IXA Biomedical Translation System at WMT16 Biomedical Translation Task

Olatz Perez-de-Viñaspre and **Gorka Labaka**
IXA NLP Group
University of the Basque Country UPV/EHU
Donostia, Basque Country
olatz.perezdevinaspre,gorka.labaka@ehu.eus

Abstract

In this paper we present the system developed at the IXA NLP Group of the University of the Basque Country for the Biomedical Translation Task in the First Conference on Machine Translation (WMT16). For the adaptation of a statistical machine translation system to the biomedical domain, we developed three approaches based on a baseline system for English-Spanish and Spanish-English language pairs. The lack of terminology and the variation of the prominent sense of the words are the issues we have addressed on these approaches. The best of our systems reached the average of all the systems submitted in the challenge in most of the evaluation sets.

1 Introduction

In this paper we present the system developed at the IXA NLP Group from the University of the Basque Country for the Biomedical Translation Task in the First Conference on Machine Translation (WMT16). This is the first shared task organized for the biomedical domain inside WMT.

The Biomedical Translation Task consists of translating scientific abstracts in health and biological domains for languages such as English, Spanish, French and Portuguese. In our case, we developed a system for English-Spanish and Spanish-English pairs.

We present a system that takes a general Moses statistical machine translation system (Koehn et al., 2007) and adapts it to the biomedical domain. The adaptation of a MT system to a specific domain comes with two main issues: i) a bigger set of out-of-vocabulary (OOV) words and ii) the variation of the prominent sense of the words.

The integration of a bilingual biomedical terminology bank to the system can mitigate great part of the lack of terminology. In any case, this may not be enough and a transliteration[1] module may be helpful. In addition, morphological variability may be a problem in no-frequent lemmas, as plurals in English or genders and plurals in Spanish.

The remaining of the paper is as follows. We first present in section 2 the resources we used. We then describe in section 3 the approaches we developed for our system and in section 4 the BLEU results for our runs. Finally, conclusions are drawn.

2 Resources

In this section we describe the resources used to train the models that will be explained in section 3. There are two main resource types involved in this work: corpora and terminological resources.

2.1 Corpora

The corpora pertains to two different sub-domains: health and biology. Thus, the corpus extracted from Scielo is separated by the domain the abstracts pertains. In the case of the Medline corpus there is a unique corpus for both sub-domains.

Although the corpora is in general bilingual and aligned at sentence level, in some cases sentences from the parallel corpora were not available, as in the Medline corpus some English sentences were marked as "[Not Available]". We removed those sentences from the parallel corpus and we created a monolingual corpus of Spanish sentences to be used for language modeling.

The Scielo corpora gives the word alignments as well as sentence alignments. Thus, in table 2.1 we show the number of sentences and words that

[1] Although transliteration is commonly used between languages with different scripts, it may also be used to adapt the spelling differences of borrowings.

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 477–482,
Berlin, Germany, August 11-12, 2016. ©2016 Association for Computational Linguistics

are aligned in the bilingual corpora for the Spanish and English pairs.

Corpus	Sentences	Words
Scielo - Biological	125,828	723,202
Scielo - Health	587,299	2,871,232
Medline	285,584	-

Table 1: Bilingual corpora

From this bilingual corpora, we excluded some sentences for development. On the one hand, we created a domain-balanced set of 2,945 sentences for tuning of the translation model as well as to interpolate the LM. This set was taken in a randomized and balanced way so we maintained the percentages of the original sets. That is, we took 361 sentences from the biological set, 1,726 sentences from the health set and 858 from the Medline set.

On the other hand, we excluded a separate set of each corpus (health, biological and Medline). In this case, we excluded 2,000 sentences from each of the subdomains, which have also been randomly selected.

In table 2.1 we show the number of sentences of the monolingual corpora. The corpora is composed by the corpus that organizers made available from the Scielo corpora, as well as the sentences we extracted from the Medline corpus that were not aligned.

