A Machine Translation Approach for Medical Terms

Alejandro Renato, José Castaño, Pilar Ávila, Hernán Berinsky, Laura Gambarte, Hee Park, David Pérez, Carlos Otero and Daniel Luna
Departamento de Informática en Salud, Hospital Italiano de Buenos Aires, Argentina

Keywords: Controlled Vocabulary, Description Terms, Machine Translation.

Abstract: We describe the task of translating clinical term descriptions from Spanish to Brazilian Portuguese. We build a statistical machine translation system (SMT) using in-domain parallel corpora and available machine learning tools. The performance of this SMT was compared with general purpose machine translation systems available online. We used different techniques to validate the result of the different systems, using reference domain terminology and the occurrence of translated descriptions in a corpus of medical scientific literature and in domain specific web pages. We also use two sets of 1000 description terms that were revised and checked by a Portuguese speaker. The performance of the SMT we built had very good preliminary results.

1 INTRODUCTION

Electronic health records (EHR) use controlled vocabularies in order to give proper semantic interpretations. SNOMED CT is a well known concept hierarchy that provides a conceptual interpretation, mapping terms descriptions to concepts in many languages. Unfortunately, there is no version for SNOMED CT in Portuguese, and the process of translating SNOMED CT (as any specific ontology) to a new language is a long and costly process that requires careful human supervision (Reynoso et al., 2000). Also given SNOMED CT is a proprietary thesaurus, it cannot be freely translated without a proper license and following a number of required steps.

In order to obtain a controlled terminology vocabulary for Portuguese, we use our interface Spanish vocabulary which extends SNOMED CT. The Hospital Italiano de Buenos Aires (HIBA) was implemented in 2002 (**?**; **?**). Each term in HIBA thesaurus is mapped via a direct relation or using compositional post-coordinated expressions to SNOMED CT as its reference vocabulary. At present, it has 207000 post-coordinated concepts in its terminology system. The HIBA terminology services have been developed over a period of more than ten years facing the challenges of mapping new terms automatically and encoding new expressions as well as characterizing those terms that provide no interpretation or partial meaning. The terminology system allows the use of free text descriptions providing more expressiveness and flexibility and adaptability to different medical practices. It provides alternative synonym expressions and fine grain information description and interpretation. Descriptions terms are short texts, mostly 3 to 5 words long. Also, it has the advantage that historical use information can be used to prioritize frequent terms and to detect new terms.

We analyze alternative possibilities to translate automatically HIBA description terms to Portuguese (see (Schulz et al., 2013) for a comparison of Human vs Machine translation of SNOMED CT in German). An initial automatic translation might provide a good starting point for a controlled vocabulary in Portuguese. Due to the characteristics of medical descriptors - short texts, specific vocabulary, simple syntax and reduced semantic ambiguity - the task of automatic translation is presented as a possibility to obtain good results. A controlled vocabulary provides a robust alternative to mitigate some of the problems of natural languages, in particular, ambiguity (Nyberg et al., 2003). Also, the possibility of translating these description terms from Spanish presents a good opportunity due to the transparency existing between both languages.

General purpose automatic translation systems, such as Microsoft (Bing) Translator, Google Translate, IBM Language Translator and many others face the complex task of translating texts without specific domain customization.

However, medical texts are not exempt from synonymy and ambiguity. Terms such as *dolor de cabeza*

Renato, A., Castaño, J., Ávila, P., Berinsky, H., Gambarte, L., Park, H., Pérez, D., Otero, C. and Luna, D.
A Machine Translation Approach for Medical Terms.
DOI: 10.5220/0006555003690378
In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) - Volume 5: HEALTHINF*, pages 369-378
ISBN: 978-989-758-281-3

(headache), and *cefalea* (cephalea) are examples of synonymy. Abbreviations and acronyms are most critical examples of ambiguity, they usually have different meaning according to medical discipline: *OD* in Spanish can mean, right ovary, right ear or right eye.

