, as it refers to *Hospital Vall Hebron*.
- Enumerations of measures and readings from medical analysis. For example, *"Usa L/C OD 85º-0.50 +1.00 0.8 /+4.00. OI 115º-1.00 +0.25 0.9 /+3.50.AO 4DP BT en VL.Rx ¿OD NG. OI NG Ad/3.00."*
- Inconsistent use of languages, since notes often combine Spanish and Catalan words, phrases or idioms. For instance, sencence *"M:febre de 39ºC tot el dia a pesar que la mare li ha donat Dalsy, vomits i mucositat nasal."* is written in Catalan but includes the Spanish expression *a pesar que* (despite of), while sentence *"E:herida mordida palma de mano D.P:neteja, steristrip..."* is written in Spanish but uses the Catalan verb *neteja* (to clean).

3.2. Protected Health Information categories

The PHI categories that we consider in this work follow the de-identification directives given by the *Institut Universitari d'Investigació en Atenció Primària* (IDIAP), a medical research center subordinated to the *Institut Català de la Salut* (ICS). A health note is considered successfully de-identified if the PHI entities listed below are replaced by their respective category name. Estimations of the proportions of tokens corresponding to each PHI category are given based on the observation of a subset of documents.

1. PERSON: Name or surname of a patient, relative, medical staff or any other person mentioned in the report. (about 0.28% of the tokens).
2. LOCATION: Physical locations or geographic subdivisions including street address, city, county, precinct, ZIP code, et cetera. This also includes public locations such as hospitals, clinics, schools and others. (about 1.12% of the tokens, 0.73% being public locations).
3. TELEPHONE: Digits of a phone number. (below 0.01% of the tokens).
4. EMAIL: E-mail address. (less than 0.01%).
5. DNI: Spanish *Documento Nacional de Identificación*. (less than 0.01% of the tokens).
6. SOCIAL_SECURITY_ID: Spanish social security number. (less than 0.01% of the tokens).
7. SANITARY_CARD_ID: Catalan sanitary card number. (less than 0.01% of the tokens).

It is also worth noting that there is a high degree of correlation between PHI in the health records. Based on the observation of a subset of documents, more than 50% of health records containing PHI include multiple instances, 44% of them having 3 or more. While the probability that a note contains any PHI is below 9%. As could be expected, health records that explicitly include personal information of a patient or doctor such as the name often include other information such as the names of relatives and partners, as well as working places, clinics etcetera.

4. Evaluation Framework

In order to evaluate the method that we are presenting, we apply both direct and indirect evaluation. First, we evaluate how the manual validation time by curators is optimized in terms of the faction of relevant examples presented to them. Additionally, we indirectly evaluate the corpus that is obtained during the iterative method in terms of the quality of a model trained with it.

	Validation	Test	Resulting Corpus
PERSON	372	282	699
LOCATION	99	680	825
TELEPHONE	7	6	17
Notes	311	5000	1051
Notes /w PHI	299	667	793

Table 2: Count of instances of PHI corresponding to categories PERSON, LOCATION and TELEPHONE in corpora

We restrict evaluation to the identification of instances of PHI that correspond to categories PERSON and LOCATION, even though the iterative process is applied to every category described in Section 3.2.. Categories TELEPHONE, EMAIL, DNI, SOCIAL_SECURITY_ID and SANITARY_CARD_ID have a formal structure and previous work in the subject has proven that simple regular expressions are enough to cover all instances (Yang and Garibaldi, 2015). Moreover, as shown in Section 3.2., the density of such entities in the full corpus reported by IDIAP is so low that instances in the evaluation corpus may not be representative enough. For these two reasons, we have opted for neglecting them in the evaluation figures.

4.1. Validation and testing partitions

Our method starts from a completely empty labelled corpus which grows at each iteration, jointly with the set of rules. In order to be able to evaluate each modification the set of rules, we have previously selected and manually labelled a small validation set of documents from the unlabelled corpus. Considering that our main goal is to embrace as much diversity as possible and we want to keep evaluation time as small as possible, this validation set is composed of just positive examples. These examples are selected by skimming the set of unlabelled documents and selecting those in which an example is spotted in order to lower the required building time.

The test set, which is used to perform the indirect evaluation of the final set of rules obtained after the iterative method, is composed of 5000 randomly selected documents from the whole set of unlabelled ones. Opposite to the resulting labelled corpus and the validation set, the test set maintains the proportion and density of PHI mentions of the unlabelled corpus.

Columns 2 and 3 of Table 1 show statistics about the number of health records and words in the Validation and Test corpora. Columns 1 and 2 of Table 2 list the amount of instances of each category of PHI that we evaluate in our work for these two corpora.

4.2. Direct evaluation of the guided labelling process

A way to measure the fraction of PHI in the health notes that are manually labelled while ensuring that an heterogeneous set of examples are retrieved is to look at the F_1 score. On the one hand, we would like the number of documents that are supposed to contain PHI to actually contain them, which means that the generated rules should be precise. On the other hand, relevant examples should not be ignored, so the rules are required to have a high *recall*. F_1 computes the harmonic mean of both these scores so it is the most suitable evaluation measure for this criteria. Note that strict evaluation must not be enforced, since the health notes are supposed to be manually labelled anyway. Exact matching of the entities' bounds is not mandatory and they will be considered as *true positive* if the labels contain any of the entities' tokens.

We evaluate the initial and final rule sets obtained after applying the iterative method using a test corpus composed of 5000 randomly selected and manually labeled health records described in Section 4.1. according to the evaluation criteria described above. Additionally, we compare the F_1 score achieved by the aforementioned context-specific sets of rules to the one using the general-purpose *Named Entity Recognition and Classification* (NERC) module included in *FreeLing*.

4.3. Indirect evaluation of the resulting corpus

The indirect evaluation of the corpus obtained as a result of the iterative process is done by using it as the training set of supervised PHI identification models. We train a Conditional Random Field (CRF) sequence tagger using morphological, part-of-speech, lemmas and clustered word-embedding input features, which is a widely-used model for PHI identification capable of achieving state of the art performance in recent de-identification shared tasks (Yang and Garibaldi, 2015), (Dehghan et al., 2015).

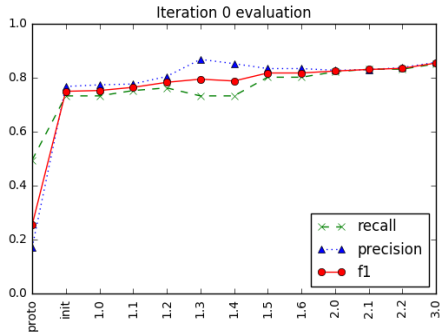
We compare the models trained with the iterative corpus to others trained using a larger corpus generated by randomly selecting and labelling examples from the unlabelled corpus. In particular, we take the test corpus disposed for the direct evaluation of the rule set and divide it into 8 folds, which are then re-purposed as 8 test corpora and 8 training corpora of 625 and 4375 health records respectively. The iteratively generated corpus is evaluated with each one of these test folds independently.

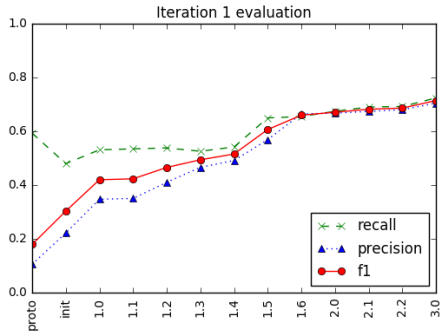
.

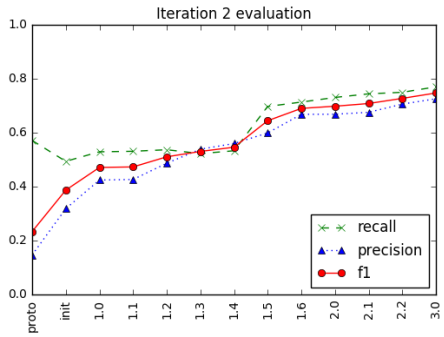
5. Results

Figure 3 shows the evolution of *recall*, *precision* and F_1 score evaluated using the training corpora obtained after iterations 0 to 2 for each update of the rule set. Given the conditions imposed by the iterative process, F_1 score in the validation set increases monotonically. Both *recall* and *precision* also have an ascending trend. This means that the set of rules is improved at each iteration, being able to cover a broader variety of common contexts of PHI while avoiding mislabelled instances.

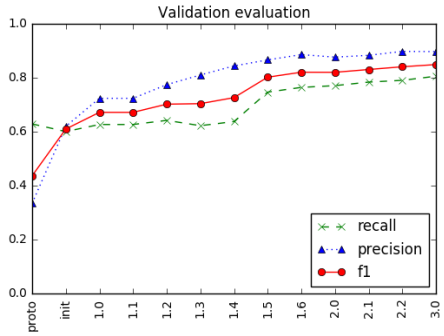
Table 3 shows the direct evaluation of *recall*, *precision* and F_1 scores in the test corpus for the *FreeLing NERC* module, as well as for the initial and final sets of rules. F_1 score is considerably lower than in the validation set, due to the fact that the latter only includes health records containing instances of PHI whereas the test corpus maintains the proportions of the full unlabelled corpus. The final recall is over 70% while precision is around 50%. This means that the set of rules is capable of retrieving training corpora including 70% of the instances of PHI in the unlabeled corpus showing 50% of false positives. Hence it can







(a) Iterative training corpora



(b) Validation corpus

Figure 3: Evaluation results for classes PERSON, LOCATION and overall for the iterative corpora built for iterations 0 to 2 (Figure 3a) and the validation corpus (Figure 3b).

considerably reduce the time required for the manual labelling process while discarding just 30% of the positive instances. Even though the iterative method achieves similar F_1 score for categories PERSON and LOCATION (0.564 and 0.509 respectively), the behaviour of *recall* and *precision* differs considerably. *Recall* for category PERSON is high compared to LOCATION (0.772 and 0.371 respectively), probably due to the fact that rules for LOCATION are less abundant but more precise, since they rely more in gazetteers.

	Eval.	NERC	initial	final
ALL	Recall	0.052	0.147	0.702
	Prec.	0.494	0.208	0.489
	F_1	0.094	0.172	0.576
PERSON	Recall	0.436	0.676	0.772
	Prec.	0.023	0.196	0.445
	F_1	0.044	0.304	0.564
LOCATION	Recall	0.517	0.013	0.371
	Prec.	0.064	0.127	0.809
	F_1	0.114	0.024	0.509

Table 3: Evaluation results in the test set for the general-purpose *Freeling* NERC module, and for the initial and final sets of hand-crafted rules.

	Eval.	Cross-Val.	Res. Corpus
ALL	Recall	0.721 (0.027)	0.699 (0.042)
	Prec.	0.839 (0.026)	0.769 (0.047)
	F_1	0.774 (0.017)	0.732 (0.039)
PERSON	Recall	0.784 (0.064)	0.759 (0.093)
	Prec.	0.909 (0.041)	0.730 (0.061)
	F_1	0.840 (0.025)	0.744 (0.057)
LOCATION	Recall	0.695 (0.040)	0.676 (0.056)
	Prec.	0.812 (0.022)	0.783 (0.061)
	F_1	0.748 (0.037)	0.726 (0.052)

Table 4: Mean *recall*, *precision* and F_1 score obtained by a CRF model trained using the labelled corpus obtained after 3 iterations of the method (1051 health records) compared to the *8-fold* cross validation of the test corpus (4350 health records) for the 8 testing partitions. Standard deviation is shown between brackets.

Table 4 shows the mean *recall*, *precision* and F_1 score of the indirect evaluation of the resulting labelled corpus after 3 iterations, compared to the 8-fold cross-validation of the test corpus used for direct evaluation. F_1 score using the iterative corpus is a 0.042 points lower compared to the traditional corpus, achieving similar *recall* (0.022 points lower) but significantly worse *precision* (0.07 points lower). This remarkable downgrade in *precision* is expectable, as the corpus has a higher density of positive examples. Nevertheless, the obtained results are promising, since they show that it is possible achieve similar *recall* after just 3 iterations. This leads us to believe that with more iterations and unsupervised re-sampling strategies to increase *precision*, the iteratively generated corpus could outperform the traditional one.

6. Conclusions

In this paper, we describe our method to build a corpus, in which protected information is labelled, from a large set of unlabelled electronic health records containing very sparse relevant information. Basically, hand-crafted rules are iteratively created for the automatic labelling of new protected information occurring in the health records. The re-

trieved documents that are considered most informative are selected and manually corrected in order to be used later to design new hand-crafted rules and refine the existing ones. Using this method, we created a bilingual Spanish/Catalan health records corpus with labelled protected information. We evaluated the resulting corpus in two ways: a direct evaluation by examining the relevance of the rules created at each iteration, and an indirect evaluation by comparing CRFs models learned using the resulting labelled corpus and a manually labeled extract of the full unlabelled corpus.

Given that we get comparable results using the iteratively generated corpus while requiring much less manual effort in terms of documents to be validated, we believe that the proposed method is appropriate for building inexpensive PHI identification training corpora. This more efficient use of resources is specially significant in subsequent iterations, as the proportion of PHI in the retrieved health notes increases and fetch rules grow in complexity.

We conclude that the presented method is a reasonable alternative to the much expensive process of uninformedly labelling random health records to build corpora when the density of target entities is low, while making minimal compromises to the final performance of the supervised models trained with it.

7. Acknowledgments

This works has been partially funded by the Spanish Goverment and by the European Union through GRAPHMED project (TIN2016-77820-C3-3-R and AEI/FEDER,UE.)

References

Dehghan, A., Kovacevic, A., Karystianis, G., Keane, J. A., and Nenadic, G. (2015). Combining knowledge-and data-driven methods for de-identification of clinical narratives. *Journal of biomedical informatics*, 58:S53–S59.

Dernoncourt, F., Lee, J. Y., Uzuner, O., and Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Huang, T. M. and Kecman, V. (2004). Semi-supervised learning from unbalanced labeled data–an improvement. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 802–808. Springer.

Kozareva, Z. (2006). Bootstrapping named entity recognition with automatically generated gazetteer lists. In *Proceedings of the eleventh conference of the European chapter of the association for computational linguistics: student research workshop*, pages 15–21. Association for Computational Linguistics.

Liu, B., Dai, Y., Li, X., Lee, W. S., and Yu, P. S. (2003). Building text classifiers using positive and unlabeled examples. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 179–186. IEEE.

Madhulatha, T. S. (2012). An overview on clustering methods. *arXiv preprint arXiv:1205.1117*.

Padró, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Saeed, M., Lieu, C., Raber, G., and Mark, R. G. (2002). Mimic ii: a massive temporal icu patient database to support research in intelligent patient monitoring. In *Computers in Cardiology, 2002*, pages 641–644. IEEE.

Saerens, M., Latinne, P., and Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41.

Settles, B. and Craven, M. (2008). An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1070–1079. Association for Computational Linguistics.

Stubbs, A. and Uzuner, Ö. (2015). Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of biomedical informatics*, 58:S20–S29.

Stubbs, A., Filannino, M., and Uzuner, Ö. (2017). De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1. *Journal of Biomedical Informatics*.

Uzuner, Ö., Luo, Y., and Szolovits, P. (2007). Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

Yang, H. and Garibaldi, J. M. (2015). Automatic detection of protected health information from clinic narratives. *Journal of biomedical informatics*, 58:S30–S38.

Building a Biomedical Chinese-English Parallel Corpus from MEDLINE

Lingyi Tang, Jun Xu, Xinyue Hu, Qiang Wei, Hua Xu
School of Biomedical Informatics, The University of Texas Health Science Center at Houston
7000 Fannin St, Suite 870, Houston, TX U.S.A 77030
{Lingyi.Tang, Jun.Xu, Xinyue.Hu, Qiang.Wei, Hua.Xu}@uth.tmc.edu

Abstract

In this paper, we present a biomedical Chinese-English parallel corpus aligned at sentence level. We collected biomedical publications that are available in both Chinese and English from MEDLINE and generated a dataset of 5,129 bilingual abstracts. We then employed the Champollion aligner, which uses both lexicon and sentence length features, to align the sentences in both languages. The aligned parallel corpus contains 61,874 English sentences and 43,866 Chinese sentences. The corpus is still under development, as we are manually checking sentence boundary and the alignment. We believe such a publically available corpus will benefit the development of cross-lingual systems and applications in the biomedical domain.

Keywords: Parallel Corpus, Biomedical, Sentence Alignment, Chinese-English

1. Introduction

A parallel corpus is a collection of texts with the translation of one or more languages besides the original one, where the texts, paragraphs, sentences and words are typically linked to each other. The most common case of parallel corpora is bitext where only two languages are studied. Parallel corpora play a vital role in many natural language processing (NLP) tasks and applications, especially in Statistical Machine Translation (SMT). In addition, it is also a valuable resource for a wide range of multi-lingual research, such as cross-language information retrieval and information extraction. Parallel corpora can also serve as a reference to check whether the element of the words or phrases is correct when dealing with bi-lingual texts for translators and researchers.

Over the past several years, parallel corpora have been constructed widely in general domain, including resources for translation between English and Chinese texts. For example, the Institute of Computational Linguistics of Peking University has developed a large-scale Chinese-English Contrastive Language Knowledge Base （CECLKB）, containing 7.5 million words/characters (Bai et al., 2002). CECLKB is a sentence-aligned parallel corpus developed for formal description of Chinese and English sub-sentential comparison. It is in the XML format with contrastive knowledge for future implementation of NLP systems in a multilingual environment. Later, Bai et al. (2014) refined CECLKB with updated architecture, improved entry selection and implementation of XML-based annotation schemes. It serves as a general knowledge base for many NLP tasks such as Computer-Assisted Translation and Second Language Acquisition. In 2010, Mohammadi & GhasemAghaee (2010) proposed a Bilingual Parallel Corpora Base using Wikipedia. Researchers show increasing interest in Wikipedia because it is cooperatively edited, which makes its content up-to-date. In addition, it is entirely free and available in most of the languages in the world. Another advantage of leveraging Wikipedia for parallel corpus development is that its structures (e.g., definition sections followed by description sections) may remain similar to the same topic written in different languages. All these attributes have made Wikipedia a great resource for multilingual parallel corpora establishment. Another sizeable parallel corpus worth notice is the Hong Lou Meng Parallel Corpus developed by Yanshan University sponsored by National Social Science Foundation of China (Liu et al., 2008). It contains the original texts of 120 Chapters in Chinese and three representative translation texts in English which can be used for both independent or collaborative research. The whole corpus is aligned at sentenced level so that it can benefit cross-lingual information extraction and information retrieval. Another critical Chinese-English corpus is the Bilingual Corpora of Tourism Texts Corpus developed at the Hong Kong Polytechnic University (Li et al., 2010). Later Sun et al. (2014) further improved the bilingual tourism texts with travel guides, tourist information and travelogues. They demonstrated the use of thematic and formal features when translating Chinese tourism discourse text to English. However, as discussed in their paper, the translation in the tourism domain could be a relatively easy task, as the average word count of the original Chinese texts is usually similar to that in the translated English texts.

