Comparing Rule-Based and Data-Driven Approaches to Spanish-to-Basque Machine Translation

Gorka Labaka[1], Nicolas Stroppa[2], Andy Way[2], Kepa Sarasola[1]

1- Informatika Fakultatea
University of the Basque Country
Donostia, Basque Country, Spain
{jiblaing ,kepa.sarasola}@ehu.es

2- National Centre for Language Technology
Dublin City University
Dublin 9, Ireland
{nstroppa,away}@computing.dcu.ie

Abstract

In this paper, we compare the rule-based and data-driven approaches in the context of Spanish-to-Basque Machine Translation. The rule-based system we consider has been developed specifically for Spanish-to-Basque machine translation, and is tuned to this language pair. On the contrary, the data-driven system we use is generic, and has not been specifically designed to deal with Basque. Spanish-to-Basque Machine Translation is a challenge for data-driven approaches for at least two reasons. First, there is lack of bilingual data on which a data-driven MT system can be trained. Second, Basque is a morphologically-rich agglutinative language and translating to Basque requires a huge generation of morphological information, a difficult task for a generic system not specifically tuned to Basque. We present the results of a series of experiments, obtained on two different corpora, one being "in-domain" and the other one "out-of-domain" with respect to the data-driven system. We show that *n*-gram based automatic evaluation and edit-distance-based human evaluation yield two different sets of results. According to BLEU, the data-driven system outperforms the rule-based system on the in-domain data, while according to the human evaluation, the rule-based approach achieves higher scores for both corpora.

1 - Introduction

Data-driven Machine Translation is nowadays the most prevalent approach carried out in Machine Translation (MT) research; translation results obtained with this approach have now reached a high level of accuracy, especially when the target language is English. Data-driven MT systems base their knowledge on bilingually aligned corpora, and the accuracy of their output depends strongly on the quality and the size of these corpora. Consequently, when pointing out the success of data-driven MT, we also need to make two additional remarks: (i) large and reliable bilingual corpora are unavailable for lots of language-pairs, (ii) translating into a morphologically rich target language makes the task of data-driven systems a lot more difficult.

When translating into Basque, we are confronted with both problems at the same time. First, few bilingual corpora are available which include Basque, which obviously limits to some extent the application of data-driven approaches. Second, Basque is a morphologically-rich agglutinative language that is difficult to translate into, in particular because of the morphological information we need to generate.

In this paper, we compare the rule-based and data-driven approaches in the context of Spanish-to-Basque translation. The rule-based system we consider has been developed specifically for Spanish-to-Basque MT, and is tuned to this language pair. On the contrary, the data-driven system we use is generic, and has not been specifically designed to deal with either of these languages. The generation of the Basque morphemes poses a particular problem for a system untuned to this language.

We present the results of a series of experiments, obtained on two different corpora, one being "in-domain" and the other one "out-of-domain" with respect to the data-driven system. We show that *n*-gram based automatic evaluation and edit-distance based human evaluation yield two different sets of results. According to BLEU, the data-driven system outperforms the rule-based system on the in-domain data, while according to the human evaluation, the rule-based approach achieves higher scores for both corpora.

The remainder of this paper is organized as follows. In Section 2, we introduce *Matxin*, a rule-based MT system designed for Spanish-to-Basque translation. In Section 3, we present *MaTrEx*, a data-driven MT system that we trained on a Spanish-to-Basque bilingual corpus extracted from magazines. In Section 4, we describe how to work at the morpheme level for Basque. In Section 5, we evaluate the two approaches mentioned above, and report and discuss our experimental results. Section 6 concludes the paper and gives avenues for future work.

2 - Matxin: a Rule-Based MT System

In this section, we describe *Matxin*, the main rule-based MT system developed at the University of the Basque Country. *Matxin* is an open source RBMT engine, whose first goal is to translate from Spanish into Basque, using the traditional transfer model. The transfer component of the translation system is based on both shallow and dependency parsing.[1]

[1] Note that *Matxin* is part of a more general project, *OpenTrad,* which implements two different translation approaches. The first one, named *Apertium* (Corbí-Bellot et al., 2005), is based on a shallow-transfer engine suited to machine translation between

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.

