Machine Translation of the Penn Treebank to Spanish

Martha Alicia Rocha[1], *Joan Andreu Sánchez*[2]

[1] Departamento de Sistemas y Computación, Instituto Tecnológico de León, México
[2]Instituto Tecnológico de Informática, Universidad Politécnica de Valencia, Spain
mrocha@dsic.upv.es, jandreu@dsic.upv.es

Abstract

In this work we explored the problem of translating the Penn Treebank corpus to Spanish. For this problem, we considered Phrase-based Machine Translation techniques. Given that there not exist parallel training data for this corpus, we used a large out-of-domain training data set, and a small "hight-quality" in-domain training data set. We studied simple and effective Domain Adaptation techniques that were used for other applications. We report experiments on a small test set of sentences manually translated from the Penn Treebank corpus.

Index Terms: Penn Treebank, Machine Translation, Domain Adaptation.

1. Introduction

The Penn Treebank corpus [1] is one of the most referred data sets that has been extensively used for different sort of Natural Language Processing problems, including but not limited to Language Modeling [2], Word Sense Disambiguation [3], PoS Tagging [4, 5], Statistical Parsing [6], Maximum Entropy techniques [5, 7], among others. Recently, it has been also successfully used for Language Modeling in Machine Translation (MT) [8].

In recent years, promising Syntax-based MT systems have been introduced [9]. This sort of systems may be benefited from the availability of parallel annotated corpus that conveys syntactic information in order to learn syntactic models. The very rich linguistic information that has the Penn Treebank corpus, that includes syntactic information, semantic information and PoS tags, makes it very interesting for Syntax-based MT. It seems clear that the availability of this corpus adequately translated would be of major interest for MT.

The translation of this corpus would be more useful as more perfect the translation was. Currently there exist powerful techniques for MT [10], and phrase-based MT approach is among the most popular [11, 12]. This approach uses automatic methods in order to learn the translation models from large parallel corpus. This technique has demonstrated to obtain moderate results for tasks of hight complexity [13].

Although interesting MT systems have been proposed in the last years, perfect translation could be just guaranteed after human supervision. However, reviewing the full translation of the Penn Treebank corpus would be a very expensive work. If the goal is to obtain perfect translations, other techniques like Computer-Assisted Translation (CAT) should be explored [14]. In this approach the user translates interactively a data set and the CAT system adapts on-line both the translation models and the search process. After some time, "hight-quality" translated data is available that can be used to change the models. This seems a quite natural scenario for translating "perfectly" the Penn Treebank corpus. But Adaptation Techniques should be also considered.

As a first approximation to this scenario, in this work we explored the feasibility of translating the Penn Treebank corpus to Spanish by means of phrase-based MT techniques. An important problem with this approach arises when "in-domain" parallel data is not available. In such case, several approaches have been explored, like Domain Adaptation techniques [15]. We show how simple adaptation techniques can be very effective when applied to the translation of the Penn Treebank corpus when.

This paper is organized as follows: next section reviews basic adaptation techniques in MT. Then, we describe how we intend to tackle the translation of the Penn Treebank to Spanish. Experiments are reported in Section 3, and some concluding remarks are given in Section 4.

2. Adaptation in MT

There exist different MT techniques, like word-based models [10], those that are based in finite-state models [16], syntax-based models [9], or phrase-based models [11, 12]. In this work we will focus on phrase-based MT.

In phrase-based MT, the source sentence is split into phrases and then a large phrase translation table that contains paired source-target phrases is used to translate the source sentence to a target sentence. Target sentences are filtered according to a language model. Each paired phrase entry (e, f) in the phrase table has associated several scores h_i: phrase translation probabilities, reordering models, lexical translation probabilities, etc. In the decoding process, hypotheses are recombined in a log-linear model and the best-scoring translation is searched according to expression:

$$score(e, f) = \exp \sum_i \lambda_i h_i(e, f). \quad (1)$$

The weights λ_i associated to each component h_i are usually adjusted with a discriminative method on development data [17] in order to optimize a standard metric, like for example the BLEU metric [18]. The most important components in expression (1) are the translations models and the language model.

The language model component in expression (1) is usually estimated from target monolingual corpora. The parameters of translations models in expression (1) are usually trained from parallel corpora by using word alignment techniques [10]. Once both models are estimated, a test set is translated by using a decoder system. Usually, both the training data set and the test data set belong to the same domain.

When parallel data from the application domain is not available, Domain Adaptation (DA) techniques may be considered to obtain good results. The basic idea in DA is to adapt the models

trained with parallel data of one domain to a different domain. We now summarize some DA techniques.