Synonymy can be partially dealt taking advantage of the linguistic proximity between Spanish and Portuguese since many of the words present in Spanish have their counterpart in Portuguese as the following examples show:

(1) dolor de cabeza (Spanish)
dor de cabeça (Portuguese)
headache

(2) cefalea (Spanish)
cefaleia
cephalea

(3) migrañas
migrâneas
migraines

(4) jaqueca
enxaqueca
headache

It is not clear whether a general purpose machine translation system can provide reasonable results in this domain and if so which of the available ones would fit best to the task. Also, it is not clear whether it is possible to build a customized SMT for this domain using machine learning tools. Building a customized SMT with reasonable performance requires the availability of parallel corpora in the corresponding languages.

Medical descriptions can be reduced to simple noun (NP) and prepositional phrases (PP). An approach that takes advantage of these characteristics has already been discussed (Koehn and Knight, 2003), demonstrating that a dedicated translation subsystem can have better performance than general purpose SMT by incorporating special modeling and features.

The need for the availability of SNOMED ontology in Portuguese has been addressed previously. (Barcelos Junior et al., 2008), propose a relational database as a form of access to SNOMED-CT through a mapping with DeCS. (Pacheco, 2009) has a similar approach but using NLP tools as the interface to different languages: English, German, Spanish or French. (Silva et al., 2015) have worked on the translation of UMLS ontologies from Portuguese to Portuguese from Brazil, through the creation of an algorithm for translation between languages. (Lucas Emanuel Silva e Oliveira, 2016) have attempted an approach based also on pattern matching, between ICD-9, DBPedia and SNOMED CT using also Google Translator.

However, we have not found previous attempts to build a dedicated SMT for this task. There is an open-source general purpose Portuguese-Spanish SMT reported (Armentano-Oller et al., 2006).

The remainder of this work is organized as follows, in Section 2, we describe a preliminary evaluation of available general purpose SMTs. Section 3 describes the implementation of an SMT for Spanish-Portuguese in the medical domain (M-SMT). Finally, in Section 5 we discuss results and future work.

2 PRELIMINARY EVALUATION OF SMTS AS TOOLS TO ASSIST HUMAN TRANSLATORS

In this section, we describe a series of experiments which aimed to assess the feasibility of using SMTs to translate a large data-set of medical terms. The goal of these experiments and its description here is to provide a set of tests that can provide a good indicator whether an automatic translation task and reference data have some chances to provide a starting point for a curated human translation of a controlled terminology (Koehn and Haddow, 2009). Currently, the source terminology system in Spanish has a population close to two million terms. We expected that general purpose SMTs could be cross-domain robust enough to provide a starting point and speed up a process of curated translation. We also considered necessary to provide some source of automatic validation of the translations proposed by a SMT. For this purpose, we compiled a database of medical terms which can be used to guarantee at least that a translated term belongs to the domain. It should be noted that being a medical term does not imply that a translated term preserves the same meaning. The second alternative considered that a high number of hits provided by Google Search could provide some confidence on the proposed translation. A low-frequency term can be considered as a strong indicator of a dubious translation and can be used to filter the translation. Both the database of medical terms, the use of a search engine on reference domain texts in the target language, and several SMT can be used as tools in a model of computer-assisted translation (Bowker and Fisher, 2010).

2.1 Compilation of Portuguese Medical Terms

We used several sources to compile a dataset of Portuguese medical description terms. The purpose of this dataset was to obtain a set of controlled reference description terms. If a translated term is in this set, it is at least an existing medical term, although it might not be a proper translation. It should be noted that many automatic translations of medical terms do not produce syntactic or semantically well-formed terms. Semantic transparency between Spanish and Portuguese provides some confidence that a translated term existing in a controlled reference set be a proper translation. A total of 190,381 unique terms was compiled. We looked for different terminological sources to compile this dataset. We detail the sources as follows in Table 1.

Table 1: Source medical vocabularies.

163,171	DeCS Health Science Descriptors
8,982	Dicionario Médico (Em Portuguese do Brasil) (pdf)
6,714	Vocabulario de Medicina (pdf)
20,907	wikipedia_medicine_2017-02.zim
6,078	Diccionario de Termos Médicos e de Enfermagem (pdf)
9,567	http://www.dicionáriomédico.com

The total terms (including repetitions) that originated from different sources was 215,419 which shows there is an intersection of 25,038 description terms coming from more than one source. It should be noted that the source of most description terms is DeCS Health Science Descriptors.