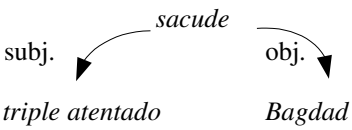
Analysis

The analysis of the Spanish source sentences into dependency trees is performed using an adapted version of the *FreeLing* toolkit (Carreras et al., 2004).[2] *FreeLing* contains a part-of-speech tagger and a shallow parser (or chunker) for Spanish. In *Freeling*, tagging and shallow parsing are performed using the Machine Learning AdaBoost models (Freund & Schapire, 1997). The shallow parses provided by *Freeling* are then augmented with dependency information, using a set of rules that identify the dependencies in the sentence. First, the relationships between chunks is established, based on their labels. As an example, consider the chunked Spanish sentence in (1):

(1) *[np] Un triple atentado* ||| *[verb-chain] sacude* ||| *[np] Bagdad* (a three-pronged attack rocked Baghdad)

Here the dependency parser identifies the verb-chain as the head of the sentence, and the two noun phrases as its children. Then, the dependencies are labelled using a second set of rules. In the previous example "*Un triple atentado*" and "*Bagdad*" are recognised to be the subject and the object respectively of the main verb "sacude". The analysis of this sentence is displayed in (2):

(2)



Transfer

The transfer component consists of lexical transfer and structural transfer.

Lexical transfer is performed using a Spanish-to-Basque dictionary compiled into a finite-state transducer. This dictionary is based on the wide-coverage dictionary *Elhuyar*.[3] This dictionary was enriched with named entities and terms automatically extracted from parallel corpora. This extraction was performed using the Consumer and EITB corpora (see Section 5 for a detailed description of these corpora). Moreover, some Spanish words (such as articles, conjunctions, etc.) do not translate into Basque words, and are translated as morphemes that will be concatenated to other words.

Note that in the actual version of the engine no word-sense disambiguation is performed (we plan to solve semantic ambiguities within a concrete domain in the near future), but a large number of multi-word units representing collocations, named entities and complex terms are included in the bilingual dictionary in order to reduce the influence of this limitation. In the case of prepositions, we adopt another strategy: we decide on the proper translation using some information about verb argument structure extracted automatically from the corpus.

Structural transfer is applied to turn the source dependency tree structure into the target dependency structure. This transformation follows a set of rules that will copy, remove, add, or reorder the nodes in the tree. In addition, specialized modules are included to translate verb chains (Alegria et al., 2005).

Generation

Generation, like transfer, is decomposed into two steps. The first step, referred to as syntactic generation, consists of deciding in which order to generate the target constituents within the sentence, and the order of the words within the constituents. The second step, referred to as morphological generation, consists of generating the target surface forms from the lemmas and their associated morphological information.

In order to determine the order of the constituents in the sentence, a set of rules is defined that state the relative order between a node in the dependency tree and its ancestors. For example, a prepositional phrase is generated before its ancestors if the latter is a noun phrase. The order of the words within the chunks is solely based on the Part-of-Speech information associated with the words.

In Basque, the declension case, number case and other features are assigned to a whole NP as a suffix of the last word of the phrase. Consequently, when generating Basque, the main inflection of a noun phrase is added to its last word. In the case of a verb chain phrase, morphological generation needs to be applied to every word in the phrase.

In order to perform morphological generation, we use the morphological generator for Basque described in (Alegria et al., 1996). This generator makes use of the morphological dictionary developed in *Apertium*, which establishes correspondences between surface forms and lexical forms for Basque. It is used in morphological generation to produce the inflected forms of Basque words. In particular, this dictionary contains:

languages showing syntactic similarities (up to now, Spanish, Catalan and Galician are handled); it can be freely downloaded from http://apertium.sourceforge.net. The second one is *Matxin*, based on a deep-transfer engine, and is focused on the Spanish-Basque language pair; it is a continuation of previous work in the IXA group (Diaz de Ilarraza et al., 2000). *Matxin* can be freely downloaded from http://matxin.sourceforge.net.

[2] Freeling can be freely downloaded from http://www.lsi.upc.edu/~nlp/freeling/.

[3] http://www1.euskadi.net/hizt_el.

- A definition of Basque paradigms (sets of correspondences between partial surface forms and partial lexical forms). Those paradigms are similar to continuation classes in two-level morphology (Koskeniemmi, 1983).
- Lists of surface form to lexical form correspondences for complex lexical units (including multi-word units).