Different DA scenarios can be described depending on the availability of *in-domain* training data. In *cross-domain* adaptation, a small sample of parallel in-domain text is available. This small parallel in-domain data is used for adapting the models. Another possible scenario is when no data is available ahead of time and is generated dynamically. In such case, *dynamic* adaptation techniques can be used. This second scenario is defined for Computer-Assisted Translation [14], in which the translation is a carried out *on-line* by a human expert. The system adapts dynamically to the user corrections. Note that in this situation the corrections introduced by the human have an added value since has been validated and can be considered as "hight quality" translations.

Other DA techniques have been defined for taking profit of latent information that could be present in the training data. Thus, Mixture Modeling was studied in [19], and Mixture-Model Adaptation was studied in[20]. The main advantage of those techniques is their capability to learn specific probability distributions that better fit subsets of training data set.

Simple and effective DA techniques were studied in [15] for a phrase-based MT system. The basic idea was to combine both the out-of-domain language model and the in-domain language model as separate components in expression (1) or to merge all data an to obtain an unique component. Something similar was carried out for the translations models.

2.1. Translation of the Penn Treebank

It should be noted that for the translation of the Penn Treebank corpus to Spanish, we intended to obtain "hight-quality" translations of the corpus, and therefore, a final supervision of an expert human would be appropriate.

In this scenario, no parallel text was available, and therefore the appropriate technique would be the one previously described in which the system adapts dynamically the models according to the corrections introduced by the human expert. However, the human cost of this translation could be very large. Therefore, it makes sense to try first other approaches less expensive.

The approach that we considered in this work was similar to [15]. In that work, an in-domain corpus was used for DA. Both the out-of-domain and the in-domain data sets were combined in different ways in order to improve translation results. The in-domain data set was not so large as the out-of-domain, but enough training data was available.

However, for the Penn Treebank there was not any sort of training data and just a small portion was manually translated. This small data was used to improve the results obtained from a standard baseline system. In this way the approach followed in this work can be considered as a combination of two approaches: DA adaptation from in-domain data together with "hight-quality" translations.

3. Experiments

The out-of-domain corpus used in the experiments was the *Europarl* corpus [13]. This is a set of parallel texts that is free available for several languages including English and Spanish. This corpus was built from the proceedings of European Parliament, which are published on the web. For our experimentation, we used the second version of this corpus [21]. This corpus is divided into four separate sets: training, development, development test, and final test. For this experiments we used only the sentences of training set to length 40 words. We called this set EU corpus. The main characteristics of this training set can be seen in Table 1.

Table 1: Characteristics of the Europarl (EU) corpus.

Sentence Pair	730,740
Running words Spa.	15,702,800
Running words Eng.	15,242,854
Vocabulary Spa.	102,821
Vocabulary neg.	64,076

As we have described in previous section, we prepared a small data set to be used as in-domain data set. For this work, we translated manually the first 300 sentences from section 23 of the Penn Treebank. We called this set *Small Parallel* Penn Treebank (SPPT) set. This is usually the section used for testing. Note that although this is a very small data set, they were manually translated, and could be considered as a "hight quality" translation set. Note also that this small corpus tried to simulate the CAT translation scenario that was previously described. Two Spanish native speakers independently translated the set of sentences. Then, each of them reviewed the translations of the other person, and finally they reached an agreement when different translations were proposed for each sentence. The main characteristics of this data set can be seen in Table 2. It is important to note that the relation between the number of running words in Spanish and the vocabulary size was 3.7. The same relation for the EU corpus was 152.8. This reveals that a great number of words of SPPT corpus could be singletons.

From SPPT corpus, 50 sentences were used for development, 100 sentences were used for test, and 150 sentences were used for training. The sentences to be included in each of these three sets considerable affected the results, as we describe below.

Table 2: Characteristics of the *Small Parallel* Penn Treebank (SPPT) corpus.

Sentence Pair	300
Running words Spa.	6,109
Running words Eng.	5,689
Vocabulary Spa.	1,664
Vocabulary Eng.	1,498

Standard free software tools were used for the experiments. For training of the language models, we used SRILM toolkit[1]. In all the experiments, 5-gram models trained with default options were used as language models. For training translation models, we used GIZA++[2] [22] with default options. Default translation models (h_i components in expr. (1)) were used. Finally, MOSES[3] [23] was used for decoding. The parameters of the model were tuned with MERT technique by improving BLEU metric.