However, there is very limited work regarding building parallel Chinese-English corpora in the biomedical domain. One significant Chinese-English parallel corpus in the medical domain is PCMW (Chen & Ge, 2011), which contains texts from 15 English medical books, with corresponding translation texts in Chinese. Apparently, it is an excellent resource for developing medical Chinese-English machine translation systems as these books have a broad coverage of medical sub-domains. However, it is not open to public at this time. Nevertheless, there are multilingual biomedical corpora available in other languages, and they have been used for different NLP tasks. For example, Deleger et al. developed a word alignment method to automatically acquire translation between medical terms in English and French using existing parallel corpora (Deléger, Merkel, & Zweigenbaum, 2009). It would be interesting to conduct similar research to expand medical terminologies in

Chinese, if such biomedical Chinese-English parallel corpora are available.

This paper describes our effort on building a biomedical Chinese-English parallel corpus using MEDLINE abstracts that are available in both Chinese and English. We applied and evaluated two different algorithms to automatically construct the aligned parallel corpus at sentence level.

2. Method

This study attempts to automatically build a parallel corpus using available bilingual abstracts from MEDLINE. We downloaded and extracted all available bilingual abstracts from PubMed and then applied two different sentence alignment approaches to align sentences. To evaluate the performance of sentence alignment, we annotated a small dataset and used it to report the performance of each method. We plan to continue manually reviewing sentence alignments, thus to improve the quality of the parallel corpus.

2.1 Data Collection

We used the query "Chinese[Language]" to search biomedical abstracts that are originally in Chinese from the PubMed. Then we downloaded all MEDLINE entries (including titles and abstracts) of relevant publications in the search results using the Entrez Programming Utilities (E-utilities) [1]. We obtained a collection consisting of 289,597 publications in XML format. We then extracted Chinese and English abstract pairs by parsing the XML documents. Most of the records contain abstracts in a single language only. Finally, we obtained a dataset containing 5,129 pairs of abstracts in both English and Chinese.

2.2 Preprocessing

After a bilingual abstract was extracted from XML document, further preprocesses were applied. For English abstracts, a regular expression-based sentence boundary detection program and a tokenization program developed in our lab were used to break each English abstract into sentences and tokens (Soysal et al., 2017).

Some Chinese abstracts in the collection were written in traditional Chinese. For consistency, we converted them to simplified Chinese (used in mainland China) using OpenCC[2], an open source simplified-traditional Chinese conversion tool. We split the abstracts into sentences using punctuation marks such as "。", "？" and "！"., which are used to indicate a full-stop of a sentence. Since Chinese is standardly written without spaces between words, we used JieBa [3] Word Segmenter to split a sentences into a sequence of words.

2.3 Sentence Alignment

Researchers have proposed a number of automatic sentence alignment approaches, mainly in two categories: length-based and lexical-based. Length-based algorithms are simple: it aligns sentences based on their length in terms of words or characters, such as the Gale-Church algorithm (Gale et al., 1993). The Gale-Church algorithm's assumption is that long sentences in original language should be long sentences in the translated texts, while short sentences tend to be short in the translated texts. The algorithm uses dynamic programming to search the best alignment using the probabilities produced for each bead based on the sentence length. It performs well on texts written in alphabetic languages.

However, length-based alignment algorithms do not perform well on aligning syllabic/logographic language (e.g., Chinese) to alphabet language (e.g., English)(Tan et al., 2014). The Champollion algorithm uses both sentence length features and lexicon features to align two sentences. It borrows the idea of *tf-idf* weight to calculate the similarity between two text segments. *tf-idf* is a statistic that reflects how important a term is to a document in a corpus, which has been widely used in information retrieval tasks. The Champollion algorithm also defines the segment-wide term frequency (*stf*), i.e. the number of occurrences of a term within the segment. Thus, the *stf* is able to measure the importance of the term within the particular segment. The algorithm uses the combination of *stf* and *tf-idf* to evaluate the importance of a translation term pair. A higher weight will be assigned to a less frequent translation term pair because these term pairs provide stronger evidence for aligning two segments. For instance, in the following sentence pair in one abstract:

Chinese: **出院** 时仅 有 5 例 （ **7.1‰** ） 诊断 慢阻肺 。

English: And only 5 patients (**7.1‰**) were diagnosed as COPD at **discharge**.

We can tell that the pair (7.1‰, 7.1‰) is a stronger piece of evidence indicating the two sentences should be aligned because they appear less frequently at the same time. The pair ("discharge" and "出院") appears much more often than "7.1‰" in bilingual biomedical abstracts. Thus, the translation pair (7.1‰, 7.1‰) has a much higher weight than the pair of (discharge, 出院). This is in concordance with the observation.

For any two segments, the Champollion algorithm calculates the similarity score based on the term pair weight, the number of sentences, and the sentence length. Moreover, the dynamic programming algorithm is used to search the path with maximum similarity, i.e., the prediction of the best alignments.

In this study, we used the Champollion algorithm as our primary sentence alignment method, as it has been reported with a high overall performance on English-Chinese sentence alignment task (Li et al., 2010). In addition, we also included the Gale-Church algorithm as a baseline.

2.4 Manual Correction

The automatic sentence alignment approach does not achieve a 100% accuracy. To build an accurate parallel corpus with sentences aligned, we implemented a manual review process on the top of the automated system. This process is still ongoing.

[1] https://www.ncbi.nlm.nih.gov/books/NBK25501/

[2] http://opencc.byvoid.com

[3] https://github.com/isuhao/jieba

2.5 Evaluation

To evaluate the sentence alignment algorithm, we constructed a gold standard dataset of 100 English-Chinese abstract pairs, which were randomly selected from the entire corpus and manually aligned. The review was performed by a native Chinese speaker who is also fluent in English. The results of sentence alignment were measured in Precision, Recall and F-measure. The formula for each measurement is listed as below.

$$\text{Precision} = \frac{\#\text{Correct\_links}}{\#\text{Predicted\_links}}$$

$$\text{Recall} = \frac{\#\text{Correct\_links}}{\#\text{Reference\_links}}$$

$$\text{F-measure} = 2\frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

3. Results and Discussion

Table 1 shows the descriptive statistics of the constructed parallel corpus. The generated corpus contains 5,129 bilingual abstracts collected from the biomedical domain. There are 61,874 English sentences and 43,866 Chinese sentences.

#Documents	5,129
# English sentences	61,874
# English tokens	1732K
# Chinese sentences	43,866
# Chinese tokens	1551K

Table 1 The statistics of the bilingual corpus

Type	Number	Percentage(%)
1→1	25,041	60.62
2→1	6,595	15.97
3→1	3,529	8.54
4→1	2,484	6.01
0→1	1,364	3.30
1→2	1,285	3.11
1→3	181	0.44
1→4	46	0.11
2→2	771	1.87
Others	13	0.03
Total	41,309	100

Table 2 The statistics of aligning English sentences to Chinese sentences in the corpus

Table 2 shows the numbers of different types of alignment between English and Chinese sentences. In this 'silver' standard corpus, we can find that almost 30% aligned sentences are the n→1 type. It is reasonable as Chinese texts tend to concatenate multiple clauses with commas. As a result, a corresponding Chinese sentence could consist of several English sentences in the original text.

Algorithm	Precision	Recall	F-measure
Champollion	0.8079	0.8338	0.8206
Gale-Church	0.2862	0.2954	0.2907

Table 3: The performance of the sentence alignment algorithms on the test dataset

Table 3 shows the results of the Champollion and Gale-Church algorithms on aligning sentences from English to Chinese using the test dataset of 100 manually reviewed abstracts. Our experiment results show that it is challenging for the Gale-Church algorithm to align syllabic/logographic language to alphabetic language, as it merely uses the statistic information from the unaligned text. The performance of the Champollion method is much better than the Gale-Church algorithm. However, the precision and recall are noticeably lower than those reported on the datasets from general domain, indicating it is more challenging to align sentences in biomedical texts. There could be several reasons for this finding, such as lack of Chinese-English medical term pairs, low performance of the Chinese segmentation program in biomedical text, mismatches of lengths of terms (e.g., abbreviations) etc. Further analysis is needed to accurately identify reasons for such errors and to identify potential solutions.

4. Conclusion

In this paper, we developed a parallel biomedical corpus with aligned sentences using a collection of bilingual abstracts collected from MEDLINE and an automated sentence alignment algorithm. The corpus contains 5,129 bilingual abstracts, 61,874 English sentences and 43,866 Chinese sentences. Our evaluation using a manually reviewed test set of 100 bilingual abstracts shows that the sentence alignment algorithm achieved an F-measure of 0.8206. We believe this parallel corpus will benefit the development of Chinese-English translation systems and applications in the biomedical domain.

The corpus is still under development: we are manually checking the sentence boundary and alignments generated by the automated algorithm. We will release the final corpus to the public once it is done. In the future, we will keep expanding it. About 80% abstracts in the original collection are written in English only. However, it is possible to further query Chinese bibliographic databases to find those abstracts in Chinese, thus expanding the parallel corpus.

5. References

Bai, X., Chang, B., Zhan, W., & Wu, Y. (2002). The construction of a large-scale Chinese-English parallel corpus. *In Recent Development in Machine Translation Studies-Proceedings of the National Conference on Machine Translation* (pp. 124-131).

Bai, X., Zähner, C., Zan, H., & Yu, S. (2014). The Chinese-English Contrastive Language Knowledge Base and Its Applications. *In Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data* (pp. 107–119). Springer, Cham. https://doi.org/10.1007/978-3-319-12277-9_10

Deléger, L., Merkel, M., & Zweigenbaum, P. (2009). Translating medical terminologies through word alignment in parallel text corpora. *Journal of Biomedical Informatics*, 42(4), 692-701.

Gale, W. A., & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational linguistics*, *19*(1), 75-102.

Gurulingappa, H., Rajput, A. M., Roberts, A., Fluck, J., Hofmann-Apitius, M., & Toldo, L. (2012). Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, *45*(5), 885-892.

Kim, J. D., Ohta, T., Tateisi, Y., & Tsujii, J. I. (2003). GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, *19*(suppl_1), i180-i182.

Li, D., & Wang, K. (2010). Development and application of bilingual corpora of tourism texts: A new approach [J]. *Modern Foreign Languages*, *1*, 007.

Li, P., Sun, M., & Xue, P. (2010, August). Fast-Champollion: a fast and robust sentence alignment algorithm. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 710-718). Association for Computational Linguistics.

Liu, Z., Tian, L., & Liu, C. (2008). The compilation of Hong Lou Meng Chinese-English parallel corpus [J]. *Contemporary Linguistics*, 4, 007.

Mohammadi, M., & GhasemAghaee, N. (2010). Building Bilingual Parallel Corpora Based on Wikipedia. *In 2010 Second International Conference on Computer Engineering and Applications* (Vol. 2, pp. 264–268).

Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., & Xu, H. (n.d.). CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*. https://doi.org/10.1093/jamia/ocx132

Tan, L., & Bond, F. (2014). NTU-MC toolkit: Annotating a linguistically diverse corpus. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations* (pp. 86-89).

Sun, Y., & Tang, F. (2014). A Parallel Corpus-based Investigation of Vocabulary Features of Tourism Translations1. *International Journal*, 2(3), 01-22.

Parallel Corpora Preparation for Machine Translation of Low-Resource Languages: Turkish to English Cardiology Corpora

Gökhan Doğru, Adrià Martín-Mor, Anna Aguilar-Amat
Universitat Autònoma de Barcelona, Universitat Autònoma de Barcelona, Universitat Autònoma de Barcelona
gokhan.dogru@uab.cat, adria.martin@uab.cat, anna.aguilar-amat@uab.cat

Abstract

High quality clean parallel corpora is a must for creating statistical machine translation or neural machine translation systems. Although high quality parallel corpora is largely available for official languages of the European Union, the United Nations and other organization, it is hard to encounter enough amount of open parallel corpora for languages such as Turkish, which, in turn, leads to lower quality Machine Translation for these languages. In this study, we use automatic and semi-automatic procedures to collect and prepare parallel corpora in cardiology domain. We crawl a journal website and obtain 6500 Turkish abstracts and their English translations by using HTTrack. By aligning these abstracts and converting them into a translation memory in a computer-aided translation tool environment, we make it possible to use the corpora for machine translation training as well as term extraction. We argue that new tools integrating and streamlining the web crawling, alignment and cleaning steps are needed in order to support the preparation of parallel corpora for low-resource languages.

Keywords: Parallel Corpora Preparation, Medical terminology, Low-resource languages

1. Introduction

Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) approaches have yielded important advances for creating automated translations with higher qualities compared to previous approaches. SMT is a mature study field and there are open source SMT systems such as Moses as well as free platforms based on Moses such MTradumàtica[1], which has a user friendly interface for training parallel corpora and creating an SMT engine. Both of these systems are corpus-based, namely, they need a large parallel corpus to learn from. As Forcada observes "[i]n both NMT and SMT, a target sentence is a translation of a source sentence with a certain probability of likelihood" (Forcada, 2017). Although research on NMT is relatively new, there are already some open source systems such as OpenNMT, AmuNMT and Nematus. Up until now, the automatic and human quality evaluation studies on the engines trained by these systems have reported that NMT engines yield slightly better quality results, though there are some studies claiming that SMT still gives higher quality. Although this discussion will likely carry on in the near future, there is one thing that everyone agrees: Since both of these approaches are data-driven, the quality of the parallel corpora used for training the systems plays a key role in the success of the respective system. Both of the approaches rely on high quality parallel corpora for the training of the SMT or NMT system. For this reason, the selection and preparation of the parallel corpora conditions the quality of resulting MT engine. This can be illustrated by the implementation of these systems to data-rich language pairs which have large amounts of high-quality open and free parallel corpora available on the internet.

With the availability of these huge amounts of data, domain-specific parallel corpora can easily be selected and the overall training data can be prepared in a short time (it is widely accepted that MT works better with domain-specific parallel corpora). However, not all language pairs have these readily available large amounts of data especially minoritized languages and less translated languages. And with the lack of parallel corpora, it becomes very difficult to create a machine translation engine for the low-resource language pairs. In the scope of this study, low-resource languages are defined as the languages with limited amount of high quality open parallel corpora on the internet. We think that with ever-increasing amount of open tools for MT, the high-quality parallel corpora selection and preparation will be very important to close the gap between low-resource languages including minoritized languages. Besides, we believe that one of the reasons why low-resource languages score lower in machine translation evaluation is that generally machine translation systems including these languages are trained with either small amount of data or low quality data.

Turkish language is spoken by nearly 90 millions of people. However, the parallel corpora having Turkish language (e.g. Turkish to English translations) in OPUS CORPUS, the biggest open parallel corpora on the internet, is very limited compared to other data-rich languages. And the available Turkish corpora is mostly obtained from volunteer or crowdsourced translation

[1] m.tradumatica.net. See Martin-Mor (2017) for the details of the project.

projects such Open Subtitle and Wikipedia. For achieving higher scores in machine translation quality evaluations, a Turkish to English machine translation engine will need domain-specific high quality parallel corpora. For example, it will be more convenient to create a medical domain English into French machine translation engine not only because these languages are grammatically similar but also because of the large amount of open parallel corpora available on the internet. By using open tools and state-of-the-art CAT tool features, we achieved to create parallel corpora of Turkish to English cardiology ready for machine translation training. We have selected cardiology because it is a very specific area with its own terminology and textual conventions.

2. Objective

The objective of this study is to show the methods that we have used in order to gather and prepare medical parallel corpora for the purpose of machine translation training. In this study, we report the automatic and semi-automatic methods we use for creating domain-specific (medical) custom translation memories as well as bilingual terminology lists which include web-crawling, document alignment in CAT tools and term extraction. We have crawled the web and obtained 6500 Turkish cardiology abstracts and their human translations into English using HTTrack, then we have aligned these abstracts using LiveDocs feature of Memoq and have created a translation memory of 1.200.000 words, of which we have extracted terms to be used both in the MT training and evaluation steps. This parallel corpora will be used in NMT and SMT training in the future, and evaluation of the quality of the engines through terminological quality. Using this method, it is possible to prepare large parallel corpora for language pairs lacking free and open training data.

3. Tools&Methods

Created in computer-aided translation (CAT) tools, translation memories are the most common parallel corpora types and constitute the most valuable data for machine translation training. They are exchanged between different CAT tools in TMX format which makes that highly reusable both in other translation projects and in machine translation or text extraction projects (and in other corpus analysis projects). However, most of the translation memories created in the industry are proprietary and are not open to be used by third parties. For this reason, finding open domain-specific parallel documents and process them to create translation memories will be beneficial for many shareholders including freelance translators, machine translation practioners as well as academic researchers.

Web crawling is an important step in collecting parallel corpora. There are many multilingual websites with open textual data. Nevertheless, although these websites may have a well-structured form, it is generally hard to first crawl them and then align their multilingual/bilingual content (say, English and Turkish) automatically. Bitextor[2] is a free and open automatic bitext generator which yields a TMX file after crawling and aligning a bilingual website. However, it only works with a few languages and does not include Turkish-English language pair. Besides, the quality of the automatic alignments is not enough to be used practically. For these reasons, we did not include this tool in our study. Yet, we think that the development of similar tools and the addition of an interface to be able to correct/validate alignment pairs will accelerate a lot the process of parallel corpora preparation.

The first step of our study is the selection of the data. Since corpus based machine translation systems function better with more specific data, we have focused on cardiology abstracts. We have used a free and open-source website copier called HTTrack to crawl the website of Archives of the Turkish Society of Cardiology[3] which both in Turkish and English. Since these abstracts have been translated from Turkish to English directly (without the intermediacy of another language), revised for an academic journal and published in this journal, we have assumed high quality in translation. In this decision, we take into consideration the "publishable quality" concept suggested by TAUS[4] for postediting machine translation output. We think that this criteria can also be applied during the selection of translations for MT training. We obtained 6500 abstracts (in total, 13000) in the form of HTML as a result of the web crawling. We used regular expression filters and cascading filters Memoq CAT tool to extract only the relevant abstracts, and finally aligned them using the LiveDocs feature of the same tool. Although segment by segment alignment has been very successful most of the time, a human intervention has been needed to fix the misalignments. However, this intervention has been needed at a minimum level because the automatic alignment has been very successful and the alignment editing interface of LiveDocs is very user-friendly. Upon completing the alignment process, it is possible to export the aligned segments directly into translation memory. We have obtained a TMX file in the end, which needed one more step to be cleaned since some HTML tags remained in our translation memory. We have used Olifant which is a translation memory editor to clean up the HTML tags. But not all statistical machine translation customisation platforms accept TMX files. Hence, depending on the platform to be used, a further step may be needed. While KantanMT[5] and Microsoft Translator

[2] http://bitextor.sourceforge.net
[3] http://www.archivestsc.com
[4] www.taus.net. Translation Automation User Society
[5] kantanmt.com

Hub[6] support TMX files, MTradumàtica currently requires aligned source language file and target language file separately. Okapi Framerwork[7], especially Tikal, can convert TMX files into two separate language files ready to be used for training. With these steps, we have created a translation memory of 1.200.000 words. This amount of corpus is of course not sufficient to train a state-of-the-art machine translation engine but by finding more data and speeding up these process, we are planning to gather more data in the future.