This dictionary is compiled into a finite-state transducer which is used to perform the morphological generation of Basque words. A more detailed description of this process can be found in (Armentano-Oller et al., 2005).

3 - MaTrEx: a Data-Driven System

The *MaTrEx* system (Stroppa & Way, 2006) used in our experiments is a modular data-driven MT engine, which consists of a number of extendible and re-implementable modules, the most important of which are:

- Word Alignment Module: takes as its input an aligned corpus and outputs a set of word alignments.
- Chunking Module: takes in an aligned corpus and produces source and target chunks.
- Chunk Alignment Module: takes the source and target chunks and aligns them on a sentence-by-sentence level.
- Decoder: searches for a translation using the original aligned corpus and derived chunk and word alignments.

The Word Alignment and the Decoder modules are wrappers around existing tools, namely Giza++ (Och & Ney, 2003), and *Moses* (Koehn et al., 2007). The chunking and alignment strategies are described in more detail below.

The translation process can be decomposed as follows: the aligned source-target sentences are passed in turn to the word alignment, chunking and chunk alignment modules, in order to create our chunk and lexical example databases. These databases are then given to the decoder to translate new sentences. These steps are displayed in Figure 1.

Chunking

In the case of Spanish, the extraction of chunks relies on the shallow parser described above (as part of *Freeling*). This shallow parser enables us to identify the main constituents in the sentence: noun phrases, verb phrases, prepositional phrases, etc.

In the case of Basque, we use the toolkit *Eusmg,* which performs POS tagging, lemmatisation and chunking (Adu riz & Díaz de Ilarraza, 2003). It recognizes syntactic structures by means of features assigned to word units, following the constraint grammar formalism (Karlsson, 1995). An example of chunked sentences is given in (3), for Spanish and Basque:

(3)

Spanish:
Un triple atentado sacude Bagdad:
=> *[np] Un triple atentado* ||| *[verb-chain] sacude* ||| *[np] Bagdad*

Basque:
atentatu hirukoitz batek Bagdad astintzen du
=> *[np] atentatu hirukoitz batek* ||| *[np] Bagdad* |||
[verb-chain] astintzen du

Note that, since each module of the system can be changed independently of the others, it is possible to use a variety of chunkers, including those of the Marker-based approach, used in other works (Gough & Way, 2004; Stroppa et al., 2006; Stroppa & Way, 2006).

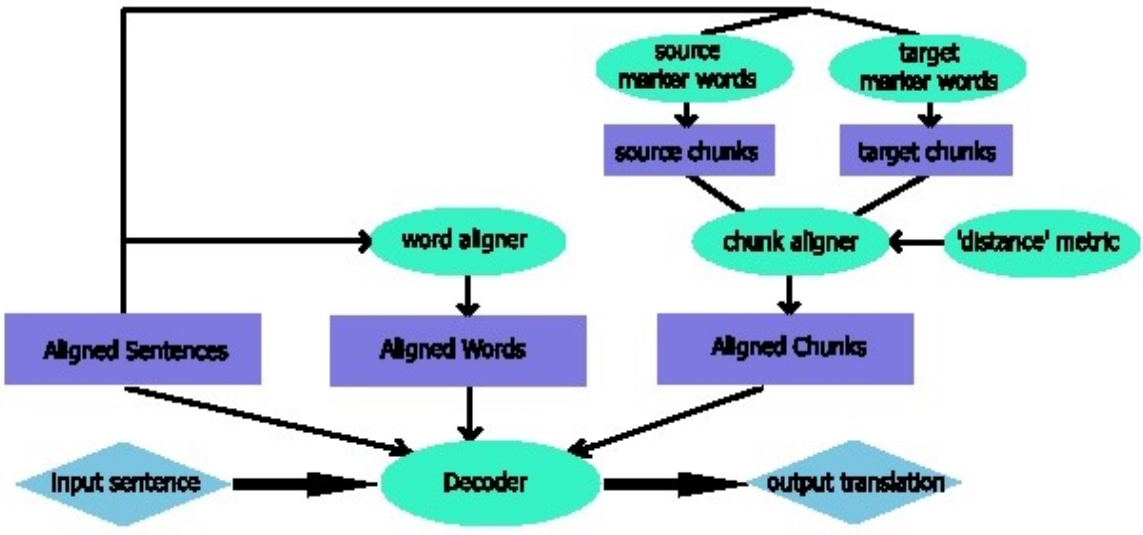


Figure 1: Translation Process in *MaTrEx*

Alignment Strategies

Word alignment
Word alignment is performed using the *Giza++* statistical word alignment toolkit and we followed the "refined" method of (Koehn et al., 2003) to extract a set of high-quality word alignments from the original uni-directional alignment sets. These along with the extracted chunk alignments were passed to the translation decoder.