Corpus	English	Spanish
Scielo - Biological	55,346	1,248
Scielo - Health	68,992	5,163
Medline	0	2,227

Table 2: Monolingual corpora

In addition to the in-domain corpora, we also included some other corpora available in other machine translation tasks inside the WMT challenge.

- Parallel Corpora:
  - Europarl[2] (Koehn, 2005): it is a corpus of parallel texts in 11 languages from the proceedings of the European Parliament. The version we used for this task has 2,218,201 English sentences and 2,123,835 Spanish sentences. For a direct alignment we excluded some of the sentences, obtaining 1,965,734 parallel sentences.
  - News commentary: this corpus consists in political and economic commentary crawled from the web site Project Syndicate[3]. It is composed of 247,966 sentences in English and 206,534 Spanish sentences. The parallel set has 174,441 sentences.
  - Common Crawl[4]: it is an open corpus of web crawl data. It has 1,845,286 parallel sentences for the English-Spanish language pair.
- Monolingual Corpora:
  - News Crawl (articles from 2007 to 2012): these 6 corpora (one per year) are articles extracted from various online news publications. In total, the English corpus has 68,521,621 sentences and the Spanish one 13,384,314.

2.2 Terminological Resources

SNOMED CT (***S**ystematized **N**omenclature of **Med**icine – **C**linical **T**erms*) (IHTSDO, 2014) is considered the most comprehensive, multilingual clinical health care terminology in the world. The use of a standard clinical terminology improves the quality of health care by enabling consistent representation of meaning in an electronic health record[5].

SNOMED CT provides the core terminology for electronic health records and contains more than 296,000 active concepts with their descriptions organized into hierarchies. (Humphreys et al., 1997) shows that SNOMED CT has an acceptable coverage of the terminology needed to record patient conditions. Concepts are defined by means of description logic axioms and are also used to group terms with the same meaning. Those descriptions are more generally considered as terms.

There are two types of descriptions in SNOMED CT: Fully Specified Names (FSN) and Synonyms. Fully Specified Names are the descriptions used to identify the concepts and they have a semantic tag in parenthesis that indicates its semantic type and, consequently, its hierarchy. Those are descriptions to unambiguously identify the concept, and they are not proper terms you can find in texts.

[2] http://www.statmt.org/europarl/

[3] https://www.project-syndicate.org/

[4] http://commoncrawl.org/

[5] http://www.ihtsdo.org/snomed-ct/whysnomedct/snomedfeatures/

Synonyms are the ones used in real texts, and SNOMED CT distinguishes between Preferred Terms and Acceptable Synonyms. As the name indicates, the Preferred Terms are the ones preferred and there is one defined for each concept in each language or dialect. In addition, there are as many Acceptable Synonyms as needed.

3 Systems

In this section we describe the systems we developed for the Biomedical Translation Task. First, we describe the baseline system. Then, we continue describing the three approaches we presented to the task, that deal with the most frequent issues of domain adaptation.

The baseline system has not been submitted for the Shared Task, and it is a reference system.

3.1 System 0: Baseline

To create our baseline system we trained a Moses statistical machine translation system on the corpora made available for the WMT Biomedical Translation Task, as well as and some general corpora publicly available on previous WMT tasks.

The system configuration is based on standard parameters: Tokenization, lowercasing and recasing using tools available in Moses toolkit, MGIZA for word alignment with the ”grow-diag-final-and” symmetrization heuristic, a maximum length of 80 tokens per sentence and 5 tokens per phrase, translation probabilities in both directions with Good Turing discounting, lexical weightings in both directions, a phrase length penalty, a ”msd-bidirectional-fe” lexicalized reordering model and a target language model. The weights for the different components were adjusted to optimize BLEU using Minimum Error Rate Training (MERT) with an n-best list of size 100.

For the Language Modeling we create a separate Language Model (LM) for each of the subcorpora we have available and interpolated all of them with the balanced development set extracted from the bilingual in-domain corpora. We must highlight that we used the monolingual corpora as well as the target language part of the bilingual corpora.