There are also two ways of extracting terminology out of the aligned segments. Firstly, it is possible to run an automatic monolingual term extraction feature to search for the most frequent words and then add their target term counterpart and validate them. The second strategy is to select the source term and target term manually in the LiveDocs interface and save them as term. Although the second option seems to be time-consuming, in fact, it can be done very fast and a glossary/terminology list can be created in a short time. And then, this terminology/glossary list can be used in the statistical machine translation training and increase the terminological quality of the output.

4. Turkish-English Parallel Corpora in Cardiology Domain

In corpus-based approaches of machine translation, the more specific the training corpus domain, the better the translation output will be. As Wolk and Marasek highlights, "SMT systems should work best in specific, narrow text domains and will not perform well for a general usage" (Wolk and Marasek, 2015). Similar observation is made by Lumeras and Way (2017): "It is well-known that MT systems work best when tested on data that is very similar to the corpora on which they are trained. For example, it would be foolish to translate weather forecasts using an SMT system that had been trained on parliamentary texts. So, for a particular company, it might not make sense to talk about their single English-to-French engine, but rather a whole suite of engines for this language pair depending on the domain (or in the field of Language for Specialised Purposes, what is known as "text genre"), e.g. patents, trials, white papers, personnel texts, product documentation, legal texts, contracts etc". However, in this point, we come across a paradox: although we will be more likely to create a better engine when we have a specific domain, the more our domain is specific, the less amount of text we will likely have. This is especially true for low-resource languages like Turkish. If we would like to create high-quality custom engines in the future, we will need to create both sufficient amount of domain-specific open parallel corpora and tools and methods to create this corpora. Since professional translators and translation companies continue to translate through their CAT tools, which, in turn, leads to the accumulation of new translation memories (which constitute the most widely used parallel corpora). Yet, these translation memories are proprietary and it is very unlikely that they will be made public due to the privacy agreements between translators/translation companies and clients. Hence, academic studies should be made to create (from openly available data) the relevant parallel corpora for different text genres for low-resource languages so that new corpus-based MT systems (SMT and NMT) can be studied in academic fields or implemented in industrial environments.

There are different text types (domains, as MT practitioners call them) with different textual conventions and terminologies. Since 1950s, many different text typologies have been suggested, one of the most well-known being Reiss and Vermeer's (2004) three-fold text type distinction where they classify texts as informative, operative and expressive texts. We believe that an understanding and use of text typology in the process of preparing parallel corpora are essential and are the added-values to be brought by the translators to the development of machine translation systems.

With this in mind, we have kept our focus very narrow in our study. Instead of concentrating on a more general text domain like medical domain, we focus on cardiology domain which is an area of expertise with its own terminology. We have crawled only Turkish and English abstracts in a scientific cardiology journal. Scientific abstracts tend to have their own text conventions including word choice, character or word limitations, and (and when applicable just like the case of medical articles) conventional headings such as "objective", "methodology", "results" and "conclusion".

We can also resort to another classification to understand better our data. Montalt (2010) groups medical genres under four classes in his social function oriented distinction: research, professional, educational and commercial. This classification is helpful in the parallel corpora preparation phase since it constrains the text type selection, and as mentioned above, the more the data is of a specific domain, the better the statistical machine translation engine is. Cardiology articles and abstracts are research-based medical texts and their translations. These are "those used by researchers to communicate their findings and arguments: original articles, case reports, doctoral theses, etc." (Montalt, 2010). Note that in such a MT engine preparation process, the work begins by selecting the specific text domain. The next step will be to search, crawl and collect open resources considering the amount and type of the domain.

5. Conclusion

Most of the studies concentrate on the evaluation of machine translation systems as well as procedures for postediting. We think that concentrating on the parallel corpora selection, collection and preparation processes is equally important and may have a positive impact on the later processes. And this is where translators can prove

[6] hub.microsofttranslator.com

[7] http://okapiframework.org

their added-value to the partial automation of translation process. The further creation of free and open parallel corpora will also make it possible for freelance translators and small language service providers to be able to use open source machine translation platforms such as MTradumatica, and to customize their own engines without paying large amounts of money. Hence, they will be able to compete in respective areas.

Another important aspect of optimizing the linguistic data (parallel corpora in our case) preparation is to help minority languages which have scarce data or no data to be able to make use of the data-driven approaches such neural machine translation and statistical machine translation, translation technologies and corpus technologies which, all together, have a potential to empower these languages.

Acknowledgements

This work has been supported by the FI-DGR-2017 grant cofinanced by European Social Fund and Generalitat de Catalunya.

6. Bibliography

1. The Opus Corpus. http://opus.nlpl.eu [last access : 05.02.2018]
2. Olifant. Okapi Translation Memory Editor. http://okapi.sourceforge.net/Release/Olifant/Help/ [last access : 05.02.2018]
3. Memoq. http://memoq.com [last access : 05.02.2018]
4. MTradumatica. http://m.tradumatica.net [last access : 05.02.2018]
5. Martín-Mor, Adrià (2017) MTradumàtica: Statistical Machine Translation Customisation for Translators. *Skase journal of translation and interpretation*, 11, p. 25-40
6. Forcada, Mikel L. (2017) Making sense of neural machine translation. *Translation Spaces*, 6, p. 291-309
7. Wołk, K. and Marasek, K. (2015) Polish - English Statistical Machine Translation of Medical Texts. In Aleksander Zgrzywa, A., Choroś, K., Siemiński, A. [eds.] *New Research in Multimedia and Internet Systems*. pp. 169-179.
8. Reiss, K. and Vermeer, H. (2004) *Towards a General Theory of Translational Action. Skopos Theory Explained*. Routledge.
9. Montalt, V. (2010) "Medical translation and interpreting". In Gambier, Y. and Doorslaer, L. [eds.] Handbook of Translation Studies. Vol. 2. pp. 79-83.
10. TAUS Postediting Guidelines. https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines (last access: 06.03.2018)

Cross-lingual Candidate Search for Biomedical Concept Normalization

Roland Roller[1], **Madeleine Kittner**[2], **Dirk Weissenborn**[1], **Ulf Leser**[2]
[1]German Research Center for Artificial Intelligence (DFKI), Language Technology, Berlin, Germany
[2]Humboldt Universität zu Berlin, Knowledge management in Bioinformatics, Berlin, Germany
{firstname.surname}@dfki.de, {surname}@informatik.hu-berlin.de

Abstract

Biomedical concept normalization links concept mentions in texts to a semantically equivalent concept in a biomedical knowledge base. This task is challenging as concepts can have different expressions in natural languages, e.g. paraphrases, which are not necessarily all present in the knowledge base. Concept normalization of non-English biomedical text is even more challenging as non-English resources tend to be much smaller and contain less synonyms. To overcome the limitations of non-English terminologies we propose a cross-lingual candidate search for concept normalization using a character-based neural translation model trained on a multilingual biomedical terminology. Our model is trained with Spanish, French, Dutch and German versions of UMLS. The evaluation of our model is carried out on the French Quaero corpus, showing that is outperforms most teams of CLEF eHealth 2015 and 2016. Additionally, we compare performance to commercial translators on Spanish, French, Dutch and German versions of Mantra. Our model performs similarly well, but is free of charge and can be run locally. This is particularly important for clinical NLP applications as medical documents underlay strict privacy restrictions.

Keywords: Candidate Search, Concept Normalization, UMLS, Multi-Lingual

1. Introduction

Concept normalization is the task of linking a text mention to a corresponding concept in a knowledge base (KB). This is useful to determine its distinct meaning and to include additional information linked through that knowledge base. The Unified Medical Language System (UMLS) (Bodenreider, 2004) is a large biomedical knowledge base which unifies different terminologies and their concepts, also across languages.Although UMLS includes synonyms and lexical variants for each concept, these are usually not exhaustive, since mentions in natural language can be expressed in many different ways which makes the task of concept normalization challenging. When dealing with non-English biomedical or clinical texts concept normalization becomes even more difficult as compared to English other languages are underrepresented in UMLS in terms of number of concepts or synonyms. Currently, UMLS[1] includes 25 different languages represented by 13,897,048 concept names (terms) which describe 3,640,132 individual concepts. The majority of concept names are English ($\approx$ 70%). Concept names in other languages make out a much smaller part: for instance, Spanish $\approx$ 10%, French $\approx$ 3%, Dutch $\approx$ 2%, and German $\approx$ 2%.

To support non-English biomedical concept normalization two approaches can be observed: (i) translating terms or whole documents from the target language to English and search in the English knowledge base or (ii) translating relevant subsets of the English knowledge base to the target language in order to expand the target knowledge base. Both approaches have been used with varying success for different languages, for instance for French (Afzal et al., 2015; Jiang et al., 2015; Van Mulligen et al., 2016) or Italian (Chiaramello et al., 2016).

All of these studies apply commercial tools such as Google Translate[2] or Bing Translator[3] for translation. Despite the good results, web-based translators can not be used when dealing with clinical documents. These data underly strict privacy restrictions and can not be shared online. Therefore, adaptable, local translation models are needed for NLP research in the biomedical and clinical domain.

In this work, we present a sequential cross-lingual candidate search for biomedical concept normalization. The central element of our approach is a neural translation model trained on UMLS for Spanish, French, Dutch and German. Evaluation on the French Quaero corpus shows that our approach outperforms most teams of CLEF eHealth 2015 and 2016. On Mantra we compare the performance of our translation model to commercial translators (Google, Bing) for Spanish, French, Dutch and German. Our model[4] performs similarly well, but can be run locally and is free of charge.

2. Related work

Concept normalization of non-English biomedical text has been the subject of several CLEF challenges. In CLEF eHealth 2015 Task 1b (Neveol et al., 2015) and 2016 Task 2 (Neveol et al., 2016) named entity recognition and normalization was performed on the French Quaero corpus containing Medline titles and EMEA abstracts. Apart from other tasks teams were asked to perform concept normalization using gold standard annotations.

The best performing team in 2015, team Erasmus (Afzal et al., 2015) used a rule based dictionary lookup approach. Erasmus expanded the French version of UMLS by translating a potentially interesting subset of English UMLS to French using Google Translator and Bing Translate. Translations were only used when both translation systems returned the same result. Additionally, they applied several post-processing rules developed from the training data to remove false positives: preferring most frequently used

[1] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html, accessed January 7th 2018

[2] https://www.translate.google.com

[3] https://www.bing.com/translator

[4] The model is available here: http://macss.dfki.de.

concept IDs for certain terms or most frequent semantic type and concept ID pairs. The winning team in 2016, team SIBM (Cabot et al., 2016) used the web-based service ECMT (Extracting Concepts with Multiple Terminologies) which performs bag of words concept matching at the sentence level. ECMT integrates up to 13 terminologies partially or totally translated into French. Compared to those teams we do not translate terminologies but terms and search in full non-English and English subsets of UMLS. Additionally, we use a similar disambiguation procedure as (Afzal et al., 2015).

One team (Jiang et al., 2015) in 2015 translated gold standard annotations to English using Google Translate and then applied MetaMap (Aronson and Lang, 2010) for normalization. This approach only yielded moderate results which has similarly been shown for Italian (Chiaramello et al., 2016). We apply a sequential search for candidate concepts. English UMLS is only used when the search in Non-English UMLS was not successful. A similar procedure is usually applied when initially annotating non-English corpora (Névéol et al., 2014; Kors et al., 2015).

Most teams in CLEF eHealth 2015 and 2016 rely on commercial online translation systems. Since the use of such tools is questionable and most probably forbidden when dealing with medical text local solutions are needed. Local machine translation models for the biomedical domain have been developed previously, for instance as part of the CLEF ER challenge 2013 (Rebholz-Schuhmann et al., 2013) based on the parallel English, Spanish, French, Dutch and German Mantra corpus. Participating teams often used phrase-based statistical machine translation models (Attardi et al., 2013; Bodnari et al., 2013; Hellrich and Hahn, 2013). However, to develop more sophisticated neural translation models large sets of parallel sentences are required, whereas the Mantra corpus is rather small. We developed a neural translation model on parallel language data of UMLS and FreeDict to be used for cross-lingual candidate search in concept normalization.

3. A Neural Translation Model for concept normalization using UMLS

The following section describes the biomedical translation model and the sequential procedure for candidate search during concept normalization.

3.1. Translation Model

The central element of our cross-lingual concept normalization solution is a character-based neural translation model (Lee et al., 2016). The model does not require any form of segmentation or tokenization at all. We chose this model because many translations of biomedical concepts can be resolved by small amends, due to the common origin of many words, that can be captured by such a system. At the same time it has enough modeling capacity to learn translations rules that cannot be captured by simple surface-form rules.

Model The model embeds lower-cased characters of the source phrase into a 256-dimensional space. The embedded character sequence is processed by a convolution layer with $N = 16 + 32 + 64 + 64 + 128 + 128 + 256 = 688$ filters of varying width resulting in 688-dimensional states for each character position. These are max-pooled over time within fixed, successive intervals of $k = 5$. This effectively reduces the number of source states by a factor of 5. To allow for additional interaction between the pooled states, they are further transformed by a 2-layer highway network. Finally, the transformed states are processed by a bidirectional recurrent neural network, in particular a bidirectional GRU (Chung et al., 2014), to produce final encoder states. A 2-layer recurrent neural network with attention (Bahdanau et al., 2015) on the encoder states subsequently produces the translation character by character. For more in-depth, technical details we refer the reader to (Lee et al., 2016). The model is trained on mini-batches comprising 32 source phrases with their respective translations, using ADAM (Kingma and Ba, 2015) as optimizer. The initial learning rate is set to 10^{-3} which is halved whenever performance on the development set drops.

Dataset We train our system on a subset of UMLS concept translations combined with various English to target language dictionaries taken from the FreeDict project [5].There might be multiple target language translations for each English phrase, thus we employ a deterministic decoder. The model is trained to minimize the perplexity only on the transformation that is most likely under the current model, i.e., the transformation with the least loss.

3.2. Concept Normalization

We approach concept normalization in two steps. First a candidate search is carried out while terms are sequentially looked up in non-English and English versions of UMLS. This first step aims at achieving a high recall. Then a disambiguation step is applied in order to reduce the number of candidates while keeping the precision high. In the following details on both steps are described.

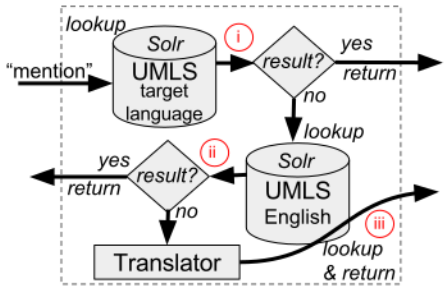


Figure 1: Sequential Candidate Search using mono-lingual (target lang.) and cross-lingual (English) UMLS subsets.

Candidate Search Concept terms of English, Spanish, French, Dutch and German versions of UMLS AB2017 are indexed and searched using Apache Solr 6.5.0[6]. A dictionary lookup always applies exact matching. If this first search does not return any results fuzzy matching is applied. Fuzzy matching uses a Levensthein edit distance of one per token for tokens larger than 4 characters otherwise the edit distance is set to zero.

[5] http://freedict.org/en/
[6] http://lucene.apache.org/solr/

As shown in Figure 1, we apply a sequential procedure for candidate search. Subsequently, the following searches are performed: (i) a mono-lingual candidate search looking up terms only in the respective non-English UMLS, (ii) a simple cross-lingual candidate search looking up the original term in English UMLS, and (iii) a cross-lingual candidate search in English UMLS while terms are first translated to English using our biomedical translation model. Once matching candidates are found the sequence is stopped. In this work we study how the addition of each search level improves concept normalization. We compare candidate searches up to level (i) mono-lingual (ML), level (ii) simple cross-lingual (CL), and level (iii) cross-lingual including translation of terms (BTM).

Disambiguation Our sequential candidate search may return a list of candidate concepts. This list is filtered by the following steps: (1) filter for known UMLS semantic groups or types, (2) prefer concepts with UMLS preferred labels, (3) filter by using densest-subgraph disambiguation (Moro et al., 2014), and (4) choose the smallest UMLS concept ID. Steps 1, 2, and 4 were similarly applied by (Afzal et al., 2015).

4. Evaluation

The performance of our sequential candidate search including our biomedical translation model is evaluated on two corpora: Quaero (Névéol et al., 2014) and Mantra (Kors et al., 2015). In the following the corpora and evaluation procedures are explained and results are presented.

4.1. Evaluation Corpora

Quaero The Quaero corpus contains French Medline titles and EMEA abstracts and has been used for named entity recognition and normalization tasks in CLEF eHealth 2015 Task 1b (Neveol et al., 2015) and 2016 Task 2 (Neveol et al., 2016). We compare our approach to results of teams performing best in the entity normalization task, Afzal et al. (2015) and Cabot et al. (2016). For that purpose we extracted gold standard annotations from the test corpora in 2015 and 2016. Note, the current (2016) version of Quaero contains a training, development and test set. The current development set is the test set of 2015.

Mantra The Mantra corpus contains Medline titles, EMEA abstracts and EPO patents for several languages including bi-lingual aligned sentences. For evaluation of our system we extracted gold standard annotations of Mantra Medline titles in Spanish, French, Dutch, and German. Note, that the Mantra Medline corpus is much smaller than the Quaero corpus. For evaluation we compare perfomance of our system to performance of commercial translators. Hereby, we apply the same sequential candidate search while instead of our biomedical translation model we use translations obtained manually from Google Translate and Bing Translator Similar to Afzal et al. (2015), only the first translation of each service was selected and used only if both systems returned the same translation.

