Mixing Approaches to MT for Basque: Selecting the best output from RBMT, EBMT and SMT

I. Alegria, A. Casillas, A. Diaz de Ilarraza, J. Igartua, G. Labaka,
M. Lersundi, A. Mayor, K. Sarasola
Ixa taldea. University of the Basque Country.
X. Saralegi
Elhuyar Fundazioa
B. Laskurain
Eleka S.L.
i.alegria@ehu.es

Abstract

We present the first steps in the definition of a mixing approach to MT for Basque based on combining single engines that follow to three different MT paradigms. After describing each engine we present the hierarchical strategy we use in order to select the best output, and a first evaluation.

1 Introduction and Basque Language

Basque is a highly inflected language with free order of sentence constituents.

It is an agglutinative language, with a rich flexional morphology. In fact for nouns, for example, at least 360 word forms are possible for each lemma. Each of the declension cases such as absolutive, dative, associative… has four different suffixes to be added to the last word of the noun phrase. These four suffix variants correspond to indefinite, definite singular, definite plural and "close" definite plural. Basque syntax and word order is very different compared with other languages as Spanish, French or English.

Machine translation is both, a real need, and a test bed for our strategy to develop NLP tools for Basque. We have developed corpus based and rule based MT systems, but they are limited.

On the one hand, corpus based MT systems base their knowledge on aligned bilingual corpora, and the accuracy their output depends heavily on the quality and the size of these corpora. When the pair of languages used in translation have very different structure and word order, obviously, the corpus needed should be bigger.

Being Basque a lesser resourced language, nowadays large and reliable bilingual corpora are unavailable for Basque. 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 present corpora for languages; e.g., Europarl corpus (Koehn, 2005), that is becoming a quite standard corpus resource, has 30 million words. So, the results obtained in corpus based MT to Basque are promising, but they are still not ready for public use.

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

Therefore it is clear that we should combine our basic hes for MT (rule-based and corpus-based) in order to build a hybrid system with better performance. As the first steps on that way, we are experimenting with two simple mixing alternative approaches used up to now for languages with huge corpus resources:

- Selecting the best output in a multi engine system (MEMT, Multi-engine MT), in our case combining RBMT, EBMT and SMT approaches.
- Statistical post-editing(SPE) after RBMT.

This paper deals with the first approach. Our design has been carried out bearing in mind the following concepts:

- Combination of MT paradigms.
- Reusability of previous resources, such as

translation memories, lexical resources, morphology of Basque and others.

- Standardization and collaboration: using a more general framework in collaboration with other groups working in NLP.
- Open-source: this means that anyone having the necessary computational and linguistic skills will be able to adapt or enhance it to produce a new MT system,

Due to the real necessity for translation in our environment the involved languages would be Basque, Spanish, French and English.

The first strategy we are testing when we want to build a MT engine for a domain, is translating each sentence using each of our three single engines (rule-based, example-based and statistical) and then choosing the best translation among them (see section 4).

In section 2 we present the corpus that we will use in our experiments, while in section 3 we explain the single engines built up for Basque MT following the three traditional paradigms: rule-based, example-based and statistical. In section 4, we report on our experiment to combine those three single engines. We finish this paper with some conclusions.

2 The corpus

Our aim was to improve the precision of the MT system trying to translate texts from a 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.

Finally the domain related to *labor agreements* was selected. The Basque Institute of Public Administration (IVAP[1]) collaborated with us in this selection, by examining some possible domains, parallel corpora available and their translation needs. The Labor Agreements Corpus is a bilingual parallel corpus (Basque and Spanish) with 585,785 words for Basque and 839,003 for Spanish. We automatically aligned it at sentence level and then manual revision was performed.

As said before, our goal is to combine different MT approaches: Rule-Based (RBMT), Example Based (EBMT) and Statistical (SMT). Once we had the corpus, we split it in three for SMT (training, development and test corpus) and in two for EBMT (development and test corpus).

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, and because it is important to measure the coverage and precision we get when translating the whole text in one agreement document and not only some sentences 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.

In RBMT and EBMT there are not parameters to optimize, and so, we consider only two corpora: one for the development (joining the training and development ones) and one for the test.