2.2 Initial Experiments on Most Frequent Terms

A set of the 1262 most frequent terms from the HIBA terminology database was selected to assess the behavior of Google Translate.[1] Each term was translated by phrase and by word. Also, the Google API was used to check the number of hits each translated term had on Google search. Each translation was checked against the data-set of compiled medical terms.

From these most frequent set, 471 were *validated*[2] against the compiled dictionary of terms, 67 were *validated* by the number of hits using Google Search: they had more than 500 hits (an arbitrary threshold selection) and 724 (57%) were not validated by either method.

It was apparent that Google translation by word had most of the times syntactic ill-formed terms. There were also some weird translations: Spanish *resfrío* (cold) was incorrectly translated as Portuguese *frío* (cold) and *Accidente cerebrovascular* (stroke) was translated as Portuguese *golpe* (stroke)[3]. Spanish *angina de pecho* was incorrectly translated as Portuguese *angina*, missing part of the information. This situation was not isolated but there were other examples in which the translation was semantically inconsistent.

This set of translations was verified and alternative translations were provided by a second language Brazilian Portuguese speaker and linguist.[4] Out of the 471 translations validated against the compiled dictionary, there were 79 (16%) which had disagreements (0.832 agreement). Clearly, most of them might be considered alternative translations. Even some of them were acronyms which were correctly translated by Google translator:

(5) TTH (cefaléia tensional)
Tension-type headache (TTH)

From the 67 additional translations that had more than 500 hits, there was more disagreement: only 27 agreed (40%). Some disagreements looked like alternative translations:

(6) dor no pé vs. podalgia
pain in the foot vs. podalgia

(7) cancro do pâncreas vs. câncer pâncreas
pancreatic cancer

The remaining 724 non validated phrase translations had much less agreement: only 227 agreed (29%) .

A subset of 933 terms was used to measure agreement between translators: using Google Translate 338 (36%) agreed with human translator and 595 disagreed. Regarding Bing, 342 translated terms agreed with the human translator and 591 disagreed. From these 342 terms, 216 (63%) agreed with Google Translate and 126 did not.

As these results show, there was only about half of

[1] The arbitrary number of 1262 corresponds to a database selection of the 1000 most frequent concepts and their synonyms.

[2] We use the term *validate* informally here, and it makes reference to the possibility of a human using the database of terms to check if the MT is correct.

[3] We suppose that these translations were caused because Google might be using English as an intermediate step in the translation. However, Spanish *resfrío común* was correctly translated as Portuguese *resfriado común*.

[4] We did not have available a domain native speaker to perform the translation.

the most common terms that could be translated with a certain confidence. These results describe more an initial assessment on using SMTs and reference terminology to assist human translation than a formal MT comparison. Later on, we made a more thorough comparison of the three automatic translators, Google, Bing and our M-SMT using SMT. It will be discussed in more detail in Section 3 (Tables 3 and 4).

2.3 Experiments with Larger Data-sets

A second experiment was performed using a set of 17000 terms which were mapped to children concepts from the 1000 concepts selected in the previous experiment. A partial corpus of scientific medical papers was collected (see next section), to use for validation. Systran (www.systranet.com), was also used as a translator. And a very preliminary version of an SMT we implemented was tested. We used Google Search for hits on a set of about ten sites of the medical domain in Brazilian Portuguese. Search was restricted to the medical domain because search without domain restriction returned many results which were not trustworthy. The following table 2 shows the results.

Table 2: Validation possibilities over 17,000 terms.

759	validated using database of terms
955	validated using medical papers
1123	showed co-incidence between three systems
360	hits in domain sites
13897	no validation

A partial conclusion of this experiment was that Bing Translator performed better than Google Translate in the domain, and at that an early stage the M-SMT developed with parallel corpora was close to Bing Translator. Systran showed an apparent lower quality translation in the domain. These results show that available online SMTs have a limited capability to assist the translation process of medical description terms.

Then we selected the 88,000 most frequent terms from the terminology database and run translation with Google, Bing, our M-SMT. From these, 3,990 could be validated as terms using the compiled term database. We didn't run Google Search. There was a number of 9,212 translated terms which showed the same translation using Google, Bing and M-SMT. We discuss this dataset of 88,000 terms and the M-SMT in the next section.