Chunk alignment
In order to align the chunks obtained by the chunking procedures introduced in Section "Chunking", we make use of an "edit-distance style" dynamic programming alignment algorithm, as described in (Stroppa et al., 2006).

This algorithm works as follows. First, a "similarity" measure is determined for each pair of source-target chunks. Then, given these similarities, we use a modified version of the edit-distance alignment algorithm to find the optimal alignment between the source and the target chunks. The modification consists of allowing for jumps in the alignment process (Leusch et al., 2006), which is a desirable property for translating between languages showing significant syntactic differences. This is the case for Spanish and Basque, where the order of the constituents in a sentence can be very different.

To compute the "similarity" between pair of chunks, we rely on the information contained within the chunks. More precisely, we relate chunks by using the word-to-word probabilities that were extracted from the word alignment module. The relationship between a source chunk and a target chunk is computed thanks to a model similar to IBM model 1 (Stroppa et al., 2006).

Integrating SMT data
Since its inception, EBMT has recommended the use of both lexical and phrasal information (Nagao, 1984); current SMT models now also use phrases in their translation models (Koehn et al., 2003). Actually, it is possible to combine elements from EBMT and SMT to create hybrid data-driven systems capable of outperforming the baseline systems from which they are derived, as shown in (Groves and Way, 2005). Therefore, we also make use of SMT phrasal alignments, which are added to the aligned chunks extracted by the chunk alignment module. The SMT phrasal alignment follows the procedure of (Koehn et al., 2003).

Decoder
The decoding module is capable of retrieving already translated sentences and also provides a wrapper around *Moses*, a phrase-based decoder. This decoder also implements Minimum-Error-Rate Training (Och, 2003) within a log-linear framework (Och & Ney, 2002). The BLEU metric (Papineni et al., 2002) is optimized on a development set. We use a log-linear combination of several common feature functions: phrase translation probabilities (in both directions), word-based translation probabilities (lexicon model, in both directions), a phrase length penalty and a target language model.

The decoder also relies on a target language model. The Basque language model is a simple 3-gram language model trained on the Basque portion of the training data, using the SRI Language Modeling Toolkit,[4] with modified Kneser-Ney smoothing.

4 - Morpheme-Based Machine Translation

Basque is an agglutinative language in which words may be made up of a large number of morphemes. For example, suffixes can be added to the last word of a noun phrase; these suffixes can represent some morpho-syntactic information associated to the noun phrase, such as number, definiteness, grammatical cases and postpositions.

As a consequence, most words only occur once in the training data, leading to serious sparseness problems when extracting statistics from the data. In order to limit this problem, one solution is to working at a different representation level, namely morphemes (cf. (Stroppa et al., 2006)). By segmenting each word into a sequence of morphemes, we reduce the number of tokens that occur only once (cf. (Agirre et al., 2006)). Furthermore, as many Basque words correspond to several Spanish words (for example, the Basque "etxeko" translates to "de la casa" in Spanish), lots of 1-to-n alignments have to be defined when working at the word level. Although 1-to-n alignments are allowed in IBM model 4, training can be harmed when the parallel corpus contains many such cases.

Working at the morpheme level within *MaTrEx* is straightforward: we only need to segment the Basque side of the training (and development) data. The *MaTrEx* system trained on these new data will generate a sequence of morphemes as output.

In the experiments we carried out, we report results obtained when working at both the word and morpheme levels.

From Words to Morphemes

Working at the morpheme level does, however, have some drawbacks. In particular, if we want to be able to generate surface word forms from morphemes, then we need to include some additional information to the morphemes. In (Agirre et al., 2006), a segmentation strategy is proposed, which does not include this information. In this paper, we build upon this strategy,

[4] http://www.speech.sri.com/projects/srilm/

but we also include the required information to recover the surface words from the morphemes.