In a similar way to [15], several systems were trained, each with a different way of combining the information of the two corpora. The different combinations were the following:

[1] http://www.speech.sri.com/projects/srilm/
[2] http://www.fjoch.com/GIZA++.html
[3] http://www.statmt.org/moses/

- **B:** baseline system. This systems was trained only with the EU corpus. It corresponds to situation in which there is not adaptation data available.
- **B+M50:** in this system, the parameters of the baseline system were adjusted with MERT on a development set composed of 50 sentences of SPPT corpus.
- **B+M50+TW:** a second translation table that was trained with 150 sentences from SPPT was added to previous system.
- **B+M50+TW+LW:** a second language model that was trained with 150 sentences from SPPT was added to previous system.
- **B+M50+TWMerg:** both EU set and 150 training sentences from SPPT were merged and then only a translation table was obtained. 50 tuning sentences of SPPT were used for MERT.
- **B+M50+TWMerg+LWMerg:** both EU set and 150 training sentences from SPPT were merged, and then only a translation table and only a language model were obtained. 50 sentences of SPPT were used for MERT.

Table 3 summarizes this information.

Table 3: Data sets and number of sentences used in the translation systems.

System	Tr. set	Dev. set	Test set
B	EU	-	SPPT 100
B+M50	EU	SPPT 50	SPPT 100
B+M50+TW B+M50+LW+TW B+M50+TWMerg B+M50+TWMerg +LWMerg	EU+ SPPT 150	SPPT 50	SPPT 100

In order to simulate CAT approach mentioned in Section 3, as a first attempt we chose the development, training, and test sets of SPPT as follows. Consecutive sentences for each set were chosen. Thus, sentences from 1 to 50 were used for development (SPPTdev set), sentences from 51 to 150 were used for test (SPPTts set), and sentences from 151 to 300 were used for training (SPPTtr set). Table 4 shows the obtained results for the translation systems that have previously described.

In this first experiment, we observed that the best translation results were obtained by the baseline system, and adjusting the translation with MERT did not achieve to improve this result. We thought that this could be due to the fact that the chosen test set included a lot of out-of-vocabulary words that prevented the system to improve the baseline result.

We tested this hypothesis in the following way. First, we trained a system with SPPTtr set, and then we translated independently both SPPTts and SPPTtr (close vocabulary in the last case), without MERT, and with MERT with SPPTdev and SPPTtr, both again independently. Table 5 shows the BLEU obtained results.

The large difference between results in column SPPTtr and column SPPTts revealed that the vocabulary of SPPTtr and SPPTts was quite different. Note also that when we used SPPTdev for MERT, the BLEU for SPPTtr decreased more than 10 points. Note in addition that when we used SPPTtr for MERT (over-learning the SPPTtr set), the BLEU decreased for SPPTts.

Table 4: BLEU, Word Error Rate (WER) and Translation Error Rate (TER) obtained for the translation systems.

System	BLEU	WER	TER
B	25.5	56.4	53.4
B+M50	22.1	61.4	58.3
B+M50+TW	17.6	64.1	62.8
B+M50+LW+TW	22.8	60.5	56.8
B+M50+TWMerg	23.3	57.8	55.0
B+M50+TWMerg +LWMerg	23.6	57.7	54.7

Table 5: BLEU results for the experiment with SPPT corpus, when we translated SPPTts and SPPTtr (closed vocabulary in the last case).

		SPPTtr	SPPTts
Without MERT		71.1	16.8
With MERT	SPPTdev	60.8	16.9
	SPPTtr	76.0	13.2

This experiment was interesting, because prevented us to use cross-validation in the experiment with SPPT corpus.

Therefore, we repeated the experiments but choosing each sentence in SPPTdev, SPPTtr, and SPPTts randomly. We tried several seeds and we chose the partition that provided the best results. Table 6 shows the obtained results. In this partition, the out-of-vocabulary words of the Spanish side of SPPTts with regard to the Spanish side of the EU set was 91 words, and this value decreased to 53 when we merged both the EU set and the Spanish side of SPPTtr.

Table 6: BLEU, Word Error Rate (WER) and Translation Error Rate (TER) obtained for the translation systems with SPPTdev, SPPTtr, and SPPTts randomly generated.

System	BLEU	WER	TER
B	20.8	61.9	58.2
B+M50	22.2	61.5	58.1
B+M50+TW	18.3	68.8	65.1
B+M50+LW+TW	18.6	65.0	61.6
B+M50+TWMerg	23.4	58.5	54.2
B+M50+TWMerg +LWMerg	23.1	59.7	55.8

In this experiment, we can see that some of the proposed systems improved the baseline results. We can also see that for this task and with this amount of data is better to merge the training data than combine them in two separate table and two separate language models as opposite to the results reported in [15]. This could be due to the fact that with the small amount of training data available for SPPT, the weights of the log-lineal model could not be better tuned.