3 A MACHINE TRANSLATION SYSTEM FOR MEDICAL TERMS (M-SMT)

In order to build an SMT for medical terms, we decided to follow several steps. First, we had to compile two corpora: a) a parallel corpus for Spanish and Portuguese and b) a reference corpus of medical texts in Portuguese to obtain an adequate language model for this domain. The purpose of the first corpus is to train the automatic translation engine and the purpose of the second one is to validate the translations based on its language model. At a second stage, the parallel corpus was used to train the automatic translation engine. Then the Spanish data-set of 88,000 terms from the HIBA thesaurus was translated into Portuguese. Finally the translations were validated according to their presence in the reference corpus.

3.1 Parallel Spanish-Portuguese Corpus

The parallel training corpus was constructed using the following corpus:

- The ICD-10 version in Portuguese of Brazil[5] and Spanish[6] 10890 sentences
- DeCS- version in Brazilian Portuguese and Spanish [7] 73515 sentences.
- EMEA(Tiedemann, 2009) version in Portuguese and Spanish[8] 1084906 sentences
- A total of 3100 additional frequent words which did not have a translation were translated by the Portuguese speaker.

3.2 A Reference Corpus of Portuguese Scientific Medical Texts (PSMT)

The PSMT corpus was compiled using three sources:

[5]ICD-10 corresponds to the 2008 version of CID that was downloaded from the site: DATASUS, "Department of Infomatica do SUS", Ministry of Health of Brazil. http://www.datasus.gov.br/cid10/V2008/cid10.htm

[6]The version in Spanish ICD 10, version 2008, was downloaded from the site of the "Pan American Health Organization" ais.paho.org

[7]The DeCS (Descriptors in Health Sciences) was created from MeSH with the objective of the use of common terminology for research in three languages (Portuguese, English and Spanish). It is a vocabulary with approximately 29,000 descriptors, of which 23,963 were MeSH, 218 Science and Health, 1,951 Homeopathy, 3,486 Public Health and 828 Health Surveillance

[8]The corpus EMEA was downloaded from http://opus.lingfil.uu.se/EMEA.php

- Medical publications from SciELO.
- The Portuguese part of the corpus of the parallel corpus corresponding to ICD-10 and BIREME.
- The vademecum in Brazilian Portuguese was converted to text.[9] This text was considered important since it contains detailed names of pathologies, pharmacological products and drugs.

We selected the set of journals under the subject Health Sciences, comprising 114 publications from the Medical publications available at SciELO "Scientific Electronic Library Online", Brazil. [10] We downloaded from Scielo the complete publications. Approximately 180,000 documents were obtained in pdf format. The documents were converted to text format using Tika (https://tika.apache.org), which was also used for language detection. Those texts in which the detected language was Portuguese or Galician (the detector tends to find texts where there are paragraphs in Portuguese and Spanish as Galician) were selected. Approximately 70,000 items were made available. The text was split into sentences and tokenized (using OpenNLP). Sentences with frequent words in English and Spanish (common in bibliographical notes and summaries) were removed, otherwise words in those languages would generate false positives when searching for Portuguese phrases.

3.2.1 Medical Descriptors from the Spanish Thesaurus

The 88000 occurrences of the most frequent descriptors have a vocabulary of 16541 words. In order not to leave words outside the vocabulary, the occurrence of each word was searched within the parallel corpus of texts in Spanish and Portuguese. Within the corpus, each text contributed the following amounts:

- EMEA (7885)
- ICD-10 (4033)
- DeCS (6392)
- Vademecum (438)

There were 10111 unique words in the above corpora vocabulary, and there were 6430 OOV (out of vocabulary) words, i.e. words that were in the Spanish source vocabulary and were not found in the parallel corpus. In order to solve the OOV words various strategies were used to obtain a translation for these words. These translations were validated using the PSMT corpus.

The first strategy consisted in searching OOV Spanish words in the Portuguese corpus, since in some cases, such as drugs or pathologies the same word is used in both languages with no significant changes, after accentuation normalization.

Second, the strategy was to change regular morphological suffixes in Spanish by the corresponding Portuguese ones. For instance, those words ending in the suffix *-ción* in Spanish usually have a Portuguese counterpart written as *-cão*. Words ending in *-itis* or *-isis* in Spanish meaning 'em inflammation usually have a Portuguese translation ending in *-ite* and *-ise* respectively. After these transformations, the resulting words were searched in the corpus.