To obtain the segmented text, the Basque text is analyzed using Eustagger (Aduriz & Díaz de Ilarraza, 2003), a two-level morphology (Koskeniemmi, 1983) analyser/tagger. After this process, each word is replaced with the corresponding lemma accompanied with a list of morphological features. A sentence and the associated segmentation are displayed in (4), where each morpheme is accompanied by the appropriate morphological information:

(4)

Original Basque sentence:
Loe berriak indarrean eusten dio lege horri .

Segmented sentence:
Loe<IZE><IZB> berri<ADJ><ARR>
+<P>+<ABS> indarrean<ADB><ARR>
eutsi<ADI><SIN>+<ADOIN>+<EZBU>
edun +<A1>
+<NR_HURA>+<NI_HARI>+<NK_HARK>
lege<IZE><ARR> hori<DET><ERKARR>
+<S>+<DAT> .

From Morphemes to Words

When working at the morpheme level, the translation of a (source) sentence obtained using *MaTrEx* is a sequence of morphemes. If we want to produce a Basque text, then we need to recover the words from this sequence of morphemes; the output of *MaTrEx* is thus post-processed to produce the final Basque translation.

This post-processing consists of using the morphological generation module of *Matxin*. This module uses the same lexicon and two level rules as *Eustagger*. However, in the context of generation, we are faced with two new additional problems:

- Unknown lemmas: some lemmas do not occur in the *Eustagger* lexicon, such as unknown proper names. To solve this problem, the synthesis component has been enriched to generate words from unknown lemmas using default rules defined for each part of speech.

- Invalid sequences of tags: the output of *MaTrEx* (a sequence of morphemes) is not necessarily a well-formed sequence from a morphological point of view. For example, the correct tags might be generated, but in the wrong order. In some cases, a nominal tag is assigned to verb; sometimes, required tags are missing. In the current work, we do not try to correct these mistakes: we simply output the lemma, and remove the inappropriate tag information. A more refined treatment is left to future work.

5 - Experimental Results

Data and Evaluation

The experiments were carried out using two different test sets.

The first, referred to as *ConsumerTest*, contains 1500 bilingually aligned sentences extracted from the *Consumer* Eroski Parallel Corpus.[5] The *Consumer* Eroski Parallel Corpus is a collection of 1036 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 the sentence level.

The second, referred to as *EitbTest*, also contains 1500 bilingually aligned sentences extracted from the EITB corpus. This corpus is a collection of news (Basque News and Information Channel, *http://www.eitb24.com/en*), available in Spanish, Basque, and English.[6] This corpus contains approximately 1,500,000 Spanish words and 1,200,000 Basque words.

While *Eitb* is a *general* news corpus (politics, economy, sport, etc.), *Consumer* is a corpus of articles comparing the quality and prices of commercial products and brands. They are consequently from two different terminological "domains". Table 1 summarizes the various statistics related to these corpora.

Since the *Matxin* system is rule-based, it does not need any kind of training, and can be directly applied to translate into Basque the Spanish test sentences. However, Matxin's bilingual lexicon was enriched with 1129 entries (entities and multi-word terms) that were automatically extracted form the *ConsumerTrain* bilingual corpora.

In order to train the *MaTrEx* system, which is data-driven and relies on bilingually aligned training material, we used approximately 50,000 aligned sentences from the *ConsumerTrain* dataset, which was extracted in a similar manner to the *Consumer* dataset. In order to tune the parameters of the *MaTrEx* system, we use an additional development set of 1292 sentence pairs (referred to as *ConsumerDev*). Training *MaTrEx* on *ConsumerTrain* makes the *ConsumerTest* dataset "in-domain", and the *Eitb* dataset "out-of-domain". We thus expect the *MaTrEx* system to perform better on the *ConsumerTest* set than on the *EitbTest* set.

[5] The *Consumer* corpus is accessible online via Universidade de Vigo (http://sli.uvigo.es/CLUVI/, public access) and Universidad de Deusto (http://www.deli.deusto.es, research intranet).

[6] EITB is the official media group in the Basque Country with four television channels and five radio stations.

