Statistical Post-Editing: A Valuable Method in Domain Adaptation of RBMT Systems for Less-Resourced Languages

A. Diaz de Ilarraza, G. Labaka, K. Sarasola
Ixa taldea
University of the Basque Country
{jipdisaa, jiblaing, jipsagak}@ehu.es

Abstract

We present two experiments with Basque to verify the improvement obtained for other languages by using statistical post editing. The small size of available corpora and the use a morphological component in both RBMT and SMT translations make different our experiments from hose presented for similar works. Our results confirm the improvements when using a restricted domain, but they are doubtful for more general domains.

1 Introduction

Corpus based MT systems base their knowledge on aligned bilingual corpora, and the accuracy of their output depends heavily on the quality and the size of these corpora. When the two languages used in translation have very different structure and word order, the corpus needed to obtain similar results should be bigger.

Basque is a highly inflected language with free constituent order. Its structure and word order is different compared with languages as Spanish, French or English.

Being Basque a lesser used language, nowadays large and reliable bilingual corpora are unavailable. At present, domain specific translation memories for Basque are not bigger than two-three millions words, so they are still far away from the size of the corpora used for other languages; for example, Europarl corpus (Koehn, 2005), that is becoming a quite standard corpus resource, has 30 million words. So, although domain restricted corpus based MT for Basque shows promising results, it is still not ready for general use.

Moreover, the Spanish>Basque RBMT system Matxin's performance, after new improvements in 2007 (Alegria et al., 2007), is becoming useful for content assimilation, but it is still not suitable enough to allow unrestricted use for text dissemination.

Therefore, it is clear that we should experiment combining our basic approaches for MT (rule-based and corpus-based) to get a better performance. As the first steps on that way, we are experimenting with two simple alternative approaches to combining RBMT, SMT and EBMT:

- Selecting the best output in a multi engine system combining RBMT, EBMT and SMT approaches. (Alegría, et al., 2008)
- Statistical post-editing (SPE) on RBMT systems.

This paper deals with the second approach, where significant improvements have been recently published (Dugast et al., 2007; Ehara, 2007; Elming, 2006; Isabelle et al., 2007; Simard et al., 2007a and 2007b).

We don't have large corpus on post editing for Basque as proposed in (Isabelle et al., 2007), because our RBMT system has recently been created. However, we could manage to get parallel

corpus on some domains with a few million of words,

We will show that the issue of domain adaptation of the MT systems for Basque can be performed via the serial combination of a vanilla RBMT system and a domain specific statistical post-editing system even when the training corpus is not very big (half a million words). Unfortunately, we could not show that RBMT+SPE combination improves the result of RBMT systems when the corpus used is not related to a restricted domain.

The rest of this paper is arranged as follows: In section 2, we position the present work with respect to our ongoing research on SMT and SPE. In section 3 we present the corpora that will be used in our experiments. Section 4 describes the basic RBMT and statistical translation systems. In section 5, we report on our experiments comparing translation results under a range of different MT conditions: SMT versus RBMT, RBMT+SPE versus RBMT, and RBMT+SPE versus SMT. We finish this paper with some conclusions and future work.

2 Related work

In the experiments related by (Simard et al., 2007a) and (Isabelle et al., 2007) SPE task is viewed as translation from the language of RBMT outputs into the language of their manually post-edited counterparts. So they don't use a parallel corpus created by human translation. Their RBMT system is SYSTRAN and their SMT system PORTAGE. (Simard et al., 2007a) reports a reduction in post-editing effort of up to a third when compared to the output of the rule-based system, i.e., the input to the SPE, and as much as 5 BLEU points improvement over the direct SMT approach. (Isabelle et al., 2007) concludes that such a RBMT+SPE system appears to be an excellent way to improve the output of a vanilla RBMT system and constitutes a worthwhile alternative to costly manual adaptation efforts for such systems. So a SPE system using a corpus with no more than 100.000 words of post-edited translations is enough to outperform an expensive lexicon enriched baseline RBMT system.

The same group recognizes (Simard et al., 2007b) that this sort of training data is seldom available, and they conclude that the training data for the post-editing component does not need to be manually post-edited translations, that can be generated even from standard parallel corpora. Their new RBMT+SPE system outperforms both the RBMT and SMT systems again. The experiments show that while post-editing is more effective when little training data is available, it remains competitive with SMT translation even when larger amounts of data. After a linguistic analysis they conclude that the main improvement is due to lexical selection.

In (Dugast et al., 2007), the authors of SYSTRAN's RBMT system present a huge improvement of the BLEU score for a SPE system when comparing to raw translation output. They get an improvement of around 10 BLEU points for German-English using the Europarl test set of WMT2007.

