Example-Based Machine Translation of the Basque Language

Nicolas Stroppa, Declan Groves, Andy Way
National Centre for Language Technology
Dublin City University
Dublin 9, Ireland
{nstroppa,dgroves,away}@computing.dcu.ie

Kepa Sarasola*
Informatika Fakultatea
University of the Basque Country
Donostia, Basque Country, Spain
kepa.sarasola@ehu.es

Abstract

Basque is both a minority and a highly inflected language with free order of sentence constituents. Machine Translation of Basque is thus both a real need and a test bed for MT techniques.

In this paper, we present a modular Data-Driven MT system which includes different chunkers as well as chunk aligners which can deal with the free order of sentence constituents of Basque.

We conducted Basque to English translation experiments, evaluated on a large corpus ($270,000$ sentence pairs). The experimental results show that our system significantly outperforms state-of-the-art approaches according to several common automatic evaluation metrics.

1 Introduction

Machine Translation (MT) has been shown to be useful for minority languages such as Basque. Currently, in the case of Basque, more than ten official institutions and private companies commonly use translation memories. Moreover, Basque, being a highly inflected language with free order of sentence constituents, makes it suitable to be used as a test bed for new MT techniques.

As a result, a rule-based MT system has recently been released with the aim of translating texts and websites from Spanish to Basque (Alegria et al., 2005).[1] Furthermore, the availability of a large Basque-English corpus with more than $270,000$ aligned translation units and with an average of more than 9 words per unit makes it possible to use state-of-the-art corpus-based MT techniques for Basque-English translation.

In this paper, we present a new Data-Driven MT system which exploits both EBMT and SMT techniques to extract a dataset of aligned chunks. For the extraction of the EBMT data resources, we make use of two different chunking methods. In the case of English, we employ a marker-based chunker based on the Marker Hypothesis (Green, 1979). For Basque, we currently use the dedicated tools developed at the University of the Basque Country while investigating the application of our marker-based chunker to Basque. The chunks are then aligned thanks to a dynamic programming algorithm which is similar to an edit-distance algorithm while allowing for block movements (Leusch et al., 2006). This aligner also relies on relationships between chunks, which we compute in several ways.

We present a set of Basque to English translation experiments, evaluated on a large corpus consisting of software manuals. We show a significant improvement over state-of-the-art SMT systems (Koehn, 2004) according to several common automatic evaluation metrics. In addition, some manual evaluations have been conducted to assess the quality of our extracted aligned resources, and we find that the high quality of our resources con-

* Work carried out while at The National Centre for Language Technology at Dublin City University in Spring 2006.

[1] See also `http://matxin.sourceforge.net` and `http://www.opentrad.org`.

Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pages 232-241,
Cambridge, August 2006.

tributes positively to the overall translation quality.

This paper is organized as follows: In section 2, we present the particularities of Basque to English Machine Translation. Section 3 introduces the system we developed and the methods used for chunks alignments. In Section 4, we give more details about the process of relating Basque and English chunks. In Section 5, experimental results on Basque to English MT and on chunks alignments quality are presented. Section 6 presents some related work. Section 7 concludes the paper and gives an outlook on future work.

2 Basque/English MT: some particularities

The particularity of the Basque language increases the complexity of the task of aligning words and phrases with linguistic units in other languages; from a morphological point of view, Basque is a highly inflected agglutinative language and at the sentence level, it is a (relatively) free constituent order language.

- Since Basque is an agglutinative language, there is no definitive divide between morphology and syntax. As a consequence, **morphemes** can be used as the basic units of analysis instead of words (Goenaga, 1980; Abaitua, 1988; Abaitua et al., 1992; Aldezabal et al., 2003). This differs from most European languages, such as English or French. A "word" in Basque can thus correspond to several words in English (cf. Figure 1).

pantailan	→	within the screen
atxikitze-puntuaren	→	of the snap point
duen	→	that has

Figure 1: Basque morphemes and English words

- **Free order of sentence constituents**. In Basque, the order of the main constituents of a sentence is relatively free. For example, the 24 possible permutations obtained by changing the order of the *subject*, *object* and *PP* in the sentence displayed in Figure 2 are all well-formed Basque sentences.