	Spanish	Basque
ConsumerTrain		
Sentences	51949	
Running words	976730	786705
Running morphemes	-	910995
Word voc. Size	44715	76292
Morph. Voc. Size	-	29805
ConsumerDev		
Sentences	1292	
Running words	24755	19978
Running morphemes	-	22554
Word voc. Size	5973	7367
Morph. Voc. Size	-	4064
ConsumerTest		
Sentences	1501	
Running words	34231	27278
Running morphemes	-	45480
Word voc. Size	7278	9258
Morph. Voc. Size	-	5999
EitbTest		
Sentences	1500	
Running words	36783	26857
Running morphemes	-	41602
Word voc. Size	7345	7918
Morph. Voc. Size	-	5706

Table 1: Corpus statistics.

In order to assess the quality of the translation obtained using both systems, we used automatic evaluation metrics as well as human evaluation. As for automatic evaluation, we report the following accuracy measures: BLEU (Papineni et al., 2002), and NIST (Doddington, 2002). For each testset, we have access to one Basque reference translation per sentence. Evaluation is performed in a case-insensitive manner. Because of the specific nature of Basque, we perform two types of evaluation: a word-based evaluation, and a morpheme-based evaluation.

Since human evaluation is an expensive process, we selected 50 sentences from the *ConsumerTest* corpus to be human evaluated; this corpus is referred to as *ConsumerTestHuman*. The same applies to *EitbTest*, yielding *EitbTestHuman*. We used the edit-distance metric (Przybocki et al., 2006) called HTER or Translation Error Rate with human-targeted references (Snover et al., 2006). Edit distance is defined as the number of modifications a native Basque professional translator has to make so that the resulting edited translation is an easily understandable Basque sentence that contains the complete meaning of the source sentence. We used the software described in (Snover et al., 2006) to compute HTER. The post-editing work took 6 hours in total.

Automatic Evaluation Results

For the *ConsumerTest* corpus, the results obtained with the *MaTrEx* system are higher than those obtained with the *Matxin* system. With respect to the BLEU score, this difference is 1.58 points absolute for the word-based evaluation (27% relative increase), and 2.47 points absolute for the morpheme-based evaluation (21% relative increase). These differences are statistically significant, with a p-value < 0.002, computed using approximate randomisation (Riezler & Maxwell, 2005).

For the *EitbTest* corpus, the results obtained with the *MaTrEx* system are much lower than those obtained with the *Matxin* system. The differences are also statistically significant, with a p-value < 0.002, for both BLEU and NIST scores. This is consistent with our intuition since with respect to *MaTrEx*, the *EitbTest* corpus is "out-of-domain" (cf. (Koehn & Monz, 2006) for a comparison between in-domain and out-of-domain results of data-driven systems).

These results show that a (generic) data-driven system can be very competitive with a (specialized) rule-based system, if suitable training data is available. The argument in favour of rule-based systems is stronger when no relevant bilingual training data are available.

Given the globally low scores obtained, it is important to make two additional remarks. First, it shows the difficulty of the task of translating to Basque, which is due to the strong syntactic differences with Spanish, and the morphological properties of this language. Second, even if a morpheme-based translation is more appropriate than a word-based translation, *n*-gram based metrics are not suited to the comparison between sequences of morphemes. In particular, the absence of morphological tags that may not affect the global understanding of a sentence are penalised: if such a tag is missing in the system's output, all the *n*-grams that could have contained it would be cut.

	ConsumerTest		EitbTest	
	BLEU	NIST	BLEU	NIST
Matxin-WB	6.31	3.66	9.30	3.13
MaTrEx-WB	8.03	3.69	9.02	2.70
Matxin-MB	12.01	4.62	12.76	3.75
MaTrEx-MB	14.48	4.63	6.25	2.89

Table 2: Automatic evaluation results.

The results obtained for the Spanish-to-Basque translation task using the *ConsumerTest* and *EitbTest* datasets are summarized in Table 2, in which WB and MB denote respectively the word-based evaluation and the morpheme-based evaluation. For the morpheme-based evaluation, we segment the reference sentences into morphemes with which we compare the output of each system (which is also a sequence of morphemes).

Human Evaluation Results

The human evaluation results, obtained using HTER, are reported in Table 3. We conducted a word-based evaluation (WB), as well as a morpheme-based

evaluation (MB). For the morpheme-based evaluation, both the reference and the translated text are divided into morphemes.