4.2. Evaluation Results for Quaero

We evaluate performance of our proposed method against best performing systems in CLEF eHealth challenges 2016 and 2015. In 2016, team SIBM (Cabot et al., 2016) performed best on the task of gold standard entity normalization, see Table 1 for their results. In 2015, team Erasmus (Afzal et al., 2015) performed far better. They achieved an F1-score of 0.872 for EMEA and 0.671 for Medline, see Table 2. Moreover, the system was able to achieve a precision of 1 for EMEA.

	Medline			**EMEA**		
Method	P	R	F1	P	R	F1
ML	**0.800**	0.594	0.682	**0.822**	0.552	0.661
CL	0.786	0.620	0.693	0.808	0.676	0.736
BTM	0.771	**0.663**	**0.713**	0.781	**0.692**	**0.734**
SIBM	0.594	0.515	0.552	0.604	0.463	0.524

Table 1: Evaluation of mono- and cross-lingual candidate search for concept normalization on Quaero Corpus of CLEF eHealth challenge 2016 Task 2. We compare against SIBM (Cabot et al., 2016), the winning system of the challenge. Methods presented include mono-lingual (ML) and cross-lingual (CL) candidate search, and cross-lingual candidate search including translation of concept terms using our biomedical translation model (BTM). BTM outperforms SIBM on both, Medline and EMEA.

	Medline			**EMEA**		
Method	P	R	F1	P	R	F1
ML	0.831	0.575	0.680	0.911	0.632	0.746
CL	**0.834**	0.611	0.705	0.919	0.764	0.834
BTM	0.831	**0.661**	**0.736**	0.909	0.772	0.835
Erasmus	0.805	0.575	0.671	**1.000**	**0.774**	**0.872**

Table 2: Evaluation of mono- and cross-lingual candidate search for concept normalization on Quaero Corpus of CLEF eHealth challenge 2015 Task 1b. We compare against Erasmus (Afzal et al., 2015), the winning system of the challenge. Methods presented include mono-lingual (ML) and cross-lingual (CL) candidate search, and translation of concept terms using our biomedical translation model (BTM). BTM outperforms Erasmus on Medline but not on EMEA.

Cross-lingual candidate search including our biomedical translation model outperforms results for entity normalization of previous teams in three out of four data sets. In 2016, mono-lingual candidate search reaches highest precision for Medline and EMEA and already outperforms SIBM, see Table 1. Why the system of SIBM yields such poor results is not clear. One reason could be that their integrated and translated terminologies have a lower coverage in terms of concepts as full UMLS. Although most terminologies they integrate are part of UMLS. For ranking extracted candidates they use a classification based on most relevant term-semantic type relations which might not be as sufficient as our disambiguation procedure. Extending the search space by using the English UMLS subset in addition (CL) leads to further improvements in terms of recall and F1. Finally, including our translation system, in combination with a cross-lingual search (BTM) leads to the highest recall and the highest F1-Score.

	SPA			FRE			DUT			GER		
Method	P	R	F1	P	R	F1	P	R	F1	P	R	F1
ML	**0.799**	0.561	0.659	**0.814**	0.469	0.595	**0.800**	0.357	0.494	**0.833**	0.493	0.620
CL	0.788	0.583	0.670	0.795	0.502	0.615	0.769	0.424	0.546	0.817	0.530	0.643
BTM	0.781	**0.619**	**0.691**	0.780	0.593	0.674	0.725	0.533	0.614	0.771	0.582	0.663
GB	0.790	0.607	0.687	0.794	**0.604**	**0.686**	0.767	**0.560**	**0.648**	0.804	**0.588**	**0.679**

Table 3: Evaluation on Medline titles of Mantra. We compare performance to commercial translation tools (GB) for Spanish, French, Dutch, and German. Methods presented include mono-lingual and cross-lingual candidate search, as well as our and biomedical translation model (BTM) or Google Translate and Bing Translator (GB).

On the dataset of 2015, we see the same pattern. Cross-lingual search improves the performance of the system in terms of recall and F1 in comparison to mono-lingual search and BTM outperforms CL. Moreover, CL and BTM both outperform Erasmus on the Medline dataset. On EMEA instead, our method cannot reach the performance of Erasmus. While recall of BTM and Erasmus are similar, the precision of our system is approximately 10% lower. A reasonable explanation for that might be the fact that Erasmus applied various pre- and post-processing steps optimized for the Quaero copus. In contrast, our method is fully generic. Using a different and more corpus specific disambiguation might improve results for our system as well.

4.3. Evaluation Results for Mantra

We compare performance of our sequential mono- and cross-lingual candidate search for different languages, Spanish, French, Dutch, and German. Additionally, we directly compare performance of our biomedical translation model to commercial translators. The same sequential procedure is applied while instead of translating terms using our translation model we use translations from Google Translate and Bing Translator. Details on how translations with Google Translate and Bing Translator were obtained are described in Section 4.1..

Results presented in Table 3 show the same pattern for all languages as previous results for French on Quaero: Cross-lingual candidate search always outperforms mono-lingual candidate search and the integration of the translator outperforms CL and ML. The main reason for that is the boost of recall which leads, in combination with a good disambiguation, to a high precision and thus to an improved F1.

Cross-lingual candidate search including our biomedical translation model (BTM) outperforms Google Translate and Bing Translator (GB) for Spanish in recall and F1. Overall, precision and recall are very similar for both systems and all languages. Differences in precision are very small for Spanish and French (0.01) and only slightly higher for Dutch and German (0.03-0.04). Differences in recall are in the same range: 0.02 for Spanish (while BTM outperforms GB), 0.01 for French, 0.03 for Dutch, and no significant difference for German.

We also compared performance of BTM and GB between languages. Although differences are small, in terms of recall and F1 the performance of both systems decreases in the following order SPA $>$ FRE $>$ GER $>$ DUT. Interestingly this order correlates well with the number of concepts for each language present in UMLS. In terms of precision, BTM shows the same order, while commercial translators yield best results for German: GER $>$ FRE $>$ SPA $>$ DUT.

5. Conclusions

In this work we present a character-based neural translation model trained on the multi-lingual terminology UMLS for Spanish, French, Dutch, and German. The model is integrated into a sequential candidate search for concept normalization. Evaluation on two different corpora shows that our proposed method significantly improves biomedical concept normalization for non-English texts.

We propose a sequential procedure for candidate search. Subsequently, terms are searched (i) in the relevant Non-English version of UMLS (ML), (ii) in English UMLS without translating the search term (CL), and (iii) translating the term using our biomedical translation model before searching in English UMLS (BTM). Once a matching concept is found the sequence is stopped. We chose this sequential procedure as ML and CL tend to result in a smaller number of false positives (data not shown here).

In all evaluations we found that already a simple cross-lingual candidate search (CL) (without translation) improves recall significantly while at the same time the loss in precision is small. This might be explained by the fact that many biomedical terms are very similar across different languages because they originated from common Greek or Latin words. Therefore, already a fuzzy search is able to detect the right concept for many terms.

We evaluated our system on previous concept normalization tasks using the French Quaero corpus. The sequential candidate search including our biomedical translation model outperformed the winning teams in 3 out of 4 data sets. In this study we focused on a novel approach for candidate search using established methods for disambiguation. Using corpus-specific disambiguation procedures might even further improve precision of our method.

Compared to commercial translation systems our biomedical translation model yields comparable results for French, Dutch, and German, and slightly outperforms on recall and F1 for Spanish. While commercial services require Internet access and are charged although with a low price, our translation model is open-source and free of charge. Additionally, it can be run locally, and therefore be used for processing patient related clinical texts. Such data underly strict data privacy restrictions and are not allowed to be processed using online services. We assume integrating our translation model into NLP pipelines will improve results for non-English biomedical information extraction tasks.

Acknowledgements

This research was supported by the German Federal Ministry of Economics and Energy (BMWi) through the project MACSS (01MD16011F) and the German Federal Ministry of Education and Research (BMBF) through the project PERSONS (031L0030B).

6. Bibliographical References

Afzal, Z., Akhondi, S. A., van Haagen, H., van Mulligen, E. M., and Kors, J. A. (2015). Biomedical Concept Recognition in French Text Using Automatic Translation of English Terms. In *CLEF (Working Notes)*.

Aronson, A. R. and Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Attardi, G., Buzzelli, A., and Sartiano, D. (2013). Machine Translation for Entity Recognition across Languages in Biomedical Documents. In *CLEF (Working Notes)*. Citeseer.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation By Jointly Learning To Align and Translate. *ICLR*.

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Bodnari, A., Névéol, A., Uzuner, Ö., Zweigenbaum, P., and Szolovits, P. (2013). Multilingual Named-Entity Recognition from Parallel Corpora. In *CLEF (Working Notes)*. Citeseer.

Cabot, C., Soualmia, L. F., Dahamna, B., and Darmoni, S. J. (2016). SIBM at CLEF eHealth Evaluation Lab 2016: Extracting Concepts in French Medical Texts with ECMT and CIMIND. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.*, pages 47–60.

Chiaramello, E., Pinciroli, F., Bonalumi, A., Caroli, A., and Tognola, G. (2016). Use of "off-the-shelf" information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes. *Journal of Biomedical Informatics*, 63:22–32.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Hellrich, J. and Hahn, U. (2013). The JULIE LAB MANTRA System for the CLEF-ER 2013 Challenge. In *CLEF (Working Notes)*.

Jiang, J., Guan, Y., and Zhao, C. (2015). WI-ENRE in CLEF eHealth Evaluation Lab 2015: Clinical Named Entity Recognition Based on CRF. In *CLEF (Working Notes)*.

Kingma, D. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. *ICLR*.

Kors, J. A., Clematide, S., Akhondi, S. A., van Mulligen, E. M., and Rebholz-Schuhmann, D. (2015). A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. *Journal of the American Medical Informatics Association*, 22(5):948.

Lee, J., Cho, K., and Hofmann, T. (2016). Fully Character-Level Neural Machine Translation without Explicit Segmentation. *CoRR*, abs/1610.03017.

Moro, A., Raganato, R., and Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: A Unified Approach. *Transactions of the Association for Computational Linguistics*.

Névéol, A., Grouin, C., Leixa, J., Rosset, S., and Zweigenbaum, P. (2014). The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. In *Proc of BioTextMining Work*, pages 24–30.

Neveol, A., Grouin, C., Tannier, X., Hamon, T., Kelly, L., Goeuriot, L., and Zweigenbaum, P. (2015). CLEF eHealth Evaluation Lab 2015 Task 1b: Clinical Named Entity Recognition. In *CLEF (Working Notes)*.

Neveol, A., Goeuriot, L., Kelly, L., Cohen, K., Grouin, C., Hamon, T., Lavergne, T., Rey, G., Robert, A., Tannier, X., et al. (2016). Clinical information extraction at the CLEF eHealth evaluation lab 2016. In *Proceedings of CLEF 2016 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS (September 2016)*.

Rebholz-Schuhmann, D., Clematide, S., Rinaldi, F., Kafkas, S., van Mulligen, E. M., Bui, C., Hellrich, J., Lewin, I., Milward, D., Poprat, M., et al. (2013). 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. Springer.

Van Mulligen, E., Afzal, Z., Akhondi, S. A., Vo, D., and Kors, J. A. (2016). Erasmus MC at CLEF eHealth 2016: Concept recognition and coding in French texts. CLEF.

English-Catalan Neural Machine Translation in the Biomedical Domain through the cascade approach

Marta R. Costa-jussà, Noe Casas and Maite Melero*
TALP Research Center - Universitat Politècnica de Catalunya, Barcelona
* OTG Plan de Tecnologías del Lenguaje
marta.ruiz@upc.edu; contact@noecasas.com; maite.melero@upf.edu

Abstract

This paper describes the methodology followed to build a neural machine translation system in the biomedical domain for the English-Catalan language pair. This task can be considered a low-resourced task from the point of view of the domain and the language pair. To face this task, this paper reports experiments on a cascade pivot strategy through Spanish for the neural machine translation using the English-Spanish SCIELO and Spanish-Catalan El Periódico database. To test the final performance of the system, we have created a new test data set for English-Catalan in the biomedical domain which is freely available on request.

Keywords: Neural Machine Translation, Biomedical, English-Catalan

1. Introduction

Neural machine translation (Sutskever et al., 2014; Cho et al.,) has recently emerged as a stronger alternative to standard statistical paradigm (Koehn et al., 2003). Among other advantages, neural MT offers an end-to-end paradigm which seems to be able to generalize better from data (Bentivogli et al., 2017). However, deep learning techniques face serious difficulties for learning when having limited or low resources and machine translation is not an exception (Koehn and Knowles, 2017).

English has become the *de facto* universal language of communication around the world. In Catalonia, out of 7.5 million population only around 30% of people have knowledge of English in all competences[1]. Therefore, there are many situations where professional or automatic translations are still necessary. One of them is in medical communication patient-physician at the level of primary health care. Also in the biomedical domain it is worth mentioning that Catalonia has become a hub of global biomedical research as proven by the nearly 1% of global scientific production, 9,000 innovative companies or the fact that the sector raised a record of 153 million of euros in 2016 [2]. Therefore, English-Catalan translation in the biomedical domain is of interest not only in health communication but to properly disseminate the work in such a relevant area for Catalan economy.

English-Catalan in general —and even more in a closed domain as the biomedical one—can be considered to be a limited resourced language pair. However, there are quite large amount of resources for English-Spanish and Spanish-Catalan language pairs. Therefore, English-Catalan could take advantage of them by using the popular pivot strategies which consist in using one intermediate language to perform the translation between two other languages.

Pivot strategies have been shown to provide a good performance within the phrase-based framework (Costa-jussà et al., 2012) and also for the particular case of English-Catalan (de Gispert and no, 2006). While in the phrase-based context, pivot strategies have been widely exploited, this is not the case for the neural approaches. Pivot studies are limited to (Cheng et al., 2017) which considers a single direct approach (the cascade) contrasted with a joint trained model from source-to-pivot and pivot-to-source. Other alternatives when having no parallel data for a language pair are the multilingual approximations where several language pairs are trained together and the system is able to learn non-resourced pairs (Wu et al., 2016).

Another related research area for this study is precisely training translation systems domain-specific tasks, where there are scarce in-domain translation resources. A common approach in these cases consists in training a system with a generic corpus and then, use a small in-domain corpus to adapt the system to that particular domain. In this direction, there is a huge amount of research in the statistical approach (Costa-jussà, 2015) and also starting in the neural approach (Chu et al., 2017). Finally, there is am emerging line of research in the topic of unsupervised neural MT (Lample et al., 2018; Artetxe et al., 2018).

This study designs and details an experiment for testing the standard cascade pivot architecture which has been employed in standard statistical machine translation (Costa-jussà et al., 2012).

The system that we propose builds on top of one of the latest neural MT architectures called the Transformer (Vaswani et al., 2017). This architecture is an encoder-decoder structure which uses attention-based mechanisms as an alternative to recurrent neural networks proposed in initial architectures (Sutskever et al., 2014; Cho et al.,). This new architecture has been proven more efficient and better than all previous proposed so far(Vaswani et al., 2017).

2. Neural MT Approach

This section provides a brief high-level explanation of the neural MT approach that we are using as a baseline system, which is one of the strongest systems presented recently (Vaswani et al., 2017), as well as a glance of its differences with other popular neural machine translation architectures.

[1] Data taken from https://www.idescat.cat/
[2] http://cataloniabio.org/ca/publicacions

Table 1: Size of the parallel training corpora

Language Pair	**Corpus**	Language	Segments	Words	Vocab
En-Es	Biomedical	En	$932 \cdot 10^3$	$20,5 \cdot 10^3$	$296 \cdot 10^3$
		Es		$21,9 \cdot 10^3$	$309 \cdot 10^3$
Es-Ca	El Periódico	Es	$6,5 \cdot 10^3$	$16,5 \cdot 10^3$	$736 \cdot 10^3$
		Ca		$17,9 \cdot 10^3$	$713 \cdot 10^3$

Table 2: Size of the test set

Language Pair	**Corpus**	Language	Segments	Words	Vocab
En-Es	Biomedical	En	1000	$26,1 \cdot 10^3$	$6,1 \cdot 10^3$
		Es		$27,4 \cdot 10^3$	$6,6 \cdot 10^3$
Es-Ca	El Periódico	Es	2244	$56 \cdot 10^3$	$12,2 \cdot 10^3$
		Ca		$60,7 \cdot 10^3$	$11,7 \cdot 10^3$

Sequence-to-sequence recurrent models (Sutskever et al., 2014; Cho et al.,) have been the standard approach for neural machine translation, especially since the incorporation of attention mechanisms (Bahdanau et al., 2014; Luong et al., 2015), which enables the system to learn to identify the information which is relevant for producing each word in the translation. Convolutional networks (Gehring et al., 2017) were the second *paradigm* to effectively approach sequence transduction tasks like machine translation.

In this paper we make use of the third *paradigm* for neural machine translation, proposed in (Vaswani et al., 2017), namely the Transformer architecture, which is based on a feed-forward encoder-decoder scheme with attention mechanisms. The type of attention mechanism used in the system, referred to as *multi-head attention*, allows to train several attention modules in parallel, combining also self-attention with standard attention. Self-attention differs from standard attention in the use of the same sentence as input and trains over it allowing to solve issues as coreference resolution. Equations and details about the transformer system can be found in the original paper (Vaswani et al., 2017) and are out of the scope of this paper.

For the definition of the vocabulary to be used as input for the neural network, we used the sub-word mechanism from `tensor2tensor` package, which is similar to Byte-Pair Encoding (BPE) from (Sennrich et al., 2016). For the English-Spanish language pair, two separate 32K sub-word vocabularies where extracted, while for Spanish-Catalan we extracted a single shared 32K sub-word vocabulary for both languages.

3. Pivot Cascade Approach

Standard approaches for making use of pivot language translation in phrase-based systems include the translation cascade. The cascade approach consists in building two translation systems: source-to-pivot and pivot-to-target. In test time, the cascade approach requires two translations. Figure 1 depicts the training of the pivot systems while figure 2 shows how they are combined to devise the final one.

4. Experimental Framework: data resources

This section reports details on data to be employed in the experiments.

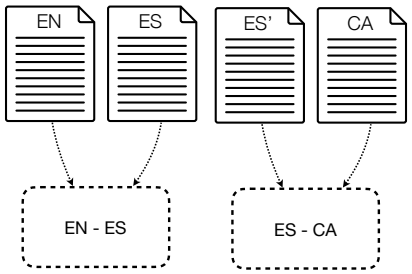


Figure 1: Pivot translation systems training.

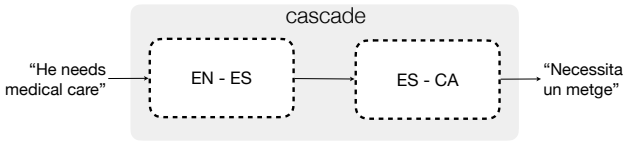


Figure 2: Pivot cascading approach at inference time.