  Consequently, unlike English or French and most other European languages, Basque sentences do not usually start with a subject followed by a verb; the subject and verb do not necessarily appear in the same order and may not even appear consecutively in Basque sentences. It is also important to mention that this general flexibility at the sentence level is much more restricted within other syntactic units (for example, inside NPs or subordinated sentences).

- Moreover, there is **agreement** in number and person between verb and subject, and object and indirect object (corresponding roughly to the ergative, absolutive and dative cases) in Basque.

3 EBMT and Hybrid Data-Driven MT

Within the field of corpus-based MT, phrase-based SMT is by far the most dominant paradigm, yet much important research continues to be carried out in EBMT. As with SMT, EBMT makes use of a corpus of source–target sententially-aligned examples to automatically extract translation resources. During translation an EBMT system:

1. Searches the source side of the corpus for "close" matches and their equivalent target language translations;

2. Identifies useful source–target fragments contained in those retrieved examples;

3. Recombines relevant target language fragments to derive the final translation of the input sentence.

The nature of the examples used in the first place may include using placeables as indicators of chunk boundaries (Brown, 1999) or aligning phrase-structure (sub-)trees (Hearne and Way, 2003) or dependency trees (Watanabe et al., 2003).

3.1 The MaTrEx System

The MATREX (Machine Translation Using Examples) system used in our experiments is a data-driven MT engine, built following an extremely modular design. It consists of a number of extendible and re-implementable modules, the most important of which are:

"The dog brought the newspaper in his mouth" $\Longrightarrow$ *"Txakurrak egunkaria ahoan zekarren"*

Txakur-rak	egunkari-a	aho-an	zekarren
The-dog	the-newspaper	in-his-mouth	brought
ergative-3-s	absolutive-3-s	inessive-3-s	
Subject	Object	Modifier	Verb

23 other possible orders:

Txakur-rak	aho-an	egunkari-a	zekarren
Txakur-rak	aho-an	zekarren	egunkari-a
Egunkari-a	txakur-rak	zekarren	aho-an
...			

Figure 2: Free order between sentence units in Basque

- 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.

In a preprocessing stage, the aligned source-target sentences are passed in turn to the word alignment, chunking and chunk alignment modules to create our chunk and lexical example databases used during translation. In our experiments we investigated a number of different chunking and alignment strategies which we describe in more detail in what follows.

3.2 Chunking Strategies

3.2.1 Marker-Based Chunking

One method for the extraction of chunks, used in the creation of the example database, is based on the Marker Hypothesis (Green, 1979), a psycholinguistic constraint which posits that all languages are marked for surface syntax by a specific closed set of lexemes or morphemes which signify context. Using a set of closed-class (or "marker") words, such as determiners, conjunctions, prepositions, possessive and personal pronouns, aligned source-target sentences are segmented into chunks (Gough and Way, 2004) during a pre-processing step. A chunk is created at each new occurrence of a marker word, with the restriction that each chunk must contain at least one content (or non-marker) word. In addition to the set of marker words used in the experiments of (Gough and Way, 2004; Groves and Way, 2005), punctuation is also used to segment the aligned sentences – with the punctuation occurring in chunk final, rather than initial, position.

3.2.2 Basque Chunking

As previously mentioned, Basque is an agglutinative language and morphemes can be used as the basic units of analysis. However, this makes it more difficult to apply the marker-based chunker, described in Section 3.2.1, to Basque. Therefore, as a starting point, we have decided to make use of existing tools developed for Basque. The Ixa group at the University of the Basque Country has developed two syntactic analyzers:

- PATRIXA is an unification grammar based on morpheme units (Aldezabal et al., 2003);
- EUSMG is a toolkit which performs POS tagging, lemmatisation and chunking (Aduriz and de Ilarraza, 2003). It recognizes syntactic structures by means of features assigned to word units, following the constraint grammar formalism (Karlsson et al., 1995).