As we had too many LMs, we grouped them in the following way to train a hierarchical interpolation. The main criterion to generate the interpolation groups has been the source/domain of the corpora. That is, we grouped all the News corpora together, the Scielo Health bilingual and monolingual together, the Scielo Biological bilingual and monolingual as well, and in the case of Spanish we grouped also the Medline bilingual with the monolingual, and in the case of English we took the Medline bilingual on its own.

The same corpora used in the language model interpolation was used to optimize the weights of the different components of the statistical machine translation system. That is, the balanced development set explained in section 2.

3.2 System 1: SNOMED CT

The adaptation of a machine translation system to a specific domain has much to do with the specialized terminology that a general system lacks. This lack of terminology is related with the quantity of unknown words or out-of-vocabulary (OOV) words.

In this first approach we faced the lack of terminology by adding a widely recognized multilingual terminology bank to our translation system. More concretely, we included the terminological content of SNOMED CT’s English and Spanish International Releases to the system as parallel corpus.

As mentioned in section 2, SNOMED has many synonyms to name a concept. In this case, we aligned all the synonyms from the source language to the Preferred Term of the target language. Thus we avoid the generation of ambiguity as we do not have resources to solve it and we take advantage of the choice made by SNOMED CT developers.

Similarly, we also used the target language Preferred Terms to train a language model that was interpolated with the previous ones.

3.3 System 2: Morphology Variability and Transliteration

In the first system, we reduced the number of OOV words by adding a terminology bank to the training corpora. Even with such a large amount of specialized terms, the number of OOV words may not be zero, as the terminology used in texts is even wider. So, we developed a module to extend the phrase tables.

We enlarged the generated phrase tables in two ways: morphology variability of the plural or feminine words to the canonical form (singular and masculine) and transliteration of the remaining words.

In regard to morphology variability, we implemented a script that checks whether the OOV word is a morphology variation of the canonical form of a term that appears in the phrase table. In the case of English words, the process is as simple as making singular the plural forms and look for the translation candidates of the singular form in the phrase table. In order to avoid inconsistencies, we extracted only the translation candidates which are also made up by a single word, and we convert them into plural.

In contrast, the Spanish morphology made the process more complex, as in addition to the number variability, we must also take into consideration the gender of the words, and even the combination of both (feminine and plural).

With respect to transliteration, (Callison-Burch et al., 2006) exposes that state-of-the-art systems usually apply two strategies to cope with OOV words, neither of them satisfactory. In the first strategy the unknown word is omitted and in the second one it is not translated. The first strategy is even excluded as solution in (Habash, 2008), because the author considers it a trick to score better precision in evaluation metrics. Nevertheless, the second approach can be a good strategy whenever the OOV word is a Named Entity, such as a proper name or an organization name. Otherwise some action is needed.

State-of-the-art shows many approaches for transliteration in machine translation, most of them based on statistical methods (Deselaers et al., 2009; Habash, 2008; Hermjakob et al., 2008; Rama and Gali, 2009).

In a previous work, we developed a system to automatically translate English medical neoclassical compounds such as “glaucoma” or “meningitis” into Basque (Perez-de Viñaspre and Oronoz, 2015). This translation system is based on affix translation and a transliteration module was also implemented. In this case, we adapted the transliteration module for the English-Spanish and Spanish-English pair for the neoclassical medical words as well as for the substances and pharmaceutical products.

The module was implemented using Foma, a free software tool to specify finite-state automata and transducers (Hulden, 2009).

3.4 System 3: Sub-domain Optimization

The organizers of the Shared Task gave two test sets for the evaluation of the systems. One of the sets corresponded to the health domain and the other to the biological domain.

Taking that into account, we optimized the System 2 to each of the sub-domains.

As explained in System 0, the optimization of the system may be done in two levels: interpolation of the Language Model and the tuning of the weights of the different statistical machine translation components in MERT.