(Ehara, 2007) presents two experiments to compare RBMT and RBMT+SPE systems. Two different corpora are issued, one is the reference translation (PAJ, Patent Abstracts of Japan), the other is a large scaled target language corpus. In the former case, RBMT+SPE wins, in the later case RBMT wins. Evaluation is performed using NIST scores and a new evaluation measure NMG that counts the number of words in the longest sequence matched between the test sentence and the target language reference corpus.

Finally, (Elming, 2006) works in the more general field called as Automatic Post-Processing (APE). They use transformation-based learning (TBL), a learning algorithm for extracting rules to correct MT output by means of a post-processing module. The algorithm learns from a parallel corpus of MT output and human-corrected versions of this output. The machine translations are provided by a commercial MT system, PaTrans, which is based on Eurotra. Elming reports a 4.6 point increase in BLEU score.

3 The corpora

Our aim was to improve the precision of the MT

system trying to translate texts from a restricted domain. We were interested in a kind of domain where a formal and quite controlled language would be used and where any public organization or private company would be interested in automatic translation on this domain. We also wanted to compare the results between the restricted domain and a more general domain such as news.

Specific domain: Labor Agreements Corpus

The domain related to *Labor Agreements* was selected. The Basque Institute of Public Administration (IVAP[1]) collaborated with us in this selection, after examining some domains, available parallel corpora and their translation needs. The Labor Agreements Corpus is a bilingual parallel corpus (Basque and Spanish) with 640,764 words for Basque and 920,251 for Spanish. We automatically aligned it at sentence level and then manual revision was performed.

To build the test corpus the full text of several labor agreements was randomly chosen. We chose full texts because we wanted to ensure that several significant but short elements as the header or the footer of those agreements would be represented. Besides it is important to measure the coverage and precision we get when translating the whole text in one agreement document and not only those of parts of it. System developers are not allowed to see the test corpus.

In SMT we use the training corpus to learn the models (translation and language model); the development corpus to tune the parameters; and the test corpus to evaluate the system.

The size of each subset is shown in Table 1.

		Sentences	Words
Training	Spanish	51,740	839.393
	Basque		585,361
Development	Spanish	2,366	41,508
	Basque		28,189
Test	Spanish	1,945	39,350
	Basque		27,214

Table 1. Statistics of Labor Agreements Corpus

1 http://www.ivap.euskadi.net

General domain: Consumer Eroski Corpus

As general domain corpus, we used the *Consumer Eroski* parallel corpus. The *Consumer Eroski* parallel corpus is a collection of 1,036 articles written in Spanish (January 1998 to May 2005, Consumer Eroski magazine, http://revista.consumer.es) along with their Basque, Catalan, and Galician translations. It contains more than one million Spanish words for Spanish and more than 800,000 Basque words. This corpus is aligned at sentence level.

In order to train the data-driven systems (both SMT and SPE systems), we used approximately 55,000 aligned sentences extracted from the Consumer dataset. Two additional sentence sets are used; 1501 sentences for parameter tuning and 1515 sentences for evaluation (see Table 2).

		Sentences	Words
Training	Spanish	54,661	1,056,864
	Basque		824350
Development	Spanish	1,501	34,333
	Basque		27,235
Test	Spanish	1,515	32,820
	Basque		34,333

Table 2. Statistics of Consumer Eroski corpus

4 Basic translation systems

Rule based system: Matxin

In this subsection we present the main architecture of an open source MT engine, named *Matxin* (Alegria et al., 2007). the first implementation of Matxin translates from Spanish into Basque using the traditional transfer model and based on shallow and dependency parsing.

Matxin is a classical transfer system consisting of three main components: (i) analysis of the source language into a dependency tree structure, (ii) transfer from the source language dependency tree to a target language dependency structure, and (iii) generation of the output translation from the target dependency structure. These three components are described in more detail in what follows.

The analysis of the Spanish source sentences into dependency trees is performed using an adapted version of the Freeling toolkit (Carreras et al., 2004). The shallow parser provided by Freeling is augmented with dependency information between chunks.

In the transfer module the Spanish analysis tree is transformed into Basque dependency tree. In this step, a very simple lexical selection is carried out, the Spanish lemma is translated by most frequent equivalent.

Finally, the dependency tree coming from the transfer module is passed on the generation module, in order to get the target language sentence. The order of the words in the final sentence is decided and morphological generation is carried out when it is necessary (in Basque: the declension case, the article and other features are added to the whole noun phrase at the end of the last word). We reused a previous morphological analyzer/generator developed for Basque (Alegria et al., 1996) adapted and transformed to our purposes.

Corpus based system

The corpus-based approach has been carried out in collaboration with the National Center for Language Technology in Dublin City University (DCU).

The system is based on a baseline phrase-based SMT system, but the dataset of aligned phrases is enriched with linguistically motivated phrase alignments. We have carried out Basque to English (Stroppa et al., 2006) and Spanish to Basque (Labaka et al., 2007) translation experiments.