For our experiments, we used the EUSMG toolkit to chunk the Basque sentences. After this processing stage, a sentence is treated as a sequence of morphemes, in which chunk boundaries are clearly visible. Morphemes denoting morphosyntactic features are replaced by conventional symbolic strings (cf. Figure 3, in which *++abs++ms* denotes the mor-

phosyntactic features absolutive, definite and singular).

Fitxategi zaharra ezin izan da irakurri

⇓

[fitxategi zahar ++abs++ms] [ezin izan da irakurri]

([The old file] [could not be read])

Figure 3: Basque chunking

After some adaptation, the chunks obtained in this manner are actually very comparable to the English chunks obtained with the marker-based chunker (see Section 4 for more details). Moreover, for future work, we also plan to consider adapting the marker-based chunker to Basque.

3.3 Alignment Strategies

Word alignment For word/morpheme alignment we used the GIZA++ statistical word alignment toolkit, and following the "refined" method of (Och and Ney, 2003), extracted a set of high-quality word/morpheme alignments from the original unidirectional 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 described in Section 3.2, we make use of a dynamic programming "edit-distance style" alignment algorithm.

In the following, a denotes an alignment between a target sequence e and a source sequence f, with $I = |e|$ and $J = |f|$. Given two sequences of chunks, we are looking for the most likely alignment $\hat{a}$:

$$\hat{a} = \underset{a}{\operatorname{argmax}}\, \mathbb{P}(a|e,f) = \underset{a}{\operatorname{argmax}}\, \mathbb{P}(a,e|f).$$

We first consider alignments such as those obtained by an edit-distance algorithm, i.e.

$$a = (t_1, s_1)(t_2, s_2)\dots(t_n, s_n),$$

with $\forall k \in [\![1, n]\!]$, $t_k \in [\![0, I]\!]$ and $s_k \in [\![0, J]\!]$, and $\forall k < k'$:

$$t_k \leq t_{k'} \text{ or } t_{k'} = 0,$$
$$s_k \leq s_{k'} \text{ or } s_{k'} = 0,$$
$$I \subseteq \cup_{k=1}^{n}\{t_k\},\ J \subseteq \cup_{k=1}^{n}\{s_k\},$$

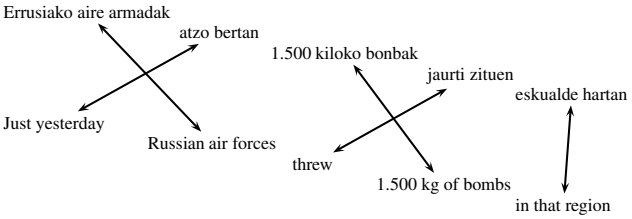

Figure 4: Subject and verb in Basque are not always contiguous

where $t_k = 0$ (resp. $s_k = 0$) denotes a non-aligned target (resp. source) chunk.

We then assume the following model:

$$\mathbb{P}(a, e|f) = \Pi_k \mathbb{P}(t_k, s_k, e|f) = \Pi_k \mathbb{P}(e_{t_k}|f_{s_k}),$$

where $\mathbb{P}(e_0|f_j)$ (resp. $\mathbb{P}(e_i|f_0)$) denotes an "insertion" (resp. "deletion") probability.

Assuming that the parameters $\mathbb{P}(e_{t_k}|f_{s_k})$ are known, the most likely alignment is computed by a simple dynamic-programming algorithm.[2] Moreover, this algorithm can be simply adapted to allow for block movements or "jumps", following the idea introduced by (Leusch et al., 2006) in the context of MT evaluation. This adaptation is necessary to take into account the potential differences between the order of constituents in Basque and English (cf. Figure 5).