	ConsumerTest Human	EitbTest Human
	HTER	HTER
Matxin-WB	43.6	40.4
MaTrEx-WB	57.9	71.8
Matxin-MB	39.1	34.9
MaTrEx-MB	49.6	76.3

Table 3: Subjective evaluation results.

For the *ConsumerTestHuman* corpus, we can observe that the error rate obtained by *Matxin* is lower than the one obtained by *MaTrEx*: 14.3 points for the word-based evaluation and 10.5 points for the morpheme-based evaluation.

Concerning the *EitbTestHuman* corpus, i.e. the "out-of-domain" corpus, the difference is even higher. While *Matxin*'s error-rate is quite similar to the one obtained with the *Consumer* corpus (40.4 points), the error-rate for *MaTrEx* becomes quite large (71.8 points).

These results are consistent with the domain independence of the rule-based system, which achieves a comparable translation quality for the two corpora. The data-driven approach is domain-dependent by construction and, as expected, it performs better on the in-domain corpus. According to the subjective evaluation, the translation quality of *Matxin* is better, irrespective of the corpus. However, it must be stressed that *Matxin* has been specifically developed and designed to translate from Spanish to Basque over a number of years, while *MaTrEx* is generic and the cost of adapting it to Spanish-Basque translation is several orders of magnitude lower.

6 - Conclusions and Future Work

In this paper, we have compared a rule-based MT system (*Matxin*) and a data-driven MT system (*MaTrEx*) in the context of Spanish-to-Basque translation. While the rule-based system we consider has been developed specifically for Spanish-to-Basque machine translation, the data-driven system we use is generic, and has not been specifically tuned to Basque.

We have introduced a translation scheme based on morphemes instead of words, in order to be able to deal with the particular agglutinative nature of Basque. This allows for the generation of the morphological information required to recover the full Basque surface word forms.

We have presented experimental results comparing the two types of approaches on two different corpora containing magazine and news articles respectively. Objective evaluation metrics such as BLEU and NIST yield different results to subjective evaluation metrics such as HTER. The automatic metrics indicate that the data-driven system outperforms the rule-based system on the in-domain data. On the contrary, the subjective evaluation indicates that the rule-based system outperforms the data-driven approach for both corpora. Note that these results are also consistent with the findings of (Callison-Burch et al., 2006) concerning objective and subjective evaluation.

Moreover, both types of evaluation confirm that *Matxin*, the rule-based system, is domain-independent while *MaTrEx*, the data-driven system, is more domain -dependent. Accordingly, if a different domain were selected which was quite different from the magazine or news articles used here (weather forecasts, say), then we would expect *MaTrEx* to win out. That said, having invested a large number of person-years in its development, it is encouraging to see the good performance of *Matxin* on out-of-domain data.

Future work consists of building upon the respective strength of both approaches, by exploring various hybridity strategies focused on the problem of Basque translation. One avenue that we would expect to bear fruit is adding into *MaTrEx* the bilingual lexicon from *Matxin.* We also plan to use automatic evaluation metrics that would be more suited to the evaluation of morpheme-based translation (cf. (Owczarzak et al., 2006)).

Acknowledgments

This work is partially supported by Science Foundation Ireland (grant number OS/IN/1732), Spanish M.E.C. (OpenMT project, TIN2006-15307-C03-01), and the Basque Government (AnHitz project, eIE06-185). Our colleagues Iñaki Alegria, Arantza Díaz de Ilarraza, Mikel Lersundi, and Aingeru Mayor are kindly acknowledged for providing their expertise on the *Matxin* system and the evaluation of the output.

References

I. Aduriz and A. Díaz de Ilarraza (2003). Morphosyntactic disambiguation and shallow parsing in Computational Processing of Basque. In *Inquiries into the lexicon-syntax relations in Basque*, B. Oiharcabal (ed.), Univ. of the Basque Country, Donostia, Spain.

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

I. Alegria, A. Díaz de Ilarraza, G. Labaka, M. Lersundi, A. Mayor, and K. Sarasola (2005). An FST grammar for verb chain transfer in a Spanish-Basque MT System. In *Proceedings of Finite-State Methods and Natural Language Processing,* pp.295—96, Helsinki, Finland.