Third, it is possible to find compound words (Compound Splitting) that can easily be broken down into simple words, such as: *vesiculo- ..*, *eritemato-*, *uretero-*, .

Fourth, we tried to relate the words by orthographic similarity. For this purpose, we used the tool *aspell*. A dictionary was created with those words obtained from Portuguese texts and corrected those words that were in Spanish (which keep similarity with the words of Portuguese, for example: generating a set of lists of suggestions. Suggestions were sought with the context as they appeared in Spanish.

In this way, it was possible to reduce the OOV from 6430 to 2450. It should be noted that these words are generally low-frequency words in the corpus of 88000 descriptions, most of them having a single occurrence within the corpus.

A final attempt to reduce even further the OOV word was performed searching for possible translations in domain dictionaries and Google search.

3.3 The Medical-SMT

In order to implement an SMT, we used Moses (Koehn et al., 2007) software, a phrasal-based probabilistic machine translation engine, which was used by many teams at the recent First Conference on Machine Translation (WMT-16)(Bojar et al., 2016). Input sequences are segmented into a number of (non-linguistic) phrases, each phrase is translated using a phrase translation table and allow for reordering of phrases in the output. No phrases may be dropped or added. The texts of the parallel corpus were divided into sentences, tokenized and converted to lowercase. Those sentences that did not meet various requirements were eliminated. The resulting corpus was 120,214 sentences. The Spanish corpus

[9] http://br.prvademecum.com

[10] The library is the result of a research project of FAPESP - Foundation for Research Support of the State of São Paulo, in collaboration with BIREME - Latin American and Caribbean Center for Information on Health Sciences. Has the support of CNPq - National Council for Scientific and Technological Development. http://www.scielo.br/

was constituted by a vocabulary of 105852 different words whereas the Portuguese vocabulary was 101506 words (excluding numbers). Then a Portuguese language model was built with the texts coming from the PSMT corpus described in 3.2. A trigram model was built using KenLM software in Moses (Koehn et al., 2007) software. The resulting model had 1748668 unigrams, 14659827 bigrams and 47825896 trigrams.

4 EVALUATION AND RESULTS

4.1 SMT Evaluation

In order to evaluate the performance of the M-SMT translator and to compare it with available general purpose translators, we used measurements from the MT scientific community. METEOR is based on "the harmonic mean of unigram precision and recall, with recall weighted higher than precision" (Denkowski and Lavie, 2014). TER (Snover et al., 2006) is an error metric *that measures the number of edits required to change a system output into one of the references*. The software used was Meteor (Denkowski and Lavie, 2014) and Multeval(Clark et al., 2011).

BLEU (Papineni et al., 2002; Och and Ney, 2003) is a metric whose objective is to show the closest proximity between machine translation and the one performed by a human being. BLEU is language independent and it is one the most widely used automated methods to determine machine translation quality.

A BLEU score ranges from 0 to 1. The more the translation correlates with human translation, the closer the score gets to one. BLEU metric is able to measure how many words overlap in a given translation and a reference translation, with sequential words being given higher scores. Scores below 15% indicate that the machine translation is unable to provide a translation and a high level of post-editing is required. Scores greater than 30% indicate that translations can be understood. Scores above 50% indicate better quality translations. It should be noted that there is a great variation between SMT performance in different corpora:

Google (Johnson et al., 2016) reports a BLEU Score of 44.40 for Portuguese to English Translation and a Score of 38.40 for English to Portuguese Translation using Google internal datasets. (Masselot and Ribiczey, 2010) reports a BLEU score of 68.31 for Spanish-Portuguese in the software domain and (Aziz and Specia, 2011) report a BLEU performance of 71.49 also for Spanish-Portuguese translation using a parallel corpus from a scientific Brazilian journal (Pesquisa FAPESP Online) using Moses toolkit.

Google Translate performance using sentences from the UniversalDoctor project in the medical domain using English as source language had the following performance: French, 24.30, Portuguese 19.51, Spanish 26.34 (Costa-jussà et al., 2012)).