Experiments will be performed for the English-Catalan language pair on the biomedical domain. Resources are the SCIELO database from the WMT 2016 International Evaluation Campaign (Costa-jussà et al., 2016). The English-Spanish corpus is the compilation of the corpora assigned for the WMT 2016 biomedical shared task, gathered from the Scielo database. The Spanish-Catalan corpus is extracted from ten years of the paper edition of a bilingual Catalan newspaper, El Periódico (Costa-jussà et al., 2014). The Spanish-Catalan corpus is partially available via ELDA (Evaluations and Language Resources Distribution Agency) in catalog number ELRA-W0053.

The size of the corpora is summarised in Table 1. The corpora has been pre-processed with a standard pipeline for Catalan, Spanish and English: tokenizing and keeping parallel sentences between 1 and 50 words. Additionally, for English and Spanish we used Freeling (Padró and Stanilovsky, 2012) to tokenize pronouns from verbs (i.e. *preguntándose* to *preguntando* + *se*), we also split prepositions and articles, i.e. *del* to *de* + *el* and *al* to *a* + *el*. This

Table 3: Sample end-to-end English-Catalan translations.

English	Catalan
crustacean diversity and population peaks were within the range of examples found in worldwide literature .	la diversitat de crustacis i els pics poblacionals van estar dins del rang d ' exemples trobat en la literatura mundial .
multivariate analysis and Post-Hoc Bonferroni tests were used and relative risk and attributable fraction were calculated .	es va utilitzar anàlisis multivariat i post-Hoc de Bonferroni i es va calcular el risc relatiu i la fracció atribuïble .
this qualitative study used semi-structured interviews , with eight coordinators of the Tuberculosis Control Program in six cities of the state of Paraíba .	es tracta d ' un estudi qualitatiu amb entrevistes semiestructurades , amb vuit coordinadors del Programa de Control de la Tuberculosi en sis municipis de l ' estat de Paraíba .
there is no statistically significant difference in global and event free survival between the two groups .	no hi ha diferència estadísticament significativa en la supervivència global i lliure d ' esdeveniments entre els dos grups .

was done for similarity to English. For Spanish and Catalan, we used Freeling to tokenize the text but no split with pronouns, prepositions or articles was done. The test sets come from WMT 2016 biomedical shared task in the case of English and Spanish. Since we required a gold standard in English-Catalan, we translated the Spanish test set from WMT 2016 biomedical shared task into Catalan. The translation was performed in two steps: we did a first automatic translation from Spanish to Catalan and then a professional translator postedited the output. This English-Catalan test set on the biomedical domain is freely available on request to authors. Details on the test sets are reported in Table 4.

5. Results

The results of each of the pivotal translation systems as well as the combined cascaded translation are summarized in table 4, which shows the high quality of the translations of the attentional architecture from (Vaswani et al., 2017).
The English-to-Spanish translation obtains a BLEU score of 46.55 in the test set of the WMT Biomedical test set while the Spanish-to-Catalan translation obtains a BLEU score of 86.89 in the El Periódico test set. The cascaded translation achives a BLEU score of 41.38 in the translated WMT Biometical test set.

Table 4: BLEU results.

Language	System	BLEU
EN2ES	Direct	46.55
ES2CA	Direct	86.89
EN2CA	Cascade	41.38

All BLEU scores are case-sensitive and where obtained with script `t2t-bleu` from the tensor2tensor framework, whose results are equivalent to those from `mteval-v14.pl` from the Moses package.

In order to illustrate the quality of the cascaded translations quality, some sample translations are shown in table 3.

6. Conclusion

This paper describes the data resources and architectures to build an English-Catalan neural MT system in the medical domain without English-Catalan parallel resources. This descriptive paper provides details on latest architectures in neural MT based on attention mechanisms and one standard pivot architecture that has been used with the statistical approach. The paper reports results on the baseline system of the cascade approach with the latest neural MT architecture of the Transformer.
Further experiments are required to fully characterize the potential of pivoting approaches. One of the future lines of research is to apply the pseudo-corpus approach, which consists in training a third translation system on a synthetic corpus created by means of the pivotal ones. A second future line of research is the use of recently proposed unsupervised machine translation approaches (Lample et al., 2018; Artetxe et al., 2018), which do not require large amount of parallel data.

7. Acknowledgements

This work is supported by the Spanish Ministerio de Economía y Competitividad and European Regional Development Fund, through the postdoctoral senior grant *Ramón y Cajal*, the contract TEC2015-69266-P (MINECO/FEDER, UE) and the contract PCIN-2017-079 (AEI/MINECO). This work is also supported in part by an Industrial PhD Grant from the Catalan Agency for Management of University and Research Grants (AGAUR) and by the United Language Group (ULG).

8. Bibliographical References

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018). Unsupervised neural machine translation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2017). Neural versus phrase-based mt quality: An in-depth analysis on english–german and english–french. *Computer Speech and Language*.

Cheng, Y., Yang, Q., Liu, Y., Sun, M., and Xu, W. (2017). Joint training for pivot-based neural machine translation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3974–3980.

Cho, K., van Merrienboer, B., and Caglar Gulcehre and Dzmitry Bahdanau and Fethi Bougares and Holger Schwenk and Yoshua Bengio, title = Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, b. . P. p. . . y. . .).

Chu, C., Dabre, R., and Kurohashi, S. (2017). An empirical comparison of simple domain adaptation methods for neural machine translation. *CoRR*, abs/1701.03214.

Costa-jussà, M. R., Henríquez Q, C. A., and Banchs, R. E. (2012). Evaluating indirect strategies for chinese-spanish statistical machine translation. *J. Artif. Int. Res.*, 45(1):761–780, September.

Costa-jussà, M. R., Poch, M., Fonollosa, J. A., Farrús, M., and Mariño, J. B. (2014). A large Spanish-Catalan parallel corpus release for Machine Translation. *Computing and Informatics Journal*, 33.

Costa-jussà, M. R., España Bonet, C., Madhyastha, P., Escolano, C., and Fonollosa, J. A. R. (2016). The talp–upc spanish–english wmt biomedical task: Bilingual embeddings and char-based neural language model rescoring in a phrase-based system. In *Proceedings of the First Conference on Machine Translation*, pages 463–468, Berlin, Germany, August. Association for Computational Linguistics.

Costa-jussà, M. R. (2015). Domain adaptation strategies in statistical machine translation: a brief overview. *Knowledge and Engineering Review*, 3(05):514–520.

de Gispert, A. and no, J. B. M. (2006). Catalan-english statistical machine translation without parallel corpus: bridging through spanish. In *Proceedings of LREC*.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54.

Lample, G., Denoyer, L., and Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Padró, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In *International Conference on language Resources and Evaluation*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems 30*, pages 6000–6010. Curran Associates, Inc.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

KabiTermICD: Nested Term Based Translation of the ICD-10-CM into a Minor Language

Olatz Perez-de-Viñaspre, Maite Oronoz, Natalia Elvira
IXA NLP Group (UPV/EHU), Osakidetza
olatz.perezdevinaspre@ehu.eus, maite.oronoz@ehu.eus, NATALIA.ELVIRAGARCIA@osakidetza.eus

Abstract

In this paper a system called KabiTermICD is presented, which automatically translates ICD-10-CM English terms into a less resourced language, Basque. The lack of a big enough specialized bilingual corpus between Basque and English inspired the creation of KabiTermICD system for the translation of medical terminology. This system is based on the semantic structures that complex terms have, where medical terms are nested into longer terms. The technology used is based on finite state transducers and previously developed multilingual medical lexicons (SNOMED CT, for instance). The results, showed a big time saving for experts to translate ICD-10-CM as around the 60% of the translated descriptions did not need to be post-edited, and most of the remaining needed a smaller post-editing effort.

Keywords: Medical term automatic translation, Finite state transducers, ICD-10-CM, Coding standards

1. Introduction and Background

In this paper KabiTermICD is presented, a system for the automatic translation of the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) from English into Basque.

The main objective of this work is to lighten the translation work of lexicographers by offering them automatically generated accurate terms in Basque. Instead of working on the translations, these professionals would review the automatically generated ones and if necessary, post-edit or correct them.

Basque is an isolated language spoken by more than 700.000 people in the Basque Country. It has not related language, and it coexists with two big languages: Spanish and French. Basque is one of the two official languages in the Basque Autonomous Community, and thus, all the official documents are published both in Spanish and Basque. By law, the Basque Sanitary System called Osakidetza should have its documentation in both languages too. Documents related to the administration are bilingual in Osakidetza but health reports are to be managed distinctively due to our context: Osakidetza has a centralized system where each doctor can have access to all the medical reports about their patients written by different services and health professionals. Not all the doctors in Osakidetza speak Basque language and, as a consequence, to safeguard patient's medical security, electronic medical records are written only in Spanish. In the future, the idea of Osakidetza is to have Electronic Health Records (EHRs) in both languages, among others, to protect patients and doctors linguistic rights while guaranteeing security. We are working in the same direction and collaborating with Osakidetza.

In order to go ahead on with the linguistic normalization of Basque inside the health system (also called language planning), works on establishing Basque medical terminology have been done in the last years. The Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) was already automatically translated from English into Basque (Perez-de Viñaspre and Oronoz, 2014; Perez-de Viñaspre and Oronoz, 2015). Different techniques were used to translate the terminological content of SNOMED CT: 1) the reuse of already developed linguistic resources, 2) finite state transducers[1] that use biomedical affixes to analyze, split, translate and build clinical terms (Perez-de-Viñaspre et al., 2013), 3) transducers that take advantage of medical terms nested into longer terms and, finally 4) the adaptation of a rule-based machine translation system, Matxin, to the medical domain, obtaining MatxinMed. The system developed to implement the third step is called KabiTerm (*Kabi* means "nested" in Basque in reference to the nested terms used for the translation). KabiTermICD is the adaptation of KabiTerm in order to translate the ICD-10-CM classification.

ICD-10 and SNOMED CT have different purposes and this reflects in the construction of some terms. ICD-10 is a very useful classification for statistical recording but the rich semantic structure of SNOMED CT makes it very valuable and adds meaning to the EHRs, besides of being adequate to detail for clinical recording.

The ICD-10, owned and published by the World Health Classification (WHO), replaced in 1999 the previous ICD-9 version with the purpose of coding and classifying mortality data from death certificates[2]. According to WHO, ICD is used by all WHO members (194 Member States), has been translated into 43 languages and in most countries (117) is used to report mortality data, a primary indicator of health status[3]. The clinical modification (ICD-10-CM) improves ICD-10 by the addition of relevant information to ambulatory and managed care encounters; by expanding injury codes; by combinations of diagnosis/symptom codes

[1] "A finite-state transducer (FST) is a finite-state machine with two memory tapes, following the terminology for Turing machines: an input tape and an output tape". In Wikipedia. Retrieved March 5, 2018, from https://en.wikipedia.org/wiki/Finite-state_transducer

[2] https://www.cdc.gov/nchs/icd/icd10cm.htm

[3] http://www.who.int/classifications/icd/en/

to reduce the number of codes to describe a condition, etc. It also increments the number of digits in the codes to allow greater coding specificity (from 3 to 5 digits or from 3 to 7 digits). Each diagnostic term in ICD is linked to one or more numeric codes.

In Spain, by law, from the 1st of January of 2016, all the hospitals must code their diagnostic terms with the CIE-10-ES classification. In fact, this is a translation of the ICD-10-CM classification created and validated by the Spanish Ministry of Health, Social Services and Equality[4]. In 1996 the Health Department of the Basque Autonomous Community translated into Basque the CIE-10 classification, obtaining the GNS-10 classification (GNS is the name in Basque for ICD). Nowadays, the expansion related to the clinical modification is needed, and we are working with Osakidetza, using KabiTermICD to obtain the terms in this extension. The amount of translations needed is still big, as the ICD-10-CM is composed of 93,830 codes, and the ICD-10 in Basque of 12,619. The 1996's ICD-10 version is not updated to the last Basque orthographic rules. In this work, ICD-10-CM is being translated into Basque to obtain the GNS-10-MK translation.

Figure 1 shows part of the ICD-10-CM classification. More specifically, the codes for "viral pneumonia". As we said before, being a classification, the medical terms appearing in it have some characteristics. For example, it gathers terms as "Viral pneumonia, **not elsewhere classified**" (main code J12), "**Other** viral pneumonia" (more specific code J12.8) or "Viral pneumonia, **unspecified**" (code 12.9). The texts "not elsewhere classified", "other" or "unspecified" do not usually appear in, for example, SNOMED CT.

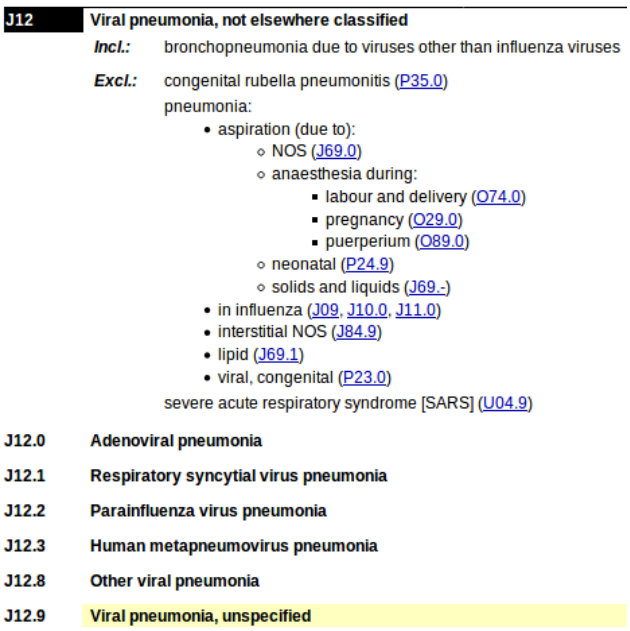


Figure 1: Example of classification of the diagnostic term 'Viral Pneumonia'.

Langlais et al. (2008) define two methods to translate terminology automatically. The "generative" methods where new target words are generated from previously unseen source words (human expertise or machine learning techniques can be used). And, non-generative methods where word translations can be found in parallel corpora (word-alignment methods). As there are not English-Basque bilingual medical corpora, the methods used to translate medical terms from English into Basque are all generative.

Some works about the translation of the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) into different languages have been performed using different techniques: i) The translation into French was done using exclusively automatic translation helping systems (Abdoune et al., 2011), ii) for Chinese, automatic translation and manual work were combined (Zhu et al., 2012), iii) the Danish translation was obtained manually (Petersen, 2011) and iv) three kinds of translations were used to translate SNOMED CT into German (Schulz et al., 2013).

However no work was found in the automatic translation of the ICD medical classification into any language, and much less, into a minority language as Basque, which does not have parallel corpora.

2. KabiTermICD

KabiTermICD is an extension of the KabiTerm system. KabiTerm is based in finite state transducers to obtain Basque equivalents of English terms and it has been implemented using the Foma library and compiler (Hulden, 2009). This system is adequate for the translation in language pairs lacking of bilingual corpus.

The current version of KabiTerm takes a complex English term and, providing the resources available, proposes Basque equivalents. It uses terms that appear within complex terms to translate those complex terms into the Basque language. Here, resources are understood to mean the English-Basque equivalents on the one hand, and the translation patterns on the other. As explained later on in this section, in order to facilitate the work carried out by KabiTerm, an analyser called AnaMed has been developed. This analyser is responsible for obtaining the information required by KabiTerm. It also identifies and prepares the nested terms, leaving KabiTerm free to focus solely on the translation into Basque.

2.1. AnaMed: Medical Term Analyser

In addition to the analysis of linguistic information, AnaMed also identifies SNOMED CT terms and eponyms in a given text. Eponyms are proper nouns that appear in the designation of certain concepts. AnaMed has been developed for English and Basque, and the aim is to adapt it also to the Spanish language in the future.

Initially, only the English version of the AnaMed analyser was developed, in response to our need for a tool to search for the information required by the KabiTerm system outlined in this section. In other words, the information gathered by AnaMed is information that may prove necessary for automatic machine translation from English into Basque. However, since AnaMed can easily be adapted to other languages, it was decided to develop a Basque-language version, since this might prove useful for the drafting of medical reports in Basque.

[4]https://www.msssi.gob.es/en/home.htm

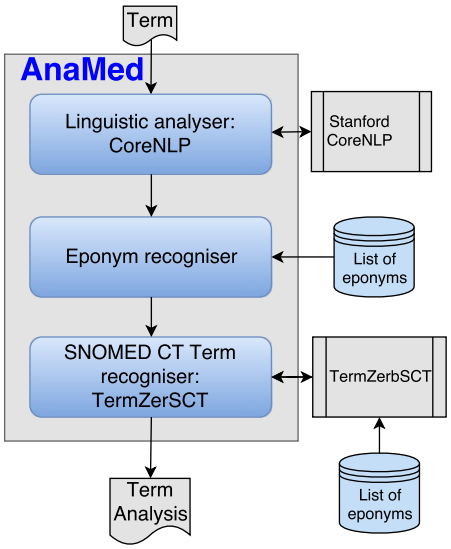


Figure 2: AnaMed analyser architecture.

AnaMed is based on an automatic analysis system and integrates the identification of both eponyms and SNOMED CT terms. The architecture of AnaMed is shown in Figure 2. The following is an explanation of the analysis process using a figure.

- **Linguistic analyser, CoreNLP:** The Stanford CoreNLP tool (Manning et al., 2014) was used as the starting point for the development of the English analyser, along with the Python wrapper for Stanford CoreNLP, developed by Dustin Smith[5]. The tokeniser, the morphological analyser and the tagger from the linguistic analyser were used to identify tokens' lemmas and parts of speech. In addition to this information, token offsets and, named entity tags, were also integrated into the analyser.

- **Eponym recogniser:** By adding a second module the analyser was given eponym identification capability. Eponyms are very common in medical terminology, particularly in the names of diseases and syndromes. The terms "Down syndrome" and "Alzheimer's disease" are good examples of this.

- **SNOMED CT term recogniser:** Finally, the SNOMED CT term identifier was added to the AnaMed analyser. We adapted the TermZerSCT terminology server to identify SNOMED CT terms. SNOMED CT contains a vast amount of terminology (around 300,000 concepts) which takes time to process. TermZerSCT enables faster terminology content management, and when the server is running the reception of information about SNOMED CT terms is almost instantaneous, with a minimum waiting period.