In the interpolation of the LM we maintained the groups done for the previous systems and we changed the interpolation corpus. In this case we replaced the balanced development set with the sub-domain tuning development set of each subdomain. That is, for the LM for health, we used the health tuning development set of health, and similarly for the biological LM, the set of biology.

Likewise, we replaced the same sets in the tuning of the whole statistical machine translation system.

4 Results

In this section we provide the results given by the organizers that measures the BLEU score of the systems submitted as the test sets are not publicly available yet. Each team was allowed to submit up to 3 runs per test file, in our case, 3 runs for the biological test sets from English to Spanish and vice versa, and 3 runs for each of the health test sets. We submitted the Systems 1, 2 and 3, and, therefore, the System 0 remained out of the evaluation.

Table 4 shows the BLEU results of the three systems we submitted for the four test sets. The results of the remaining systems have not been published yet, so we can not compare our systems to the others. In any case, we can compare them with the average of all the runs submitted for the language pair for each sub-domain.

System	Biological		Health	
	en-es	es-en	en-es	es-en
System 1	**31.57**	**30.66**	28.09	27.96
System 2	31.32	30.59	28.06	27.97
System 3	29.61	29.51	**28.13**	**28.12**
Average	31.34	30.17	28.3	27.79

Table 3: BLEU results of our systems.

The results obtained do not show any significant improvement of the different systems and in general we are close to the average. We obtained a small improvement from the average in three of the sets and we are very close to it in the fourth one (English to Spanish translation on the Health domain).

If we consider each System on its own, we can conclude that the System 2 does not give any advantage on what BLEU results regards as it decreases the results of the first system in most of the cases. In any case, we will need to check the manual evaluation that will be published in the overview paper to be sure about this conclusion.

In the case of the Biological sets, in both language pairs the best system seems to be the first one, as it outperformed the System 3 in one BLUE point and is above the average. On the contrary, the Health sets show that the last system improves a bit the results but nothing significant.

5 Conclusions

We present the IXA system for the Biomedical Translation Task from the WMT16 challenge which meets all the requirements established by the organizers. We implemented a system that translates biological and health science text from English to Spanish and Spanish to English.

We used all the corpora offered by the organizers as well as more corpora available for other tasks. In addition, we included a widely recognized multilingual terminology called SNOMED CT and a transliteration module that also solved the morphological variability of non-canonical words (plurals and feminines).

Our systems showed to be close to the average of all the submitted systems, and in three out of four of the cases even above the average. Overall we are pleased with the results even if we are surprised with the lack of improvement shown by the second system. We would like to try a new run training the optimization system based on the first system that only extends the OOV words with the terminology from SNOMED CT, so the optimization may be better on overall results.

The organizers will provide more details and additional results in the WMT'16 overview paper, such as manual evaluation of the runs submitted.

References

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved Statistical Machine Translation Using Paraphrases. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 17–24, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thomas Deselaers, Saša Hasan, Oliver Bender, and Hermann Ney. 2009. A deep learning approach to machine transliteration. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 233–241. Association for Computational Linguistics.

Nizar Habash. 2008. Four Techniques for Online Handling of Out-of-vocabulary Words in Arabic-English Statistical Machine Translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, pages 57–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ulf Hermjakob, Kevin Knight, and Hal Daumé III. 2008. Name Translation in Statistical Machine Translation-Learning When to Transliterate. In *ACL*, pages 389–397.

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

Betsy L Humphreys, Alexa T McCray, and May L Cheh. 1997. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. *Journal of the American Medical Informatics Association*, 4(6):484–500.

International Health Terminology Standards Development Organisation IHTSDO. 2014. SNOMED CT Starter Guide. February 2014. Technical report, International Health Terminology Standards Development Organisation.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, 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.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

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

Taraka Rama and Karthik Gali. 2009. Modeling Machine Transliteration As a Phrase Based Statistical Machine Translation Problem. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, NEWS '09, pages 124–127, Stroudsburg, PA, USA. Association for Computational Linguistics.