The size of each subset is shown in Table 1 (eu= Basque, es = Spanish).

		Doc	Sentences	Words
Training	es	81	51,740	839.393
	eu	81		585,361
Development	es	5	2,366	41,508
	eu	5		28,189
Test	es	5	1,945	39,350
	eu	5		27,214

Table 1. Labor Agreements Corpus

3 Single MT engines for Basque

In this section we present three single engines for Spanish-Basque translation following the three traditional paradigms: rule-based, example-based and statistical. The first one has been adapted to the domain corpus, and the other two engines have been trained with it.

3.1 The rule-based approach

In this subsection we present the main architecture of an open source MT engine, named *Matxin* (Alegria et al., 2007), the first implementation of which translates from Spanish into Basque using the traditional transfer model and based on shallow and dependency parsing. Later on, in a second step, we have specialized it to the domain.

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

The design and the programs of Matxin system are independent from the pair of languages, so the software can be used for other projects in MT. Depending on the languages included in the adaptation, it will be necessary to add, reorder and change some modules, but this will not be difficult because a unique XML format is used for the communication among all the modules.

The project has been integrated in the *OpenTrad*[2] initiative, a government-funded project shared among different universities and small companies, which include MT engines for translation among the main languages in Spain. The main objective of this initiative is the construction of an open, reusable and interoperable framework.

In the *OpenTrad* project, two different but coordinated architectures have been carried out:

- A shallow-transfer based MT engine for similar languages (Spanish, Catalan and Galician).
- A deeper-transfer based MT engine for the Spanish-Basque and English-Basque pair. It is named *Matxin* and it is stored in *matxin.sourceforge.net.* It is an extension of previous work in IXA group.

In the second engine, following the strategy of reusing resources, another open source engine, *FreeLing* (Carreras et al., 2004), was used for analysis.

The transfer module is divided into three phases dealing at the level of the three main objects in the translation process: words or nodes, chunks or phrases, and sentences.

- First, lexical transfer is carried out using a bilingual dictionary compiled into a finite-state transducer.
- Then, structural transfer at sentence level is applied, some information is transferred from some chunks to others, and some chunks may disappear. For example, in the Spanish-Basque transfer, person and number information of the object and the type of subordination are imported from other chunks to the chunk corresponding to the verb chain.
- Finally the structural transfer at chunk level is carried out. This process can be quite simple (e.g. noun chains between Spanish and Basque) or more complex (e.g. verb chains between these same languages).

The XML file coming from the transfer module is passed on the generation module.

- In the first step, syntactic generation is performed in order to decide the order of chunks in the sentence and the order of words in the chunks. Several grammars are used for this purpose.
- Morphological generation is carried out in the last step. In the generation of Basque, the main inflection is added to the last word in the phrase (in Basque, the declension case, the article and other features are added to the whole noun phrase at the end of the last word), but in verb chains other words need morphological generation. A previous morphological analyzer/generator for Basque (Alegria et al., 1996) has been adapted and transformed to the format used in *Apertium*.

	BLEU	Edit-distance TER
Corpus1 (newspapers)	9.30	40.41
Corpus2 (web magazine)	6.31	43.60

Table 2. Evaluation for the RBMT system

The results for the Spanish-Basque system using *FreeLing* and *Matxin* are promising. The quantitative evaluation uses the open source evaluation tool IQMT and figures are given using Bleu and NIST measures (Giménez et al., 2005). An additional user based evaluation has been carried out too, using Translation Error Rate (Snover, 2006). The results using two corpora without very long sentences are shown in Table 2 (Mayor, 2007).

We have to interpret the results having in mind that the development of this RBMT system was based on texts of newspapers.