  Using the TermZerSCT server, AnaMed identifies the nested terms located within complex terms, enabling us to analyse the structure of said complex terms. Moreover, it also groups nested terms together using underscores ("_"). For example, in the complex term "unstable diabetes mellitus" it identifies two nested terms: the qualifier "unstable" and the disorder "diabetes mellitus". Thanks to this identification, in addition to providing the complete analysis, AnaMed also gives us the structure (QUALIFIER+DISORDER) and the grouping ("unstable diabetes_mellitus"), information which is extremely useful for KabiTerm.

[5]https://github.com/dasmith/stanford-corenlp-python (accesed May 9, 2017)

2.2. KabiTerm's transducers

KabiTerm's operating process is shown in Figure 3:

1. First of all, AnaMed analyses the input term, identifying and grouping any nested term contained within it. In the case of the term "malignant neoplasm of small intestine, unspecified", "malignant_neoplasm" is a complex term and "small" and "intestine" are simple terms.

2. Secondly, the system calls the transducer responsible for identifying the Basque translation patterns and tagging the nested terms. Thus, this transducer applies the appropriate Basque translation rule and it attaches the tags required to translate the nested terms into Basque. That is to say, the translation rules are applied by means of new tags that imply information to generate the Basque equivalent. In such case, the transducer identifies the structure DISORDER+of+QUALIFIER+BODYSTRUCTURE and applies the corresponding Basque translation rule. It tags "malignant_neoplasm" with "—DIS+a" because it is a disorder and it appends the singular article mark ("+a") required in Basque, it tags "small as a qualifier ("—QUA") and it tags "intestine" with "—BOD+areM" as in addition to being a body part, the term also requires a declension in Basque ("+areM" in this particular case). Besides, the transducer also adds a change of order tag, indicating that the last two terms (excluding everything after the coma) should be moved to the beginning ("&AzkenBiakLehenera"). At this point, there are 72 translation patterns defined.

3. The next step involves rearranging the nested term, following the instructions provided in the tag added during the previous step ("&AzkenBiakLehenera"). Thus, "small" and "intestine" are moved to the beginning as this is the correct place for them in Basque.

4. In the fourth step the system calls up the transducer responsible for translating nested terms into Basque. This transducer provides us with two Basque equivalent terms: *"txiki&&&ADJK heste+areM neoplasia_gaizto+a , zehaztugabea"* (the semantic tags disappear and the output is the Basque equivalent of each English term).

5. Next, rearrangement of adjectives and determination of plurals is executed. There are two kinds of adjectives in the Basque language, those that go after the noun (*izenondo*) and those that go before it (*izenlagun*). In this example, *"txiki"* is an *izenondo* and since

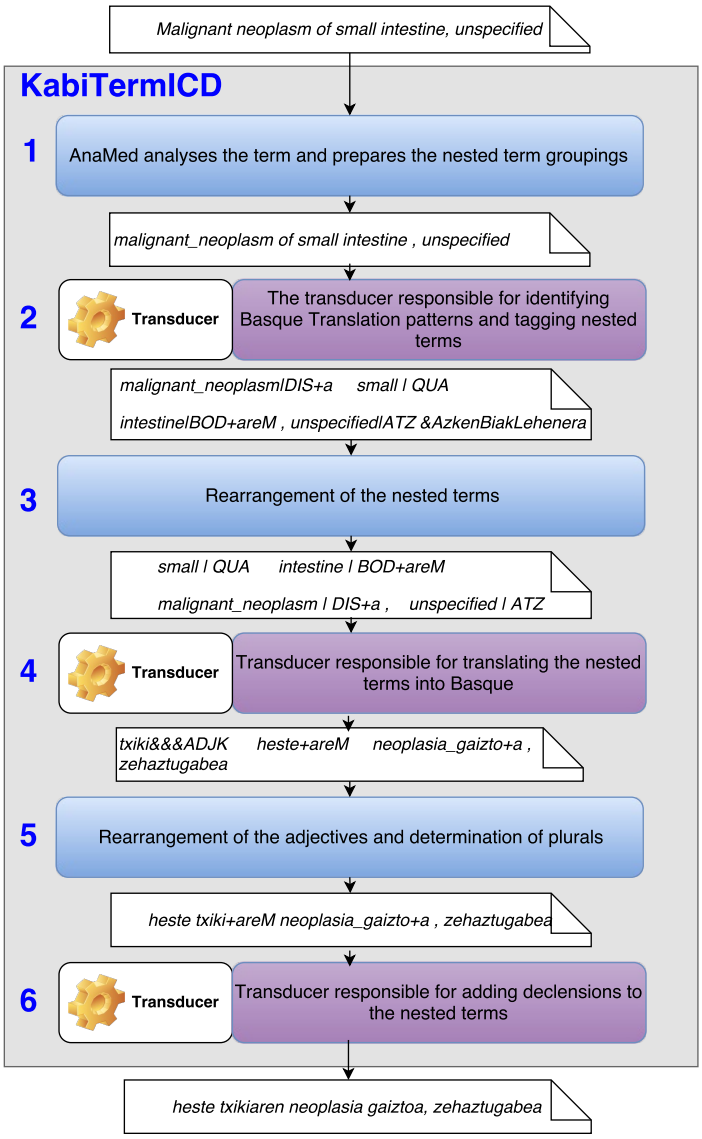


Figure 3: Examples of KabiTermICD's architecture and functioning.

it must go after the name (the &&&ADJK mark expresses that information), said adjective is rearranged after the following term (*"heste"*, translation of "intestine"). Even though this is not the case in the example used here, if the terms contain a nested term in plural in the original English term, in this step its declensions are updated to reflect its plural status.

6. Finally, a new transducer is called up to add the declensions to the nested terms, and thus, it obtains the compound Basque description *"heste txikiaren neoplasia gaiztoa, zehaztugabea"*.

2.3. Translation of ICD-10-CM

As mentioned before, KabiTerm was originally created to translate SNOMED CT's terms into Basque. The characteristics that the descriptions of ICD-10-CM have regarding classification needs, made us include 16 new rules describing the most general structures by now. Being an ongoing process, those structures are still being set and the inclusion of more rules is foreseen in the near future.

The translation of ICD-10-CM was designed as an incremental process that takes advantage of the already translated terms and checked descriptions to generate new ones (Figure 4). For instance, having "malignant" and "neoplasm" terms equivalents in Basque (*"gaizto"* and *"neoplasia"* respectively) KabiTermICD is able to translate "malignant neoplasm" into Basque as *"neoplasia gaizto"*. Similarly, from "lower" and "lip" KabiTermICD generates *"beheko ezpain"* and *"ezpain azpiko"* in Basque. In this case, the expert took *"beheko ezpain"* as the valid translation. This new equivalences may be useful to translate more complex terms, and therefore, the lexicons are enriched with this new equivalences, and the transducers are recompiled. In a second execution of the incremental algorithm (see the second columnn in Figure 4) using the "malignant neoplasm" and "lower lip" nested terms, KabiTermICD is able to translate "malignant neoplasm of lower lip" as *"beheko ezpainaren neoplasia gaizto"*, and in this case the expert post-edited the translation to improve it (*"ezpaineko"* instead of *"ezpainaren"*). Once this new term is included in the lexicons and the transducers recompiled, in subse-

quent applications of the algorithm KabiTermICD translates terms as "malignant neoplasm of lower lip, inner aspect" as shown in Figure 5.

At this point, there are around 100,000 terms included in the lexicons, divided by their corresponding hierarchy, such as disorder, or body structure. Those lexicons were initialized with the term-equivalent pairs generated during the translation of SNOMED CT, and are extended with the new pairs generated during the translation of ICD-10-CM.

As there is not a Gold Standard so the translated descriptions could be evaluated, an expert is validating one by one all the descriptions to create an ICD-10-CM version in Basque. The work was evaluated by comparing the validated descriptions with the ones created automatically. That is to say, the chosen Basque translation is checked whereas it has been automatically generated, or otherwise it has been post-edited by the expert. Validation means accepting or post-editing a description in a process.

For the above mentioned validation, the web page shown in Figure 5 was developed. This web page shows a list of codes that have been translated automatically. Each of the codes is linked with its corresponding hierarchy inside ICD-10-CM offered by the National Cancer Institute's Enterprise Vocabulary Services[6].

In addition to the information regarding the code, the corresponding descriptions in English and Spanish ("Malignant neoplasm of lower lip, inner aspect" and *"Neoplasia maligna de labio inferior, cara interna"* respectively) are also shown, as reference for the validation. Even if the source is the English description, the sociolinguistics context makes Spanish a compulsory reference for the validation. Finally, all the Basque candidates are disclosed so the expert can choose one single candidate as the referenced one (in this case *"beheko ezpaineko neoplasia gaiztoa, barrualdea"*). If any of the candidates is good enough, the expert can edit the text to get the good translation. The web page also allows the expert to let some of the descriptions to be reviewed later.

The validation, and so, the translation, is being made by phases. In the first phase, the translations obtained from dictionaries were loaded, and so a first set of around 6,000 codes was given to the expert. For example, the ICD-10-CM term "Tuberculosis of lung" corresponding to the code A150, was translated by means of the Basque Terminology Bank called Euskalterm, and given its Basque equivalent *"arnas tuberkulosi"*. Once this set of descriptions was corrected, the lexicons of the transducers were updated obtaining the transducers with the corrections included. In this case, the lexicons were not extended but corrected. It is worth to remember, that the lexicons were initially loaded by the automatic translation of SNOMED CT, and so, around 85,000 term-equivalent pairs were already loaded on those lexicons (excluding synonyms). Even if they were included in the initial lexicons, they were not validated by experts, so they need to be uploaded with the corrections made by experts.

In the following phases, the translations obtained by KabiTermICD are loaded to the validation interface. In each phase, a new set of validated translation pairs is included into the lexicons, and in addition, new rules that describe ICD-10-CM descriptions are written to improve the recall of our system.

At this point, 16,512 descriptions have been translated out of 93,830, and from them, 11,104 have been already validated by the expert.

3. Results and discussion

In this section, the focus will be set on the translations validated by the expert. The expert is a translator employed by Osakidetza with an extend background on medical terminology management.

From the 11,104 validated descriptions available nowadays, 5,663 were initially translated by KabiTermICD, and that is the set used to calculate the results.

Around 60% of the descriptions were accepted without any kind of post-editing work (3,379 descriptions). Regarding the remaining 40% with post-editing needs, Table 1 shows the edit distance between the corrected description and the source translation. Edit distance is a measure to quantify the similarity between two words. For that purpose, edit distance counts the minimum number of operations required to transform one word into the other. Even if there are different definitions for the operations, the most common metric is the Levenshtein Distance (Levenshtein, 1966), and the operations are the removal, insertion and substitution of a single character. The edit distances showed in this work were measured according Levenshtein's definition.

Edit distance	Quantity	Percentage
1	109	4.77
2	39	1.71
3	216	9.46
4	986	43.17
5	74	3.24
6	59	2.58
7	63	2.76
8	113	4.95
9	71	3.11
10 - 14	225	9.85
15 - 19	198	8.67
20 - 24	67	2.93
25 - 29	38	1.66
30 - 34	13	0.57
35 - 39	9	0.39
40 - 44	3	0.13
45	1	0.04
Total	2,284	100

Table 1: Edit distance of the post-edited descriptions

After analysing the corrections made to the source translations, correlation between the edit distance and the correction type was found out. In the following lines, the main correction types found on them are listed.

- Edit distance 1: two main mistakes were corrected, spelling errors (row 1 in Table 2) and overproduction

[6]https://nciterms.nci.nih.gov/ncitbrowser/pages/hierarchy.jsf

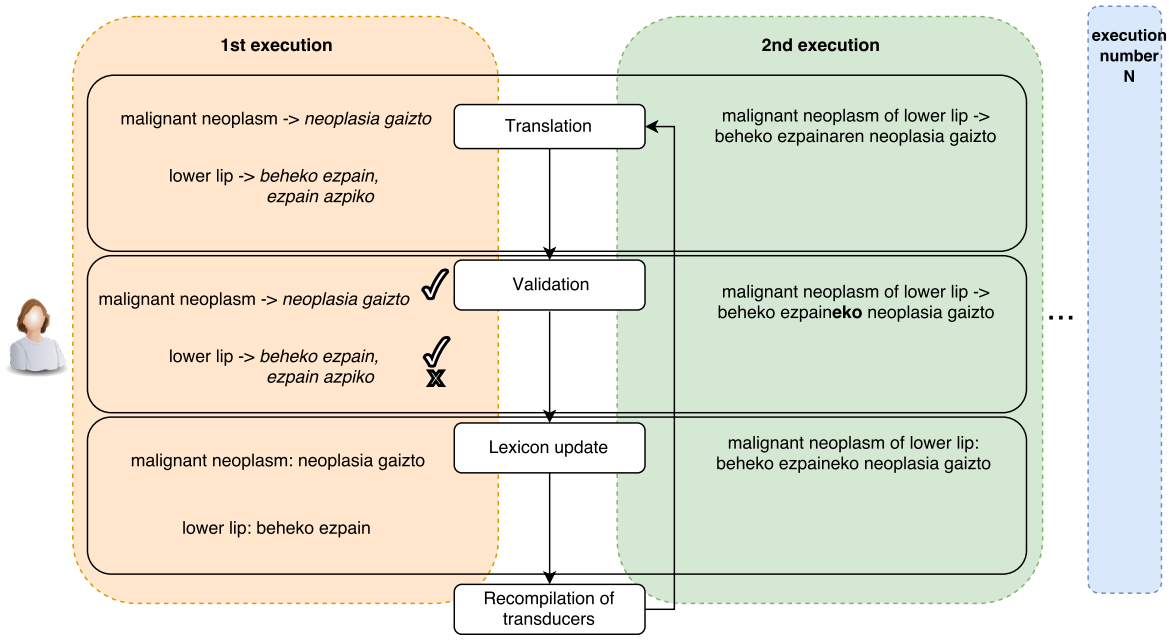


Figure 4: Diagram of the incremental functioning of the translation process.

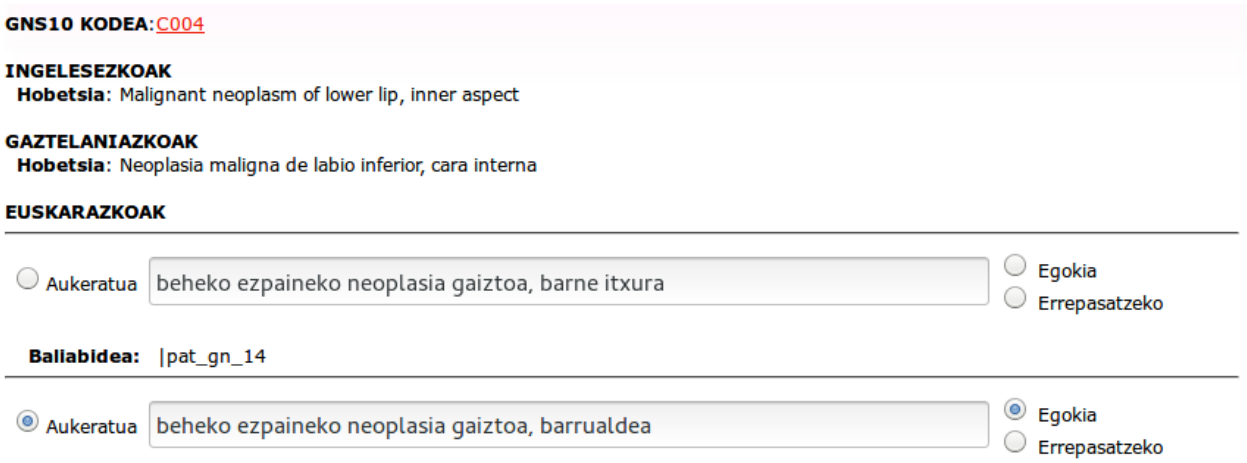


Figure 5: Screenshot of the web page for the validation of ICD-10-CM.

of the absolutive singular declension mark (row 2 in Table 2) when it is not needed.

- Edit distance 2: most of the cases were caused by an unnecessary plural mark in one of the words (row 3 in Table 2).
- Edit distance 3 and 4: in all the analysed cases it was made clear that the automatic translation has an incorrect mark of the genitive. Basque has two types of genitive that are equivalent to English preposition "of" . One is used to describe possession (possessive genitive, "-ren" mark) and the other location (locative genitive, "-ko" mark). By default, KabiTerm uses the possessive genitive to avoid massive overproduction as it is considered more general and in most of the cases it is the best option. The errors found with 3 and 4 edit distance are the descriptions in which the locative genitive should be used instead of the possessive one (row 4 in Table 2).
- Edit distance 5 and higher: From 5 edit distance on, the mistakes found are a combination of previously mentioned errors (row 5 in Table 2), or a bad choice of a nested term (row 6 in Table 2), or a combination of both.

Even if it remains impossible to measure the time saving for the generation of ICD-10-CM, the estimation is made considering the usual time spent on this kind of tasks and the time spent in this case, and the savings are around 4 times better than the time needed to translate the descriptions manually. For instance, the expert would need from 5 to 10 minutes to translate each code without any automatic translation, whereas with automatic translation the whole process takes from 1 to 2 minutes in case the translation is accurate.

4. Conclusions

In this paper, a system for the automatic generation of ICD-10-CM descriptions in a minor language such as Basque is

	Error type	Source	Correction
1	Spelling error	des**c**emetozelea, eskuineko begia	deszemetozelea, eskuineko begia
2	Singular article overproduction	atrofia optiko**a** primarioa, aldebikoa	atrofia optiko primarioa, aldebikoa
3	Incorrect plural mark	beste parafili**ak** batzuk	beste parafili batzuk
4	Incorrect genitive mark	kornea**ren** neoplasia gaizto	kornea**ko** neoplasia gaizto
5	Error combination	orbita**ren** zelulitis**ak**	orbita**ko** zelulitis
6	Incorrect nested term	**urradura**, eskuineko aldaka	**abrasioa**, eskuineko aldaka

Table 2: Examples of post-edited terms.

presented. The system called KabiTermICD is based on nested terms inside those descriptions and using translation patterns generates Basque equivalents.

As far as we know, this is the first work published on the automatic translation of ICD-10, and shows that automatic translation of ICD-10-CM is a promising investment to localise it. The work published here is specially useful for minor languages that can not afford a manual translation and its costs. Being a rule-based system, it that can be adapted to any other language pair, specially to the ones with a corpus not big enough to allow the implementation of corpus-based models. In any case, the lack of other systems makes the results not comparable at this point.

The results show the system to be robust enough, even if it is still open to improvements. Around 60% of correct translations proves a good accuracy in order to translate, and in the cases in which post-editing has been necessary, they were mostly small changes that did not require of big time spent by the expert.

For the future, in order to improve the precision of KabiTerm, the identified errors will be corrected, developing new patterns to improve its recall.

This work showed us that KabiTerm can be easily adapted to translate new medical vocabularies.