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

C. Armentano-Oller, A. Corbí-Bellot, M. L. Forcada, M. Ginestí-Rosell, B. Bonev, S. Ortiz-Rojas, J. A. Pérez-Ortiz, G. Ramírez-Sánchez, and F. Sánchez-Martínez (2005). An open-source shallow-transfer Machine Translation toolbox: consequences of its release and availability. In *Proceedings of Open-Source MT workshop, MT Summit X*, Phuket, Thailand.

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

C. Callison-Burch, M. Osborne and P. Koehn (2006). Re-evaluating the Role of Bleu in MT Research. In P*roceedings of EACL 2006*, pp.249—256, Trento, Italy.

A. Corbí-Bellot, M. Forcada, S. Ortiz-Rojas, J. Pérez-Ortiz, G. Ramírez-Sánchez, F. Sánchez-Martínez, I. Alegria, A. Mayor and K. Sarasola (2005). An Open-Source Shallow-Transfer Machine Translation Engine for the Romance Languages of Spain. In *Proceedings of 10th EAMT Conference: Practical Applications of Machine Translation*, Budapest, Hungary, pp.79—86.

G. Doddington (2002). Automatic evaluation of Machine Translation quality using n-gram cooccurrence statistics. In *Proceedings of HLT 2002*, pp. 128—132, San Diego, CA.

N. Gough and A. Way (2004). Robust large-scale EBMT with marker-based segmentation. In *Proceedings of TMI 2004*, pp.95—104, Baltimore, MD.

D. Groves and A. Way (2005). Hybrid data-driven models of MT. *Machine Translation* **19**(3,4):301—323.

F. Karlsson, A. Voutilainen, J. Heikkila, and A. Anttila, editors (1995). *Constraint Grammar: A Language-independent System for Parsing Unrestricted Tex*t. Mouton de Gruyter, Berlin, New-York.

Y. Freund and R. Schapire (1997). A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting. *Journal of Computer and System Sciences* **55**(1):119—139.

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 (2007). Moses: Open source toolkit for SMT, in *Proceedings of the ACL 2007 Demo and Poster Session*, Prague, Czech Republic, pp.177—180.

P. Koehn and C. Monz (2006). Manual and Automatic Evaluation of MT. In *Proceedings of HLT-NAACL Workshop on SMT*, pp.102—121, New York.

P. Koehn, F. Och, and D. Marcu (2003). Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pp. 48-54, Edmonton, Canada.

K. Koskenniemi (1983). Two-level Model for Morphological Analysis. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pp.683—685, Karlsruhe, Germany.

G. Leusch, N. Ueffing, and H. Ney (2006). CDER: Efficient MT evaluation using block movements. In *Proceedings of EACL 2006*, pp.241—248, Trento, Italy.

M. Nagao (1984). Framework of a mechanical translation between Japanese and English by analogy principle. In *Artificial and Human Intelligence*, A. Elithorn and R. Banerji, Eds. Amsterdam, The Netherlands: North-Holland, pp.173—180.

F. Och, (2003). Minimum error rate training in statistical machine translation. In *Proceedings of 41st ACL*, pp. 160—167, Sapporo, Japan.

F. J. Och and H. Ney (2002). Discriminative training and maximum entropy models for SMT. In *Proceedings of 40th ACL*, pp. 295—302, Philadelphia, PA.

K. Owczarzak, D. Groves, J. Van Genabith and A. Way (2006). Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation. In *Proceedings of HLT-NAACL Workshop on SMT*, pp.86—93, New York.

K. Papineni, S. Roukos, T. Ward, and W. Zhu (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th ACL*, pp. 311—318, Philadelphia, PA.

M. Przybocki, G. Sanders, and A. Le (2006). Edit distance: a metric for MT evaluation. In *Proceedings of 5th LREC*, pp. 2038—2043, Genoa, Italy.

S. Riezler and J. Maxwell (2005). On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the the ACL Workshop on Intrisic and Extrinsic Evaluation Measures for MT and/or Summarization*, pp. 57—64, Ann Arbor, MI.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA 2006*, pp.223—231, Cambridge, MA.

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

Stroppa, N. and A. Way. MaTrEx: DCU Machine Translation System for IWSLT 2006. In *Proceedings of IWSLT 2006*, pp.31—36, Kyoto, Japan.