Table 7 shows some translation results. Section 23 of the Penn Treebank includes a lot of sentences closely related to the exchange stock market and real state companies.

Table 7: Translation example

Source	kaufman & broad , a home building company , declined to identify the institutional investors .
System output	kaufman & broad , un hogar edificio compañía , desciende identificar a los inversores institucionales .
Reference	kaufman & broad , una compañía de construcción de casas , declinó identificar los inversores institucionales .

4. Conclusions

We have presented a work for the translation to Spanish of the Penn Treebank. A phrase-based model has been used for this task. Out-of-domain training data was used, together a very small "hight-quality" in-domain training set, simulating a CAT system. The obtained results cleary improved when "hight-quality" in-domain training data was included in the system. For future work we intend to use CAT systems for the same problem.

5. Acknowledgements

This work has been partially supported by the EC (FEDER) and the Spanish MEC under grant TIN2006-15694-CO2-01, and by the Spanish research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018). The first author is supported by "División de Estudios de Posgrado e Investigación" and by "Metrología y Sistemas Inteligentes" research group of Instituto Tecnológico de León.

6. References

[1] M. Marcus, B. Santorini, and M. Marcinkiewicz, "Building a large annotated corpus of english: the penn treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[2] B. Roark, "Probabilistic top-down parsing and language modeling," *Computational Linguistics*, vol. 27, no. 2, pp. 249–276, 2001.

[3] D. Bikel, "A statistical model for parsing and word-sense disambiguation," in *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, 2000, pp. 155–163.

[4] E. Brill, "Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging," *Comput. Linguist.*, vol. 21, no. 4, pp. 543–565, 1995.

[5] A. Ratnaparkhi, "A maximum entropy model for part-of-speech tagging," in *Proc. Empiriral Methods in Natural Language Processing*, University of Pennsylvania, May 1996, pp. 133–142.

[6] M. Collins, "Head-driven statistical models for natural language parsing," *Computational Linguistics*, vol. 29, no. 4, pp. 589–637, 2003.

[7] E. Charniak, "A maximum-entropy-inspired parser," in *Proc. of NAACL-2000*, 2000, pp. 132–139.

[8] E. Charniak, K. Knight, and K. Yamada, "Syntax-based language models for statistical machine translation," in *Proc. of MT Summit IX*, New Orleans, USA, September 2003.

[9] D. Chiang, "Hierarchical phrase-based translation," *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, 2007.

[10] P. Brown, S. D. Pietra, V. D. Pietra, and R. Mercer, "The mathematics of statistical machine translation: parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[11] R. Zens, F. Och, and H. Ney, "Phrase-based statistical machine translation," in *Proc. of the 25th Annual German Conference on Artificial Intelligence, LNAI, 2479*, 2002, pp. 18–32.

[12] P. Koehn, "Pharaoh: a beam search decoder for phrase-based statistical machine translation models," in *Proc. of AMTA*, 2004.

[13] ——, "Europarl: A parallel corpus for statistical machine translation," in *Proc. of MT Summit*, 2005.

[14] S. Barrachina, O. Bender, F. Casacuberta, J. Civera, E. Cubel, S. Khadivi, A. L. H. Ney, J. Tomás, and E. Vidal, "Statistical approaches to computer-assisted translation," *Computational Linguistics*, vol. 35, no. 1, pp. 2–28, 2009.

[15] P. P. Koehn and J. Schroeder, "Experiments in domain adaptation for statistical machine translation," in *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 224–227.

[16] F. Casacuberta and E. Vidal, "Machine translation with inferred stochastic finite-state transducers," *Computational Linguistics*, vol. 30, no. 2, pp. 205–225, 2004.

[17] F. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan: Association for Computational Linguistics, July 2003, pp. 160–167.

[18] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. 40th Annual meeting of the ACL*, 2002, pp. 311–318.

[19] J. Civera and A. Juan, "Domain adaptation in statistical machine translation with mixture modelling," in *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 177–180.

[20] G. Foster and R. Kuhn, "Mixture-model adaptation for SMT," in *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 128–135.

[21] P. Koehn and C. Monz, Eds., *Proceedings on the Workshop on Statistical Machine Translation*. New York City: Association for Computational Linguistics, June 2006.

[22] F. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–52, 2003.

[23] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proc. of the ACL Companion Volume Proceedings of the Demo and Poster Sessions*, June 2007, pp. 177–180.