The results reported in First Conference on Machine Translation (WMT16) used also English as a main language (Bojar et al., 2016) and reported results in the Biomedical domain with separate tests for health and biological articles. The results corresponding to health articles are as follows: English/Portuguese (19.01), Portuguese/English 21.50, English/Spanish 29.47, Spanish/English 29.05 and English/French 22.75.

We performed two evaluations were performed: (a) on the corpus of 1262 most frequent terms (see Tables 3 and 4) and (b) on a corpus of 1000 descriptions randomly selected from the 88014 descriptions described in Section 2 (see Tables 5 and 6).

Table 3: Machine Translation comparison - 1262 most frequent descriptions.

Parameters	**M-SMT**	**Google**	**Bing**
Test Words	2677	2509	2695
Ref. Words	2612	2612	2612
Chunks	378	704	717
Precision	0.8501	0.6681	0.6850
Recall	0.8719	0.6513	0.6892
f1	0.8608	0.6596	0.6871
fMean	0.8652	0.6534	0.6880
Fragmentation Penalty	0.0024	0.083	0.078
Final Score	**0.8443**	**0.6016**	**0.6338**

Table 4: Metric scores on 1262 most frequent descriptions.

Metric	**System**	**Avg**	$\bar{s}_{\mathbf{sel}}$
BLEU ↑	M-SMT	58,9	2,1
	Bing	22,4	1,8
	Google	18,7	1,6
METEOR ↑	M-SMT	46,9	0,8
	Bing	30,4	0,6
	Google	26,6	0,6
TER ↓	M-SMT	23,8	1,2
	Bing	51,5	1,4
	Google	51,3	1,2
Length	Moses	104,9	0,6
	Bing	103,4	0,9
	Google	96,0	0,8

In the second evaluation, we used 1000 manually corrected sentences to evaluate the performance of the three translation systems M-SMT, Google Trans-

late and Bing Translator. The results show that an SMT trained with domain-specific corpora significantly outperforms general purpose translators.

Table 5: Machine Translation Comparison.

Parameters	Moses	Google	Bing
Test Words	4654	4782	4230
Ref. Words	4561	4561	4561
Chunks	359	1397	1345
Precision	0.9488	0.6595	0.7615
Recall	0.9606	0.6791	0.7296
f1	0.9547	0.6692	0.7452
fMean	0.9570	0.6731	0.7389
Fragmentation Penalty	0.0089	0.6096	0.6778
Final Score	**0.9485**	**0.6096**	**0.6778**

Table 6: Metric scores on 1000 random descriptions.

Metric	System	Avg	$\bar{s}_{sel}$
BLEU ↑	Moses	86,7	0,9
	Bing	41,8	1,4
	Google	37,9	1,4
METEOR ↑	Moses	60,7	0,6
	Bing	36,5	0,5
	Google	33,1	0,5
TER ↓	Moses	8,0	0,6
	Bing	40,5	1,1
	Google	44,5	1,0
Length	Moses	102,0	0,3
	Bing	101,5	0,5
	Google	92,7	0,6

It can be observed that there are differences between the performance in the two data-sets. M-SMT and Bing performed better in the random selected descriptions than in the most frequent description set. There are two reasons for this. Most frequent descriptions, due to their frequency might have more synonym expressions, which might not be captured in the evaluation. The second reason is that there might be a bias in the random set because the translations were corrected using as a source M-SMT translation. Looking at the BLEU metrics 58.9% for most frequent descriptions and 86,7% for the random set are consistent with the Spanish-Portuguese scores reported on other domains (68.31% (Masselot and Ribiczey, 2010) and 71.49% (Aziz and Specia, 2011)) mentioned above.

4.2 Error Analysis

General purpose machine translation produced various types of errors. They can be mentioned in order of importance:

a) OOV words are usually translated into English or left in Spanish

(8) gastrinoma relapse (”incorrect”)
recidiva de gastrinoma (”correct”)
gastrinoma relapse

(9) laringobronquitis (”incorrect”)
laringo-bronquite (”correct”)
laringobronchitis

b) Translation to a hyperonym phrase,

(10) dor na parte de trás de ambas as mãos (”incorrect”)
dor no dorso de ambas as mãos (”correct”)
pain in the back of both hands

(11) ferida nas pernas (”incorrect”)
leg injury
ferida na região pré-tibial (”correct”)
pre-tibial injury

c) Translate to most common synonym in colloquial language instead of medical term