Freely available tools are used to develop the SMT systems:

- GIZA++ toolkit (Och and H. Ney, 2003) is used for training the word/morpheme alignment.
- SRILM toolkit (Stolcke, 2002) is used for building the language model.
- Moses Decoder (Koehn et al., 2007) is used for translating the sentences.

Due to the morphological richness of Basque, when translating from Spanish to Basque some Spanish words, like prepositions or articles, correspond to Basque suffixes, and, in case of ellipsis, more than one of those suffix can be added to the same word. Example of concatenation of two case suffixes:

```
  puntuarenean =
= puntu  +  aren    +  ean  =
= point  +  of the  +  in the  =
= in the one(ellipsis) of the point
```

In order to deal with these features a morpheme-based SMT system was developed.

Adapting the SMT system to work at the morpheme level consists on training the basic SMT on the segmented text. The system trained on these data will generate a sequence of morphemes as output. In order to obtain the final Basque text, we have to generate words from those morphemes.

To get the segmented text, Basque texts are previously analyzed using Eustagger (Aduriz & Díaz de Ilarraza, 2003). After this process, each word is replaced with the corresponding lemma followed by a list of morphological tags. The segmentation is based on the strategy proposed on (Agirre et al., 2006).

Both systems (the conventional SMT system and the morpheme based), were optimized decoding parameters using a Minimum Error Rate Training. The metric used to carry out the optimization is BLEU.

The evaluation results for the general domain Consumer corpus (also used in this paper) are in Table 3. The morpheme based MT system gets better results for all the measures except BLEU.

	BLEU	NIST	WER	PER
SMT	**9.85**	4,28	82,72	63,78
Morpheme-based SMT	9,63	**4,43**	**80.92**	**62,27**

Table 3. Evaluation for SMT systems

RBMT and Statistical Post-Editing

In order to carry out experiments with statistical post-editing, we have first translated Spanish sentences in the parallel corpus using our rule-based translator (Matxin). Using these automatically translated sentences and their

corresponding Basque sentences in the parallel corpus, we have built a new parallel corpus to be used in training our statistical post-editor.
The statistical post-editor is the same corpus-based system explained before. This system is based on freely available tools but enhanced in two main ways:

- In order to deal morphological richness of Basque, the system works on morpheme-level, so a generation phase is necessary after SPE is applied.
- Following the work did in collaboration with the DCU, the phrases statistically extracted are enriched with linguistically motivated chunk alignments.

5 Results

We used automatic evaluation metrics to assess the quality of the translation obtained using each system. For each system, we calculated BLEU (Papineni et al., 2002), NIST (Doddington, 2002), Word Error Rate (WER) and Position independent Error Rate (PER).

Besides, our aim was to evaluate performance using different corpora types, so we tested the output of all systems applied to two corpora: one domain specific (Labor Agreements Corpus), and a general domain corpus (Consumer corpus).

	BLEU	NIST	WER	PER
Rule-based	4,27	2,76	89,17	74,18
Corpus-based	12,27	4,63	77,44	58,17
Rule-based + SPE	**17,11**	**5,01**	**75,53**	**57,24**

Table 4. Evaluation on domain specific corpus

Results obtained on the Labor Agreements Corpus (see Table 4) shows that the rule-based gets a very low performance (rule-based system is not adapted to the restricted domain), and the corpus-based system gets a much higher score (8 BLEU points higher, a 200% relative improvement). But if we combine both systems using the corpus-based system as a statistical post-editor, the improvement is even higher outperforming corpus-based system in 4.48 BLEU point (40% relative improvement).

	BLEU	NIST	WER	PER
Rule-based MT	6,78	3,72	81,89	66,72
Corpus-based MT	**9,63**	**4,43**	80,92	**62,27**
Rule-based + SPE	8,93	4,23	**80,34**	63,49

Table 5. Evaluation on general domain corpus

Otherwise, results on the general domain corpus (see Table 5) do not indicate the same. Being a general domain corpus, the vanilla rule-based system gets better results, and those approaches based on the corpus (corpus-based MT and RBMT+SPE) get lower ones. Furthermore, the improvement achieved by the statistical post-editor over the rule-based system is much smaller and it does not outperforms the corpus-based translator.

6 Conclusion

We performed two experiments to verify the improvement obtained for other languages by using statistical post editing. Our experiments differ from other similar works because we use a morphological component in both RBMT and SMT translations, and because the size of the available corpora is small.

Our results are coherent with huge improvements when using a RBMT+SPE approach on a restricted domain presented by (Dugast eta al., 2007; Ehara, 2007; Simard et al., 2007b). We obtain 200% improvement in the BLEU score for a RBMT+SPE system working with Matxin RBMT system, when comparing to raw translation output, and 40% when comparing to SMT system.