Adaptation to the domain

The adaptation to the domain has been out in three main ways:

2 www.opentrad.org

- Terminology. Semiautomatic extraction of terminology using Elexbi, a bilingual terminology extractor for noun phrases (Alegria et al., 2006). Additionally, an automatic format conversion to the monolingual and bilingual lexicons is carried out for the selected terms. More than 1,600 terms were extracted from the development corpus, manually examined, and near to 807 were selected to be included in the domain adapted lexicon.
- Lexical selection. Matxin does not face the lexical selection problem for lexical units (Matxin only does it for the preposition-suffix translation); just the first translation in the dictionary is always selected (the other possible lexical translations are stored for the post-edition). For the domain adaptation, a new order for the possible translations has been calculated in the dictionary, based on the parallel corpus and using GIZA++.
- Resolution of format and typographical variants which are found frequently in the administrative domain.

After this improvements this engine is ready to process the sentences from this domain.

3.2 The example-based approach

In this subsection we explain how we automatically extract translation patterns from the bilingual parallel corpus and how we exploit it in a simple way.

Translation patterns are generalizations of sentences that are translations of each other in that various sequences of one or more words are replaced by variables (McTait, 1999).

Starting from the aligned corpus we carry out two steps to automatically extract translation patterns.

First, we detect some concrete units (entities mainly) in the aligned sentences and then we replace these units by variables. Due to the morphosyntactic differences between Spanish and Basque, it was necessary to execute particular algorithms for each language in the detection process of the units. We have developed algorithms to determine the boundaries of dates, numbers, named entities, abbreviations and enumerations.

After detecting the units, they must be aligned, to relate the Spanish and Basque units of the same type that have the same meaning. While in the case of numbers, abbreviations and enumerations the alignment is almost trivial, in the case of named entities, the alignment algorithm is more complex. It is explained in more detail in (Martinez et al., 1998). Finally, to align the dates, we use their canonical form.

Table 3 shows an example of how a translation pattern is extracted.

Once we have extracted automatically all the possible translation patterns from the training set, we store them in a hash table and we can use them in the translation process. When we want to translate a source sentence, we just have to check if that sentence matches any translation pattern in the hash table. If the source sentence matches a sentence of the hash table that has not any variable, the translation process will immediately return its translation. Otherwise, if the source sentence does not exactly match any sentence in the hash table, the translation process will try to generalize that sentence and will check again in the hash if it finds a generalized template. To generalize the source sentence, the translation process will apply the same detection algorithms used in the extraction process.

In a preliminary experiment using a training corpus of 54.106 sentence pairs we have extracted automatically 7.599 translation patterns at sentence level.

Aligned sentences	**Aligned sentences with generalized units**	**Translation pattern**
En Vitoria-Gasteiz, a 22 de Diciembre de 2003.	En <rs type=loc> Vitoria-Gasteiz </rs> , a <date date=22/12/2003> 22 de Diciembre de 2003</date> .	En <rs1> , a <date1>.
Vitoria-Gasteiz, 2003ko Abenduaren 22.	<rs type=loc> Vitoria-Gasteiz </rs> , <date date=22/12/2003> 2003ko Abenduaren 22</date>.	<rs1>, <date1>.

Table 3. Pattern extraction process

These translation patterns cover 35.450 sentence pairs of the training corpus. We also think that an aligned pair of sentences can be a transla-

tion pattern if it does not have any generalized unit but it appears at least twice in the training set.

As this example based system has a very high precision but quite low coverage (see Table 6 and Table 7), it is very interesting to combine with the other engines specially in this kind of domain where a formal and quite controlled language is used.

3.3 The SMT approach

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

The system exploits SMT technology to extract a dataset of aligned chunks. We have conducted Basque to English (Stroppa et al., 2006) and Spanish to Basque (Labaka et al., 2007) translation experiments, based on a quite large corpus (270,000 sentence pairs for English and 50,000 for Spanish).

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, in translation from Spanish to Basque some Spanish words, like prepositions or articles, correspond to Basque , and, in case of ellipsis, more than one of those suffixes can be added to the same word. In order to deal with this features a morpheme-based SMT system has been built.

Adapting the SMT system to work at 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 obtain the segmented text, Basque texts are previously analyzed using *Eustagger* (Aduriz and 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 in a quite general domain (for the same type of texts) are in Table 4.

	BLEU	NIST	WER	PER
SMT	9.51	3.73	83.94	66.09
morpheme-based SMT	8.98	3.87	80.18	63.88

Table 4. Evaluation for SMT systems

Details about the system and its evaluation can be consulted in (Díaz de Ilarraza et al., 2008).