(12) caroço na glabela (”incorrect”)
carotid in the glabella
tumoração na glabela (”correct”)
tumor in the glabella

(13) dormência no braço (”incorrect”)
Numbness in the arm
parestesia no braço (”correct”)
paresthesia in the arm

d) Change of grammatical category

(14) queimando em ambos os olhos (”incorrect”, verb)
ardência em ambos os olhos (”correct”, noun)
burning in both eyes

(15) queimar na mão e nádega (”incorrect”, infinitive)
burn hand and buttock
queimadura da mão e nádega (”correct” , noun)
burn of hand and buttock

e) Syntactic ill formed constructions, an adjective depends on the wrong noun. Penalty in terms of fragmentation is indicative of these errors. It is related to word order, which tends to have a tendency toward English syntax in online translators.

(16) menor alergia a amoxicilina (”incorrect”)
lower allergy to amoxicillin
alergia menor a amoxicilina (”correct”)
allergy lower to amoxicillin

(17) direito de plantação de dermatite (”incorrect”)

Planting right dermatitis
dermatite plantar direita (”correct”)
plantar dermatitis right

M-SMT using Moses trained translation presented the following problems:

a) OOV words

(18) alergia ao óleo da polpa de durazno (”incorrect”)
allergy to peach pulp oil

(19) alergia ao óleo da polpa de pêssego (”correct”)
allergy to peach pulp oil

b) Oscillating orthography (a part of the corpus had words with spelling in European Portuguese)

(20) abcesso séptico intra-raquidiano (”incorrect”)
intra-spinal septic abscess
abscesso séptico intra-raquidiano (”correct”)
intra-spinal septic abscess

c) Compound medical terms, especially drugs with a hyphen, possibly misaligned in training.

(21) alergia ao mesilato de alfa diidroergocristina (”correct”)
allergy to alpha dihydroergocristine mesylate
alergia ao mesilato de alfadihidroergocristina (”incorrect”)
allergy to alpha dihydroergocristine mesylate

4.3 Automatic Validation of Results using Reference Corpus

In order to validate the translation of the set of 88014 medical descriptors we searched for the translation in the reference PSMT corpus of medical publications described in 3.2, and then the nominal phrases contained in these descriptors were searched in the same corpora. We show results in Table 7, description length in number of words has a variation between 1 and 28.

Table 7: Medical Descriptors found in Medical Literature. (MTL: mean term length, # words).

	Percent	Total Terms	MTL
Found	31.96%	28131	3.55
Not Found	68.04%	59883	5.15
Total	100%	88014	4.64

In a third evaluation, we searched nominal chunks in the reference corpus. We have verified that more complex nominal phrases can be decomposed into simpler phrases. So those descriptions that were not found in the reference corpus were split into nominal chunks, for instance:

(22) (NP carcinoma basocelular lobulado) (PP da pele da região malar esquerda)
(NP lobular basal cell carcinoma) (PP of the skin of the left malar region)

was split in:

(23) carcinoma basocelular lobulado
lobular basal cell carcinom

(24) pele
skin

(25) região malar esquerda left malar region

Another difficulty we have observed is due to the fact that in medical descriptions laterality must be recorded (if the body location is right or left), but this information is not relevant in scientific texts. Then, we removed those words indicating laterality to perform search.[11]

(26) contusão no antebraço [esquerdo]
Contusion in the forearm [left]

Simplified nominal chunks from descriptions had a word length between 1 and 12. As in the previous case, these chunks were searched in the corpus of medical literature. Table 8 shows the results for total unique nominal chunks (type) and Table 9 for total nominal chunk occurrences (tokens).

Table 8: Nominal Phrase from Descriptors found in Medical Corpora (Unique terms).

	Percent	Total Terms	MTL
Found	50.21%	23099	2.08
Not Found	49.79%	22903	2.65
Total	100%	88014	2.37

Table 9: Nominal Phrase from Descriptors found in Medical Corpora (Total of terms).