Bibliographical References

Abdoune, H., Merabti, T., Darmoni, S. J., and Joubert, M. (2011). Assisting the Translation of the CORE Subset of SNOMED CT Into French. In Anne Moen, et al., editors, *Studies in Health Technology and Informatics*, volume 169, pages 819–823.

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

Langlais, P., Yvon, F., and Zweigenbaum, P. (2008). Analogical translation of medical words in different languages. In *Advances in Natural Language Processing*, pages 284–295. Springer.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Perez-de Viñaspre, O. and Oronoz, M. (2014). Tanslating SNOMED CT Terminology into a Minor Language. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, pages 38–45, Gothenburg, Sweden, April. Association for Computational Linguistics.

Perez-de-Viñaspre, O., Oronoz, M., Agirrezabal, M., and Lersundi, M. (2013). A Finite-State Approach to Translate SNOMED CT Terms into Basque Using Medical Prefixes and Suffixes.

Perez-de Viñaspre, O. and Oronoz, M. (2015). SNOMED CT in a language isolate: an algorithm for a semiautomatic translation. *BMC medical informatics and decision making*, 15(2):S5.

Petersen, P. G. (2011). How to Manage the Translation of a Terminology. Presentation at the IHTSDO October 2011 Conference and Showcase, October.

Schulz, S., Bernhardt-Melischnig, J., Kreuzthaler, M., Daumke, P., and Boeker, M. (2013). Machine vs. Human Translation of SNOMED CT Terms. In C.U. Lehmann et al., editor, *MEDINFO 2013*, pages 581–584.

Zhu, Y., Pan, H., Zhou, L., Zhao, W., Chen, A., Andersen, U., Pan, S., Tian, L., and Lei, J. (2012). Translation and Localization of SNOMED CT in China: A Pilot Study. *Artificial Intelligence in Medicine*, 54(2):147–149, February.

The MeSpEN Resource for English-Spanish Medical Machine Translation and Terminologies: Census of Parallel Corpora, Glossaries and Term Translations

Marta Villegas[1,2], Ander Intxaurrondo[1,2], Aitor Gonzalez-Agirre[1,2], Montserrat Marimon[1], Martin Krallinger[1,2]*

[1]Barcelona Supercomputing Center. Carrer de Jordi Girona, 29, 08034 Barcelona.
[2] Spanish National Cancer Research Center. Calle de Melchor Fernández Almagro, 3, 28029 Madrid
martin.krallinger@gmail.com

Abstract

Clinical and biomedical text mining research efforts have so far focused mainly on documents written in English. These efforts benefited significantly from the availability, not only of domain-specific components such as a tokenizers or Part-of-Speech taggers, but particularly from the access to very large training corpora and terminological resources like UMLS. In order to exploit terminological resources currently restricted to English, it is necessary to promote more systematic translation efforts into other languages, be it manual or by means of machine translation techniques. An initial barrier not only for generating medical machine translation models is the actual identification of relevant datasets that could be exploited to derive glossaries and parallel corpora. Usually relevant datasets weren't constructed as a language technology resource and thus are often overseen by the natural language processing community. This article describes an exhaustive effort to identify and characterize heterogeneous types of documents and glossaries useful to build parallel corpora for Spanish-English medical machine translation systems, including: (1) the combination and harmonization of various bibliographic datasets of biomedical and clinical literature from Spain and Latin America, (2) technical specifications and package leaflets of medicines generated by the pharmaceutical industry, (3) medical and medicinal chemistry patent translations, (4) web-content with trusted information sources about diseases, conditions, and wellness issues for patients, (5) a joined medical multilingual glossary produced by over 500 professional translators and free online medical dictionaries, and (6) keywords derived from bilingual/multilingual medical questionnaires.

Keywords: medical language resources, glossaries, parallel corpora, machine translation, biomedicine

1. Introduction

Currently, a wealth of medical-related information resources do exist in English, not only for patients but also for healthcare professionals, biomedical researches or clinical language technology experts. Such resources include large infrastructures and knowledgebases providing terminologies, biomedical literature or medically related content produced or consumed by either patients or medical professionals. Not surprisingly, this scenario motivated the development of computational tools to improve access, processing and automatic extraction of relevant information by means of natural language processing techniques (Meystre et al., 2008, Krallinger et al., 2008), together with the development of Gold Standard corpora and associated shared tasks and evaluation campaigns (Neves et al., 2014).

The use of specialized medical machine translations techniques may represent a more systematic alternative to generate translations, not only of medical terms, but also of medically related natural language content in general. Efficient medical machine translation systems would be useful not only for generating term translations, but also to assist medical translators and interpreter services in their labor, while patients and healthcare professionals could make better use of clinical information locked in documents subjected to language barriers. Practical adoption of medical machine translation systems could ultimately result in potential improvements in patient safety, diagnostic aid, integration of multilingual clinical information sources and cross-language detection of cases of rare diseases. Moreover, errors in medical interpretation could potentially be reduced by means of medical machine translation assistance (Flores et al., 2003).

Parallel and comparable corpora are a key resource for the development of state-of-the-art corpus-based machine translation (MT), like statistical MT and example-based MT. However, MT systems trained on general-domain data perform poorly in the biomedical domain (Zeng-Treitler et al., 2010), and more recent works use biomedical corpus, sometimes combined with non-medical corpora, to produce higher quality translations (Wu et al., 2011;Yepes et al., 2013; Lui et al., 2015; Neves et al., 2016).

Here we present the MeSpEN (Medical Spanish English) Resource, to our knowledge the first attempt to systematically characterize heterogeneous, complementary, resources and relevant datasets for implementing medical English-Spanish machine translation technologies. The aim of MeSpEN was not to construct the MT technology itself, but to compile medically related bilingual datasets covering different scopes, end users and content types, including bibliographic databases (PubMed[1], Scielo[2] and IBECS[3] along with certified patient-oriented web-content like MedlinePlus[4]).

Figure 1 provides a general overview of the type of resources examined within MeSpEN, which will be detailed throughout this manuscript.

MeSpEN was constructed as part of a framework of the Plan de Impulso de las Tecnologías del Lenguaje de la Agenda Digital (PlanTL) (Plan for Promotion of Language

[1] https://www.ncbi.nlm.nih.gov/pubmed/
[2] http://scielo.org/php/index.php?lang=en
[3] http://ibecs.isciii.es
[4] https://medlineplus.gov/

Technologies), launched by Spanish Ministry of State for Telecommunications with the aim of providing specialized technical support to research and development of software solutions adapted to the field of biomedicine (Villegas et al., 2017). These resources are publicly available for download at http://temu.bsc.es/mespen.

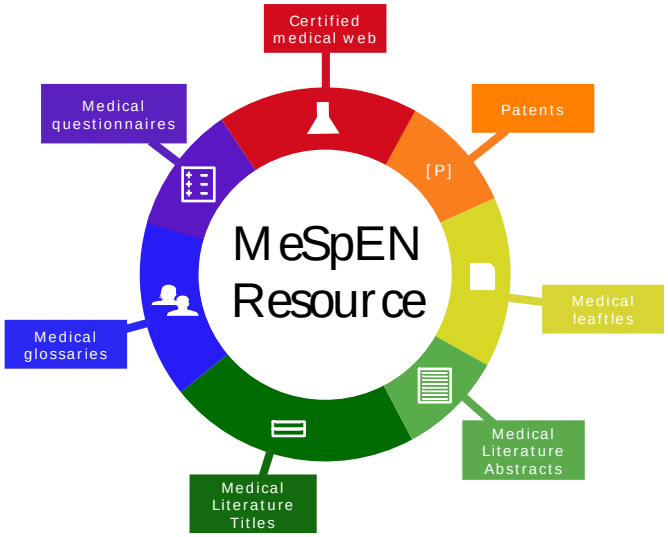


Figure 1: MedSpEN content overview.

2. Related Work

Several efforts were made to generate parallel biomedical corpora and to train statistical medical MT systems. A valuable resource for generating parallel biomedical corpora are medical publications, as it is common to provide, in addition to the full text medical article in any particular language, both the article title as well as its abstracts additionally in English. Although there is a general tendency to directly publish the entire manuscript in English, in case of Spanish medical articles, a growing number of publications published every year can be observed (see figure 2).

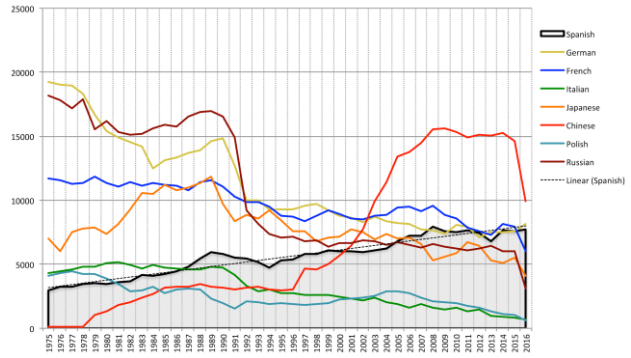


Figure 2: New records/year added to PubMed for non-English articles.

corpus from the foreign language titles (French, Spanish, German, Hungarian, Turkish and Polish) and their corresponding human English translations of Medline/PubMed articles. Yepes et al. (2013) obtained a parallel corpus of article titles from MEDLINE and abstract texts automatically retrieved from journal websites while Neves et al. (2016) presented a parallel corpus of biomedical titles and abstracts from the SciELO database in three pair languages: French/English, Portuguese/English and Spanish/English. Other efforts explored the use of EMEA (Eurpoean Medicines Agency) documents to derive multilingual parallel corpora (Tiedemann 2009) resulting in the OPUS corpus for machine translation. The access to biomedical parallel corpora also motivated carrying out machine translation shared tasks such as the WMT'16 Biomedical translation task (Yeped et al., 2017) and the CLEF-ER (Rebholz-Schuhmann et al., 2013).

3. Methodology and datasets

In order to construct the MeSpEN resource several datasets relevant for Spanish-English medical MT were retrieved, preprocessed and analyzed. A detailed description of each resource can be found in this section.

3.1 Biomedical and clinical literature

The MeSpEN resource includes, among others, bilingual (Spanish and English) data taken from different Biomedical and clinical literature sources, namely: IBECS, SciELO and PubMed. This section describes briefly the way these source data were collected and harmonized into the Dublin Core format

3.1.1 IBECS

The *Índice Bibliográfico Español en Ciencias de la Salud* (IBECS) (Spanish Bibliographic Index in Health Sciences) is a bibliographic database, maintained by the Biblioteca Nacional de Ciencias de la Salud (BNCS, National Health Sciences Library) of the *Instituto de Salud Carlos III* (Carlos III Health Institute) that collects scientific journals covering multiple fields in health sciences (including medicine, pharmacology, nursing, psychology, odontology and physiotherapy, among others) published in Spain from year 2000 onward.

Currently, IBECS includes mainly journals with monolingual content in Spanish[5]: 168,198 records, with an annual increase of more than 12,000 records. Database updates are made weekly and are freely accessible online. IBECS also includes 28,919 links to complete articles through SciELO Spain node[6]. To generate the IBECS subset of the MeSpEN resource we directly collaborated with the Carlos III Health Institute to obtain an XML file including all metadata records in IBECS in December 2017. The file was encoded following the LILACS model, a Virtual Health Library component and is composed by standards, manuals and software, which guide the identification, selection, bibliographic description, document indexing and databases development. The model is widely used in Latin American and the Caribbean countries for health documents indexation.

The mapping from LILACS into Dublin Core was quite straightforward. The following table shows the

[5] http://ibecs.isciii.es/iah/online/E/help/revistas.pdf

[6] http://scielo.isciii.es

correspondences between both models as shown in figure 3.

For some records, however, the language of the title and/or abstract was not specified in the original data. In these cases, we applied a simple language detection algorithm and included this information in the resulting records. Essentially, the system counts the occurrences of Spanish and English stop words in the title and abstract of a given record to discriminate between languages. In case of tie, the system checks if the texts have Spanish characters like accents (á, é,...) or inverted question and exclamation marks ("¿" and "¡").

ab_es	*<dc:description xml:lang="es">*
ab_en	*<dc:description xml:lang="en">*
au	*<dc:creator>*
id	*<dc:identifier>*
ti_es	*<dc:title xml:lang="es">*
ti_en	*<dc:title xml:lang="en">*
is	*<setSpec>*
type	*<dc:type>*
da	*<dc:date>*
mh	*<dc:subject>* (DeCS codes are kept here)
pu	*<dc:publisher>*
fo	*<dc:source>*
la	*<dc:language>*

Figure 3: LILACS Dublin Core mapping.

Initially the corpus had 9.699 titles and 10.252 abstracts with undefined language. After applying the language identification algorithm this number dropped to 487 titles and 62 abstracts. The java program used to convert MeSpEN-IBECS records into Dublin Core can be found in the MeSpEN website (http://temu.bsc.es/mespen).

3.1.2 SCIELO

The Scientific Electronic Library Online (SciELO) is an electronic library supported by the Sao Paulo Research Foundation (FAPESP) and the Brazilian National Council for Scientific and Technological Development (BIREME). This initiative gathers electronic publications of complete full text articles from scientific journals of Latin America, South Africa and Spain.

Currently, SciELO Network is present in 15 countries. Each country operates a collection of open access journals that are published nationally by scientific societies and associations, academic and research related institutions. The network operates over one thousand journals which publishes about 50 thousand articles per year and accumulate more than 700 thousand articles. The journals belong to different disciplines. The SciELO contents are used by different users, specially academic related communities, including students, teachers and researchers as well as professionals and the general public. During the first semester of 2017 the SciELO Network collections served a daily average of more than 1.5 million download. SciELO Spain is supported by the Spanish National Health Sciences Library (Biblioteca Nacional de Ciencias de la Salud, BNCS) .

Most SciELO nodes are available through their OAI-PMH servers and this allows the usage of standard metadata harvesting methods to collect the records. The Argentina, Chile, Peru and Venezuela nodes do not have OAI-PMH nodes and, therefore, they were not included in the corpus. Table 1 shows the number of journals currently available in SciELO for each country together with the number of articles and the number of articles written in Spanish

Although the web pages of the full text articles do have the corresponding translations into English, these translations are not included in the metadata records of the OAI-PMH servers. Thus, the source SciELO's records only contain titles and abstracts in the original language of the article. In order to collect the translations, we used a script that (i) gets the URL in the dc:identifier field of the metadata records pointing to the HTML page of the full article, (ii) retrieves the XML version and, finally, (iii) gets the translated title and abstract7.

With this approach, the MeSpEN corpus extends the original Dublin Core metadata records with translated titles and abstracts. This enrichment implied adding xml:lang attributes in the original records.

Country	Journals	Publications	Publications in Spanish
Bolivia	23	4,728	4,568
Brazil	396	2,554	73
Colombia	228	53,987	50,783
Costa Rica	37	8,510	6,894
Cuba	70	32,702	31,745
Mexico	200	54,728	45,074
Paraguay	14	1,723	1,657
Portugal	70	10,704	482
South Africa	69	23,565	7
Spain	60	34,820	28,919
Uruguay	27	4,016	3,754
Total	**1,194**	**210,205**	**173,956**

Table 1: Publications in the SciELO network

Note however that about 12,000 out of the 171,000 XML files with full articles had some codification error and were not processed. This explains why the eventual MeSpEN-SciELO resource is smaller that the complete SciELO collection. Table 2 shows the number of current publications in the MeSpEN corpus by country:

Country	Publications in Spanish
Bolivia	4,034
Brazil	39
Colombia	45,658
Costa Rica	6,871
Cuba	25,694
Mexico	44,807
Paraguay	1,648
Portugal	1,648
South Africa	7
Spain	28,843
Uruguay	3,639
Total	**161,710**

[7] The html pages include a link to their XML version where titles and abstracts are in both the original language and English (when available) and the full text in the original language.

3.1.3 PubMed

PubMed is a free search engine used to access Medline database, a bibliographical database of references and

Table 2: the number of publications the MeSpEN in Spanish.

abstracts on life sciences and biomedical topics. It is maintained by the U.S. National Library of Medicine. It contains more than 28 million publications up to 39 languages, including Spanish. The online PubMed search does not display directly non-English content, but it is provided in the actual XML record.

The Medline database stores more 330,000 records for articles published originally in Spanish. Most of the these records do additionally also provide title of the original publication in Spanish. It is possible to retrieve the abstract for at least 127,619 of the records.

PubMed allows downloading search results in XML format using the 'send to' option. We used this functionality to get all XML records in Spanish. These records were easily converted into Dublin Core as follows

<*OtherAbstract Language="spa"*>	<*dc:description xml:lang="es"*>
<*Abstract*>	<*dc:description xml:lang="en"*>
<*Author*>	<*dc:creator*>
<*PMID*>	<*dc:identifier*>
<*VernacularTitle*>	<*dc:title xml:lang="en"*>
<*ArticleTitle*>	<*dc:title xml:lang="es"*>
<*ISSN*>	<*setSpec*>
<*PublicationType*>	<*dc:date*>
<*MedlineDate*>	<*dc:date*>
<*Keyword*>	<*dc:subject*>
<*Title*> (inside tag <*Journal*>)	<*dc:publisher*>
<*Language*>	<*dc:language*>

3.2 Trusted web-content for patient information: MedlinePlus

MedlinePlus is an online information service provided by the U.S. National Library of Medicine (NLM), and gives free information about health in both English and Spanish. MedlinePlus provides the following encyclopedic information:

- **health topics[8]:** summaries about disorders, therapies and body locations, among others.
- **drugs and supplements[9]:** prescription drugs, over-the-counter medicines, dietary supplements and herbal remedies, side effects, dosage and drug interactions.
- **laboratory test information[10]:** information about what the test is used for, or why doctors order it, among other**.**
- **medical encyclopedia11:** articles about diseases, tests, symptoms, injuries and surgeries. This is similar to health topics but with more extended and elaborated content

The current version of the MeSpEN-MedlinePlus corpus only includes the health topics part because it is the only one that includes Dublin Core metadata, while additional content will be added to future releases of MeSpEN

3.2.1 3.2.1 Health topics

MedlinePlus includes 1,063 documents on health topic. The structure of these documents is quite simple: they contain a summary of the topic with links to related sites that provide more information to the user about illnesses' diagnosis, risk factors, treatments, etc. The summary is available in the English and Spanish versions of MedlinePlus.

The source code of the site stores metadata information about each topic. Metadata elements (meta) are used to include Dublin Core labels and this makes the conversion to Dublin Core records simple and fast.

To create the Dublin Core records, we checked the metadata in the English and Spanish versions of the topic, and joined them. Most of the information in the metadata has its translation to Spanish in the Spanish site. For instance, the title and keywords are always translated, but relation names and MeSH terms always remain untranslated.