4 Combining the approaches and evaluation

van Zaanen and Somers (2005) and Matusov et al. (2006) review a set of references about MEMT (Multi-engine MT) including the first attempt by Frederking and Nirenburg (1994), Macherey and Och (2007)

All those papers reach the same conclusion: combining the outputs results in a better translation.

Most of the approaches generate a new consensus translation using different language models. They have to train the system on those language models. Some of the approaches require confidence scores for each of the outputs. This approach is being used in several works (Macheret&Och, 2007; Sim et al., 2007), and some of them are used inside the GALE research program.

MEMT for Basque

Bearing in mind that huge parallel corpora for Basque are not available we decided to combine the different methods in a domain where translation memories were available.

Because confidence scores are not still available for the RBMT engine, we decided, for a first attempt, to combine the three approaches in a very simple hierarchical way: processing each sentence by the three engines (RBMT, EBMT and SMT) and then trying to choose the best

translation among them.

In a first step the text is divided into sentences, then each sentence is processed using each engine (parallel processing is possible). Finally one of the translations is selected.

In order to make this selection the facts we can deal with are the followings:

- Precision for the EBMT approach is very high, but its coverage low.
- The SMT engine gives a confidence score.
- RBMT translations are more adequate for human postedition than those of the SMT engine, but SMT gets better scores when BLEU and NIST are used with only one reference (Labaka et al., 2007).

	BLEU RBMT	BLEU SMT	HTER RBMT	HTER SMT
EiTB corpus (news)	9.30	9.02	40.41	71.87
Consumer (magazine)	6.31	8.03	43.60	57.97

Table 5. Evaluation using Bleu and HTER for RBMT and SMT (Labaka et al., 2007)

We can see in Table 5 that automatic evaluation (BLEU) with one reference and user-driven evaluation (HTER) yield different results.

Bearing this in mind, in this first attempt, we decided to apply a hierarchical strategy:

- If the EBMT engine covers the sentence its translation is selected.
- Else we chose the translation from the SMT engine if its confidence score is higher than a given threshold.
- Otherwise the output from the RBMT engine will be taken.

The results on the development corpus appear in Table 6.

The best results, evaluated using automatic metrics with only one reference, are obtained combining EBMT and SMT. But bearing in mind our previous evaluation trials with human translators (Table 5), we think that a deeper evaluation is necessary.

Table 7 shows the results on the test corpora.

	Coverage	BLEU	NIST
RBMT (domain adapted)	100%	7.97	3.21
SMT	100%	14.37	4.43
EBMT+RBMT	EBMT 42% RBMT 58%	26.85	5.15
EBMT+SMT	EBMT 42% SMT 58%	30.44	5.93
EBMT+SMT+ RBMT	EBMT 42% SMT 33% RBMT 25%	29.41	5.68

Table 6. Results for the MEMT system using the development corpus

	Coverage	BLEU	NIST
RBMT (domain adapted)	100%	5.16	3.08
SMT	100%	12.71	4.69
EBMT+RBMT	EBMT 58% RBMT 42%	26.29	5.40
EBMT+SMT	EBMT 58% SMT 42%	29.11	6.25
EBMT+SMT+ RBMT	EBMT 58% SMT 28% RBMT 14%	28.50	6.02

Table 7. Results for the MEMT system using the test corpus

5 Conclusions

We have presented a hierarchical strategy to select the best output from three MT engines we have developed for Spanish-Basque translation.

In this first attempt, we decided to apply a hierarchical strategy: First application of EBMT (translation patterns), then SMT (if its confidence score is higher than a given threshold), and then RBMT.

The results of the initial automatic evaluation showed very significant improvements. For example, 129% relative increase for BLEU when comparing. EBMT+SMT combination with SMT single system. Or 124% relative increase for BLEU when comparing. EBMT+SMT+RBMT combination with SMT single system.

Anyway the best results, evaluated using automatic metrics with only one reference, are obtained combining just EBMT and SMT.