	Percent	Total NPs	MTL
Found	57.71%	97984	1.50
Not Found	42.28%	71776	1.99
Total	100%	169760	1.90

It can be observed that those strings that were

[11]There are other properties required in medical descriptions that are not relevant in scientific texts, such as, the finger specification or month of pregnancy, but we did not implement any simplification to search simplified nominal chunks.

not found correspond to those descriptions of greater word length. In some cases, this may be due to the fact that longer medical descriptions generally have a sentence structure with bare verbs and a telegraphic style. This type of constructions is not very common neither in the training corpus nor in the corpus of scientific texts. Also, additional human translators should translate the two test sets and a proper measurement for human inter-translator agreement should be obtained.

5 CONCLUSIONS

We explored the use of alternative tools to assist the translation of medical terminology from Spanish to Portuguese. General purpose SMTs showed a number of deficiencies which limited their use for this purpose. We implemented an M-SMT for Spanish-Portuguese translation which showed much better performance than general purpose ones. The work described here showed that an approach based on using parallel corpora and linguistic mappings to reduce out of vocabulary words have been successful to address the problem with very good performance. In future work, we will consider the use of other tools and techniques to improve the results of a SMT for the medical domain.

REFERENCES

Armentano-Oller, C., Carrasco, R. C., Corbí-Bellot, A. M., Forcada, M. L., Ginestí-Rosell, M., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., Sánchez-Martínez, F., and Scalco, M. A. (2006). Open-source portuguese–spanish machine translation. In *International Workshop on Computational Processing of the Portuguese Language*, pages 50–59. Springer.

Aziz, W. and Specia, L. (2011). Fully automatic compilation of portuguese-english and portuguese-spanish parallel corpora.

Barcelos Junior, C. L., Andrade, R., Ribeiro, L. A., and von Wangenheim, A. (2008). Busca semântica aplicada a informações clínicas. In *XI Congresso Brasileiro de Informática em Saúde*.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016). Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Bowker, L. and Fisher, D. (2010). Computer-aided translation. *Handbook of translation studies*, 1:60–65.

Clark, J. H., Dyer, C., Lavie, A., and Smith, N. A. (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 176–181. Association for Computational Linguistics.

Costa-jussà, M., Farrús, M., and Serrano-Pons, J. (2012). Machine translation in medicine. a quality analysis of statistical machine translation in the medical domain. *Proceedings of the 1st Virtual International Conference on Advanced Research in Scientific Areas (ARSA-2012)*.

Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2016). Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Koehn, P. and Haddow, B. (2009). Interactive assistance to human translators using statistical machine translation methods. *MT Summit XII*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Koehn, P. and Knight, K. (2003). Feature-rich statistical translation of noun phrases. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 7-12 July 2003, Sapporo Convention Center, Sapporo, Japan.*, pages 311–318.

Lucas Emanuel Silva e Oliveira, Sadid A. Hasan, O. F. e. C. M. C. M. (2016). Translation of umls ontologies from european portuguese to brazilian portuguese. In *XV Congresso Brasileiro de Informática em Saúde 27 a 30 de novembro - Goiânia - Brasil.*

Masselot, F. and Ribiczey, P. (2010). Using the apertium spanish-brazilian portuguese machine translation system for localization.

Nyberg, E., Mitamura, T., and Huijsen, W.-O. (2003). Controlled language for authoring and translation. *Computers and Translation: A Translator's Guide*, 35:245.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Pacheco, E. J. (2009). Morphomap: mapeamento automático de narrativas clínicas para uma terminologia médica.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Reynoso, G. A., March, A. D., Berra, C. M., Strobietto, R. P., Barani, M., Iubatti, M., Chiaradio, M. P., Serebrisky, D., Kahn, A., Vaccarezza, O. A., et al. (2000). Development of the spanish version of the systematized nomenclature of medicine: methodology and main issues. In *Proceedings of the AMIA Symposium*, page 694. American Medical Informatics Association.

Schulz, S., Bernhardt-Melischnig, J., Kreuzthaler, M., Daumke, P., and Boeker, M. (2013). Machine vs. human translation of snomed ct terms. In *Medinfo*, pages 581–584.

Silva, M. J., Chaves, T., and Simões, B. (2015). An ontology-based approach for SNOMED CT translation. In *Proceedings of the International Conference on Biomedical Ontology, ICBO 2015, Lisbon, Portugal, July 27-30, 2015.*

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

Tiedemann, J. (2009). News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248.