3.2.2 MedlinePlus pages

For all the topics of the library, we used the TEI12 standard to create the parallel corpus. This standard is widely used to digitalize long texts in the academic field or digital humanities. The lack of Dublin Core data, and the high amount of information found in these articles motivated the use of this standard.

The corpus is composed by four different files per topic, instead of a unified one like in the previous resources:

- Clean raw text in English.
- TEI file with text in English.
- Clean raw text in Spanish.
- TEI file with text in Spanish.

The clean raw text is structured by sections and paragraphs. At the beginning of each line, a code indicates if the line belongs to a section title, section subtitle, paragraph or listed text. These files were created after extracting the complete text from the HTML source code of each article. TEI files contain the same text content of the raw files, but structured in XML format. Each section is divided in subsection, paragraph and lists, following the TEI schema. This collection contains a total of 6,292 articles.

3.2.3 Incoherences in the MedlinePlus corpus

To check the quality of the corpus, we analyzed all articles of the library, comparing each article in English with their corresponding Spanish version. Unfortunately, this corpus is not totally parallel: there are situations where some sections are missing in one language, or a paragraph is splitted in two or more paragraphs in the other language, and occasionally the title of a new subsection is marked as paragraph in the other language. Table 3 shows the number of section titles, paragraphs, subsection titles and list elements found for each language, and the number of the documents where we can find these issues.

[8] Link: https://medlineplus.gov/healthtopics.html
[9] Link: https://medlineplus.gov/druginformation.html
[10] Link: https://medlineplus.gov/labtests.html
[11] Link: https://medlineplus.gov/encyclopedia.html
[12] Link: http://www.tei-c.org/index.xml

	English	Spanish	Incoherent articles
Total articles	6,292	6,929	-
Sections found	57,337	57,293	120
Paragraphs found	121,954	125,827	4,248
Subsections found	5,258	5,288	38
List items found	179,059	178,721	532

Table 3: MeSpEN-MedlinePlus subset document structure statistics. The last column shows the number of records with alignment issues.

3.3 EMEA corpus

The OPUS - EMEA corpus (Tiedemann, 2009) is a corpus of biomedical documents retrieved from the European Medicines Agency (EMEA). The corpus includes documents related to medicinal products and their translations into 22 official languages of the European Union. It contains roughly 1,500 documents for most of the languages. In particular, the English-Spanish language pair is composed of 1,667 documents, 998,015 sentences and 13,818,929 tokens.

The original EMEA corpus has been compiled out of PDF documents available online. After downloading these documents they were converted to text and sentence aligned. However, the authors did not evaluate the quality of the alignment, and an inspection of the corpus reveals that, unfortunately, the quality of the alignment is not very good. The corpus including all sentence alignments is available at: http://opus.nlpl.eu/download.php?f=EMEA.tar.gz

3.4 COPPA corpus

The COPPA corpus seeks to help users and researchers to overcome the language barrier when searching patents published in different languages and to stimulate research in Machine Translation and in language tools for patent texts. The segments included in the corpus are obtained by aligning the sentences of the abstracts and titles of published PCT applications with their translations, the translations having been produced by professional patent translators. The parallel corpus contains 18.303 documents, 62,057 sentences, 2,328,713 tokens and 14,624,745 characters for the English-Spanish language pair. The corpus is available for free for research purposes and for a nominal fee for other purposes, order form and details are available at: http://www.wipo.int/patentscope/en/data/products.html#coppa

3.5 Bilingual glossaries

Hand crafted glossaries are a particularly valuable resource for the medical translator community and have shown to boost performance of MT systems. We generated 46 bilingual glossaries for various language pairs from free online medical glossaries and dictionaries made by over 500 professional translators. Glossaries were encoded in standard tab-separated values (tsv) format and 26 include English terms, 8 include Spanish terms and 13 files include other languages.

Table 7 shows the number of entries each glossary contains. As can be observed, the largest glossary is the English-Spanish one, with 123,788 terms, followed by English-Korean, with 69,368 terms and Chinese-English, with 66,939 terms.

To evaluate the quality of the glossaries we used a cTakes pipeline to identify UMLS concepts that appear in the glossaries. For time constrictions, we have only been able to process a randomly selected subset of 2% of the glossary in English. In this subset we identified 2,340 CUIs, of which 1,485 are unique (not repeated). We can estimate that in total, for the 100,245 unique terms in the English glossary, we will have about 116,000 UMLS concepts, although it is hard to anticipate how many of them will be unique (the multiplication gives 74,250, but as we increase the number of concepts it gets more difficult to get new/different ones).

Language pair	Frequency	Language pair	Frequency	Language pair	Frequency
English-Spanish	123,788	Latin-Russian	2,486	German-Russian	225
English-Korean	69,368	German-Swedish	2,208	English-Romanian	205
Chinese-English	66,939	German-Portuguese	2,028	Italian-Spanish	196
English-Italian	24,155	English-Swedish	1,067	Danish-English	193
English-German	18,534	German-Italian	976	French-Spanish	179
English-Japanese	18,320	English-Slovenian	945	Danish-Polish	166
Arabic-English	9,384	Bengali-English	841	Russian-Spanish	122
English-Turkish	7,675	English-Thai	835	English-Hindy	120
German-Spanish	7,004	Dutch-French	585	French-German	119
Dutch-English	6,878	English-Indonesian	491	French-Italian	117
English-French	6,571	Bulgarian-English	347	Croatian-German	115
English-Russian	4,346	Croatian-English	339	German-Romanian	109
English-Polish	3,727	Polish-Spanish	271	Dutch-Spanish	70
English-Hungarian	2,711	Dutch-Turkish	238	Portuguese-Spanish	61
English-Greek	2,626	Latin-Polish	237	English-Norwegian	44
English-Portuguese	2,517	Croatian-French	235	**TOTAL**	**390,713**

Table 4: Number of entries in bilingual glossaries

3.6 Keywords derived from bilingual/multilingual medical questionnaires

The MDM-Portal (Medical Data Models[13]) is a metadata registry for creating, analyzing, sharing and reusing medical forms. It contains forms with more than 350,000 data-elements and numerous core data sets, common data elements or data standards, code lists and value sets. Some of the source forms in the system include translations in other languages than English and constitute a potentially interesting multilingual resource. We have directly requested the support of language subset selection by teh MDM-Portal, which is now supported. The translations of concepts in MDM are provided by humans and sometimes they include a reference to the relevant UMLS CUI:

```
<ItemDef OID="I.101" Name="Toxicity" …
    <TranslatedText xml:lang="es">Toxicidad</
   <Alias Context="UMLS CUI" Name="C0013221"/>
</ItemDef>
```

Note however that there are only 199 data-element translated into Spanish and, in some cases, the translations are rather odd:

```
<ItemDef OID="I.104" Name="Comment" …
  <TranslatedText xml:lang="es">Comentarios, notas</
  <Alias Context="UMLS CUI [1]" Name="C0947611"/>
```

[13] https://medical-data-models.org/

4. MeSpEN statistics

In this section we show different statistics of the created corpus, and also classifying them by source.

Table 5 displays the total number of MeSpEN literature subsets entries together with source information, and how many entries had titles and abstracts in different languages. This table also shows how many publications could be detected having titles and abstracts in both languages. In general, we can find more parallel titles in English and Spanish, while parallel abstracts were less frequent. It is important to stress that PubMed records generally lack parallel abstracts, as the number of abstracts in Spanish is minimal.

Table 6 shows the number of words of titles and abstracts from different corpora by language; the average number of words is also shown. We used the IXA pipeline (Agerri et al, 2017) tokenizer to detect words prior to count the their amounts in each title and abstract. We can find the word averages of IBECS, SciELO and PubMed to be alike between them, finding a similar nature with MedlinePlus abstracts. Meanwhile, the difference between MedlinePlus titles and other sources is bigger, which provides information about the different nature of this source and the other three; all articles in MedlinePlus use health terms as titles.

Collection	IBECS	SciELO	PubMed	MedlinePlus
Total publications	168,198	161,710	330,928	1,063
Publications with titles in English and Spanish	114,798	119,820	300,690	1,018
Publications with abstracts in English and Spanish	100,824	119,722	4,147	1,063
Publications with titles and abstracts in English and Spanish	98,132	103,684	3,971	1,018
Number of titles in Spanish	155,094	158,600	300,690	1,018
Number of titles in English	166,469	120,913	330,928	1,018
Number of abstracts in Spanish	149,742	121,774	4,340	1,063
Number of abstracts in English	118,972	120,546	125,703	1,063

Table 5: Parallel corpus obtained from IBECS, SciELO, PubMed and MedlinePlus.

Collection	IBECS	SciELO	PubMed	MedlinePlus
Number of words in Spanish titles	149,742	2,191,228	3,686,570	3,001
Word average in Spanish titles	12	13	12	2
Number of words in English titles	118,972	1,539,469	4,338,153	2,178
Word average in English titles	10	12	13	2
Number of words in Spanish abstracts	23,499,026	23,978,427	1,023,621	214,154
Word average in Spanish abstracts	156	196	235	201
Number of words in English abstracts	18,934,750	21,883,710	25,306,325	198,839
Word average in English abstracts	159	181	201	187
Unique words in Spanish publications	184,936	159,997	20,942	5,099
Unique words in English publications	163,141	172,808	158,032	3,810

Table 6: Number of words and word average in titles and abstracts.

Table 7 the number of words, sentences, and their averages per document. Looking at the averages, we can find that these statistics are very similar for both the English and the Spanish collection.

Library	Health topics	Medications	Natural supplements	Laboratory tests	Patient instructions	Encyclopedia articles
Total articles	1063	1394	100	50	873	3553
Total words in Spanish	219,979	2,570,505	877,280	74,850	842,707	3,116,360
Average words in Spanish	103	921	4,386	748	482	438
Total words in English	204,013	2,318,356	856,359	71,498	773,554	2,815,448
Average words in English	95	831	117	714	443	396
Total sentences in Spanish	8,947	93,061	23,231	2,641	47,949	198,333
Average sentences in Spanish	4	33	116	26	27	27
Total sentences in English	9,136	93,392	23,532	2,642	47,037	194,902
Average sentences in English	4	33	117	26	26	27
Unique words in Spanish	5,206	8,455	48,150	2,124	13,699	25,991
Unique words in English	3,943	5,082	46,129	1,767	9,738	19,994

Table 7: Number of words, word averages and unique words in MedlinePlus documents.

4.1 Explorative use of MeSpEN

One straightforward use of a parallel corpus is to employ it for the enrichment of terminological resources that are less developed in one of the languages. We propose to use the MeSpEN with the objective of enriching UMLS (https://www.nlm.nih.gov/research/umls/) automatically, that is generating candidate term pairs and medical vocabulary in Spanish, both suggesting completely new terms or synonyms of already existing terms in the original UMLS terminological resource.

To achieve this goal we tested titles of PubMed publications for which we had a title in English and its corresponding translated title in Spanish (total 298,040 pairs of titles). Our strategie comprised the following steps:

1. Identify UMLS terms in English titles using cTakes (http://ctakes.apache.org/).
2. Align the words of the titles in English to the words of the titles in Spanish.
3. Using the previous alignment we detected the terms in the titles in Spanish, and we assigned them to their corresponding candidate terms in English.

In the first step, to identify the UMLS terms in the English titles we will use a cTakes pipeline that queries UMLS. This pipeline identified the UMLS terms in the unstructured text and assigned to them the most likely Concept Unique Identifier (CUI). Thus, once the titles in English were processed, we were able to extract UMLS terms in English that appear in each title (see Figure 3).

6 years of the International Union of Societies of Immunology. Presidential report (Brighton 1974)
Correlation between the status of the newborn infant and maternal dissociative anesthesia
Adrenergic beta receptor blockaders in arterial hypertension
Septic shock. II. Therapy
Treatment of acute schizophrenia with sulforidazine
Panarteritis nodosa associated with positive Australia antigen findings

Figure 3: UMLS terms identified by the cTakes Clinical pipeline.

In the second step, we used our PubMed titles in English and Spanish as the parallel corpus. Using this parallel corpus we trained a word alignment model using GIZA++. After this process, we obtained the words in English aligned to their correspondences in Spanish. For example, for the titles "*Adregenetic beta receptor blockaders in arterial hypertension*" and "*Drogas betabloqueantes en hipertensión arterial*" we obtain the alignment shown in Figure 4.

Sentence pair (1955197) source length 7 target length 5 alignment score : 3.89774e-08
Drogas betabloqueantes en hipertensión arterial
NULL ({ }) Adrenergic ({ 1 }) beta ({ }) receptor ({ }) blockaders ({ 2 }) in ({ 3 }) arterial ({ 5 }) hypertension ({ 4 })

Figure 4: Resulting alignment between "Adregenetic beta receptor blockaders in arterial hypertension" and "Drogas betabloqueantes en hipertensión arterial" using GIZA++.

In the last step, we used the UMLS terms in English detected in the first step, along with the alignments detected in the second step, to predict the candidate text span containing the potential translated terms in Spanish. Once the terms in Spanish were detected, they were extracted and assigned to the same UMLS concept identifier (CUI) as their counterpart in English. In the case of the previous example, the spans detected would be the ones shown in Figure 5. As we can see, using the spans detected by the alignment we can align "*Adregenetic beta receptor blockaders*" to "*Drogas betabloqueantes*" and "*arterial hypertension*" to "*hipertensión arterial*".

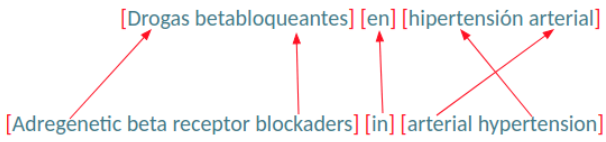


Figure 5: Final translations from English to Spanish using PubMed titles.

A sample set of 200 candidate terms were manually validated by a domain expert. 47% were correct translations, 22% corresponded to either a more general term or more narrow term (hypernym/hyponym) and the remaining pairs were either substrings of the correct translation or wrong translations. The average validation time per term was of just 2.03 seconds, using the MyMiner (Salgado et al., 2012) annotation tool.

5. Discussion

We present a resource for machine translation that is unique in the sense of integrating heterogeneous types of resources for medical machine translation, and harmonizing all the medical literature resources to a common standardized format. Our corpus is composed of publications from three different sources. Right now not OAI-PMH servers work for the SciELO network. It also covers variants of medical Spanish, as it comprises resources from several countries including Spain and Latin American countries such as Argentina, Chile and Venezuela. Other co-official languages in Spain that is Basque, Catalan and Galician are currently not well covered in our resource. Note that we could only find very few publications in Catalan in SciELO (total of 8) and PubMed (total of 88). We are currently analyzing additional information sources to better cover parallel corpora for Galician and Catalan. Moreover, we also explore to directly derive medical glossaries from UMLS for Spanish and Basque. One additional aspect that will deserve a more in depth exploration is to actually compare the lexical characteristics and mentioned UMLS concepts across the various resources that we have gathered to characterize differences in the more formal, scientific language of medical publications, attributes of intellectual property texts found in medicinal chemistry patents and of language expressions in documents whose primary readers are patients, as in the case of MedlinePlus. Although the results of extracting candidate term pairs from our resource describe din section 4.1 is still very preliminary, it is already clear that improving the candidate terms detection followed by manual validation is extremely efficient to quickly expand terminological resources for the medical domain.

6. Acknowledgements

We acknowledge the Encomienda MINETAD-CNIO/OTG Sanidad Plan TL and OpenMinted (654021) H2020 project for funding.

7. Bibliographical References

Chiao, Y. and P. Zweigenbaum, Looking for French-English translations in comparable medical corpora, in: Proceedings of AMIA Symposium, 2002.

Deleger, L., M. Mergel, and P. Zweigenbaum, Translating medical terminologies through word alignment in parallel text corpora, Journal of Biomedical Informatics 42 (4) (2009) 692-701.

Deleger, L., T. Merabti, T. Lecrocq, M. Joubert, P. Sweigenbaum, and S. Darmoni, A Twofold Strategy for Translating a Medical Terminology into French, in: AMIA Annual Symposium Proceedings, 2010.

Flores, G., Laws, M. B., Mayo, S. J., Zuckerman, B., Abreu, M., Medina, L., & Hardt, E. J. (2003). Errors in medical interpretation and their potential clinical consequences in pediatric encounters. Pediatrics, 111(1), 6-14

Huang, C. C., & Lu, Z. (2015). Community challenges in biomedical text mining over 10 years: success, failure and the future. Briefings in bioinformatics, 17(1), 132-144.

Krallinger, M., Valencia, A., & Hirschman, L. (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome biology, 9(2), S8.

Liu, W., S. Cai, B. P Ramesh, G. Chiriboga, K. Knight, and H. Yu. Translating Electronic Health Record Notes from English to Spanish: A Preliminary Study, in Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), pages 134–140, Beijing, China. Association for Computational Linguistics.

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform, 35(8), 128-144.

Neves, M. (2014). An analysis on the entity annotations in biological corpora. F1000Research, 3.

Neves, M., Yepes, A. J., and Névéol, A. (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. European Language Resources Association (ELRA).

Rebholz-Schuhmann, D., Clematide, S., Rinaldi, F., Kafkas, S., van Mulligen, E. M., Bui, C., ... & Jimeno-Yepes, A. (2013, September). 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 (pp. 353-367). Springer, Berlin, Heidelberg.

Rodrigo Agerri, Josu Bermudez and German Rigau (2014): "IXA pipeline: Efficient and Ready to Use Multilingual NLP tools", in: Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014), 26-31 May, 2014, Reykjavik, Iceland.

Tiedemann, J. (2009). News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In Recent advances in natural language processing (Vol. 5, pp. 237-248).

Villegas, M., S. de la Peña, M. Krallinger, A. Intxaurrondo, J. Santamaría. Esfuerzos para fomentar la minería de textos en biomedicina más allá del inglés: el plan estratégico nacional español para las tecnologías del lenguaje, Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp 141-144.

Wu, C., F. Xia, L. Deleger, and I. Solti. Statistical Machine Translation for Biomedical Text: Are We There Yet? AMIA Annual Symposium Proceedings:1290–1299 2011.

Yepes, A. J., Névéol, A., Neves, M., Verspoor, K., Bojar, O., Boyer, A., ... & Pecina, P. (2017). Findings of the WMT 2017 biomedical translation shared task. In Proceedings of the Second Conference on Machine Translation (pp. 234-247).

Yepes, A.J., É. Prieur-Gaston, and A. Névéol. Combining MEDLINE and publisher data to create parallel corpora for the automatic translation of biomedical text. BMC Bioinformatics, 14(1):146, April. 2013.