The consequence of the inclusion of a final RBMT engine (to translate just the sentences not covered by EBMT and with low confidence score

for SMT) has a small negative contribution of 2% relative decrease for BLEU. But based on previous evaluations we think that a deeper evaluation based on human judgements is necessary.

For the near future we plan to carry out new experiments using combination of the outputs based on a language model. We are also plan defining confidence scores for the RBMT engine (penalties when suspicious or very complex syntactic structures are present in the analysis, penalties for high proportion of ignored word senses, promoting translations that recognize multiword lexical units, …)

Acknowledgments

This work has been partially funded by the Spanish 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 Multingual Interaction in Intelligent Environments., IE06-185). Gorka Labaka is supported by a PhD grant from the Basque Government (grant code, BFI05.326)

Reference

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

Agirre, E., D´ de Ilarraza, A., Labaka, G., and Sarasola, K. (2006). Uso de información morfológica en el alineamiento Español-Euskara. In *XXII Congreso de la SEPLN*.

Alegria I., Artola X., Sarasola K. 1996. Automatic morphological analysis of Basque. *Literary & Linguistic Computing* Vol. 11, No. 4, 193-203. Oxford University Press. Oxford. 1996.

Alegria I., Gurrutxaga A., Saralegi X., Ugartetxea S. 2006. ELeXBi, A Basic Tool for Bilingual Term Extraction from Spanish-Basque Parallel Corpora. *Proc. of the 12th EURALEX International Congress*. pp 159-165

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. *LNCS 4394*. 374-384. Cicling 2007.

Carreras X.,, Chao I., Padró L., Padró M. 2004. FreeLing: An open source Suite of Language Analyzers, in P*roceedings of the 4th International Conference on Language Resources and Evaluation* (LREC'04).

Díaz de Ilarraza A., Labaka G., Sarasola K.. 2008. Spanish-Basque SMT system: statistical translation into an agglutinative language. (Submitted to LREC 2008)

Frederking R., Nirenburg S. 1994. Three heads are better than one. *Proc. of the fourth ANLP*. Stuttgart,

Giménez J., Amigó E., Hori C. 2005. Machine Translation Evaluation Inside QARLA. In *Proceedings of the International Workshop on Spoken Language Technology* (IWSLT'05)

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

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. *MT Summit X.* Phuket.

Labaka G., Stroppa N., Way A., Sarasola K. 2007 Comparing Rule-Based and Data-Driven Approaches to Spanish-to-Basque Machine Translation *Proc. of MT-Summit XI,* Copenhagen

Macherey W., Och F, 2007. An Empirical Study on Computing Consensus Translations from Multiple Machine Translation Systems. *Proc. of the EMNLP and CONLL* 2007. Prague.

Martínez R., Abaitua J., Casillas A. Alingning Tagged Bitext. *Proceedings of the Sixth Workshop on Very Large Corpora,* 1998.

Mayor A. 2007. *Matxin: erregeletan oinarritutako itzulpen automatikoko sistema*. Ph. Thesis. Euskal Herriko Unibertsitatea.

Matusov E., Ueffing, N, Ney H. 2006. Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment. *Proc. of EACL 2006*, Trento.

McTait K. A Language-Neutral Sparse-Data 1999. Algorithm for Extracting Translation Patterns". *Proceedings of 8th International Conference on Theorical and Mathodological Issues in Machine Translation*,.

Och F. and Ney H. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics,* 29(1): 19–51.

Sim K., Byrne W., Gales M., Sahbi H. 2007., Wood-

land P. Consensus network decoding for statistical machine translation system combination. Proc. of ICASSP, 2007

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 AMTA-2006, Cambridge, USA.

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

Stroppa N., Groves D., Way A., Sarasola K. 2006.Example-Based Machine Translation of the Basque Language. *7th conf. of the AMTA*.

van Zaanen M. and Somers H. 2005. DEMOCRAT: Deciding between Multiple Outputs Created by Automatic Translation. *MT Summit X*. Phuket.

Way A. and Gough N. 2005. Comparing Example-Based and Statistical Machine Translation. *Natural Language Engineering*, 11(3):295–309.