Our results also are coherent with a smaller improvement when using more general corpora as presented by (Ehara, 2007; Simard et al., 2007b).

We can not work with manually post-edited corpora as (Simard et al., 2007a) and (Isabelle et al., 2007) because there is no such a big corpus for Basque, but we plan to collect it and compare results obtained using a real post-edition corpus and the results presented here.

We also plan automatic extracting rules to

correct MT output by means of a post-processing module (Elming, 2006).

Acknowledgments

This work has been partially funded by the Spanish Ministry of Education and Science (OpenMT: Open Source Machine Translation using hybrid methods,TIN2006-15307-C03-01) and the Local Government of the Basque Country (AnHITZ 2006: Language Technologies for Multilingual Interaction in Intelligent Environments., IE06-185). Gorka Labaka is supported by a PhD grant from the Basque Government (grant code, BFI05.326).
Consumer corpus has been kindly supplied by Asier Alcázar from the University of Missouri-Columbia and by Eroski Fundazioa.

References

Aduriz, I. and Díaz de Ilarraza, A. (2003). Morphosyntactic disambiguation and shallow parsing in Computational Processing of Basque. In Inquiries into the lexicon-syntax relations in Basque. Bernard Oyharçabal (Ed.), Bilbao.

Alegria, I., Artola Zubillaga, X. and Sarasola, X. (1996). Automatic morphological analysis of Basque. Literary & Linguistic Computing 11(4):193—203.

Alegria I., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., Sarasola K. (2007) Transfer-based MT from Spanish into Basque: reusability, standardization and open source. Cicling.

Alegria, I., Casillas, A., Díaz de Ilarraza, A., Igartua, J., Labaka, G., Lersundi, M., Mayor, A., Sarasola, K., Saralegi, X., Laskurain, B. (2008). A Simple Mixing Approach to MT for Basque. To be presented in MATMT08 workshop: Mixing Approaches to Machine Translation. Donostia.

Agirre, E., Díaz de Ilarraza, A., Labaka, G., and Sarasola, K. (2006). Uso de información morfológica en el alineamiento Español-Euskara. In XXII congreso de la SEPLN, Zaragoza, Spain.

Carreras, X., Chao, I., Padró, L., Padró, M. (2004). FreeLing: An Open-Source Suite of Language Analyzers. In Proceedings of 4th LREC, Lisbon, Portugal.

Doddington, G. (2002). Automatic evaluation of Machine Translation quality using n-gram cooccurrence statistics. In Proceedings of HLT 2002, San Diego, CA.

Dugast, L., Senellart, J., & Koehn, P. (2007). Statistical post-editing on SYSTRAN's rule-based translation system. ACL 2007: proceedings of the Second Workshop on Statistical Machine Translation, 23, 2007, Prague, Czech Republic; pp. 220-223

Elming, J. (2006). Transformation-based correction of rule-based MT. 11th Annual Conference of the European Association for Machine Translation, Oslo, Norway.

Isabelle, P., Goutte, C., & Simard, M. (2007). Domain adaptation of MT systems through automatic post-editing. MT Summit XI, 10-14 September 2007, Copenhagen, Denmark. pp.255-261

Koehn, Ph. (2005). Europarl: A parallel corpus for statistical machine translation. Proc. of the MT Summit X ,pp. 79–86, September.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan N., Shen, W. Moran, C. Zens, R. Dyer, C. Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In Annual Meeting of the Association for Computational Linguistics (ACL), Prague, Czech Republic.

Labaka, G., Stroppa, N., Way, A. and Sarasola, K. (2007). Comparing Rule-based and Data-driven Approaches to Spanish-to-Basque Machine Translation. In Proceedings of the MT-Summit XI, Copenhagen, Denmark.

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

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of 40th ACL, Philadelphia, PA.

Simard, M., Goutte, C., and Isabelle, P.. (2007a). Statistical phrase-based post-editing. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pages 508–515, Rochester, USA, April. Association for Computational Linguistics.

Simard, M., Ueffing, N., Isabelle, P., & Kuhn, R.

(2007b). Rule-based translation with statistical phrase-based post-editing. ACL 2007: proceedings of the Second Workshop on Statistical Machine Translation, June 23, 2007, Prague, Czech Republic; pp. 203-206

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. In Proc. Intl. Conf. Spoken Language Processing, Denver, Colorado.

Stroppa, N., Groves, D., Way, A., and Sarasola, K. (2006). Example-base Machine Translation of the Basque Language. In Proceedings of AMTA 2006, pp. 232—241, Cambridge, MA.

Ehara Terumasa (2007). Rule based machine translation combined with statistical post editor for Japanese to English patent translation. MT Summit XI Workshop on patent translation, 11 September 2007, Copenhagen, Denmark; pp.13-18.