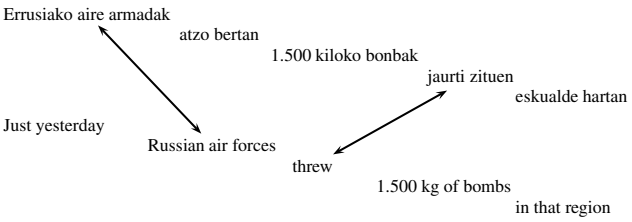

Figure 5: Equivalence between components in Basque and English

Instead of using an Expectation-Maximization algorithm to estimate these parameters, as commonly done when performing word alignment (Brown et al., 1993; Och and Ney, 2003), we directly compute these parameters by relying on the information contained within the chunks. In our experiments, we

[2] This algorithm is actually a classical edit-distance algorithm in which distances are replaced by inverse-log-conditional probabilities.

have considered three main sources of knowledge: (i) word-to-word translation probabilities, (ii) word-to-word cognates and (iii) chunks labels.

Word-to-word probabilities are simply extracted from the word alignment module, as described above. Relationships between chunks are then computed thanks to the following simple model:

$$\mathbb{P}(e_i|f_j) = \sum_{a_c} \mathbb{P}(a_c, e_i|f_j) \simeq \max_{a_c} \mathbb{P}(a_c, e_i|f_j)$$
$$= \Pi_k \max_l \mathbb{P}(e_{i_l}|f_{j_k}).$$

The same kind of model applies to cognates. In the case of chunk labels, a simple matching algorithm is used.

It is possible to combine several sources of knowledge in a log-linear framework, in the following manner:

$$logP(e_i|f_j) = \sum \lambda_k logP_k(e_i|f_j) - logZ,$$

where $P_k(.)$ represents a given source of knowledge, λ_k the associated weight parameter and Z a normalization parameter.

Integrating SMT data Whilst EBMT has always made use of both lexical and phrasal information (Nagao, 1984), it is only recently that SMT has moved towards the use of phrases in their translation models and decoders (Koehn et al., 2003; Koehn, 2004). It has, therefore, become harder than ever to identify the differences between these two data-driven approaches (Way and Gough, 2005). However, despite the convergence of the two paradigms, recent research (Groves and Way, 2005; Groves and Way, 2006) has shown that by combining elements from EBMT and SMT to create hybrid data-driven systems capable of outperforming the baseline systems from which they are derived. Therefore, SMT phrasal alignments are also added to the aligned chunks extracted by the chunk alignment module, in order to produce higher quality translations.

3.4 Decoder

The decoder is also a hybrid system which integrates EBMT and SMT. It is capable of retrieving already translated sentences and also provides a wrapper around the PHARAOH SMT decoder (Koehn, 2004).

4 Relating Basque and English chunks

In this section we will describe the adaptation of the output of the Basque chunker to enable the relating of chunks in Basque to those in English. In particular, we describe in more detail the structure of NPs and PPs (see also (Aldezabal et al., 2003)).

NPs and PPs usually end with a case-morpheme that contains information about case, number and definiteness. For example, in the NP *gizon handi-arekin* (*gizon handi ++arekin*, "with the big man"), the case-morpheme *arekin* at the end is not really syntactically linked to the adjective *handi* (big), but to the whole noun phrase *gizon handi* (big man). In such an example, it makes sense to separate the case-morpheme information from the last word. This information is mapped to a conventional symbolic string, *++soz+ms* in the case of *++arekin* for example, which denotes the associative case (*soz*), the definiteness (*m*) and number singular (*s*).

In NPs and PPs with a common noun as head (see examples 1, 2 and 3 in Figure 7), some quantifiers and determiners may appear before the noun, and others after the noun. Those components used in English or Spanish as marker words are not at the beginning or at the end of chunks in Basque (cf. Figure 6). It is, therefore, more difficult to tag a chunk based on the Part-Of-Speech of the marker word used to produce a chunk.

Errusiako aire armadak
atzo bertan
1.500 kiloko bonbak
jaurti zituen
eskualde hartan

Errusia ++ko aire armada ++ak
atzo bertan
1.500 kiloko bonba ++ak
jaurti zituen
eskualde hura ++an

Figure 6: Suffixes at the end of NPs and PPs

Nevertheless, it is possible to get comparable information from the output of the Basque chunker, which also provides us with morphosyntactic annotations. For example, in the case of quantifiers, personal pronouns and determiners, the chunker is able to detect those components even if they are not the first word in the chunk and it is able to tell when to tag a chunk with this information.

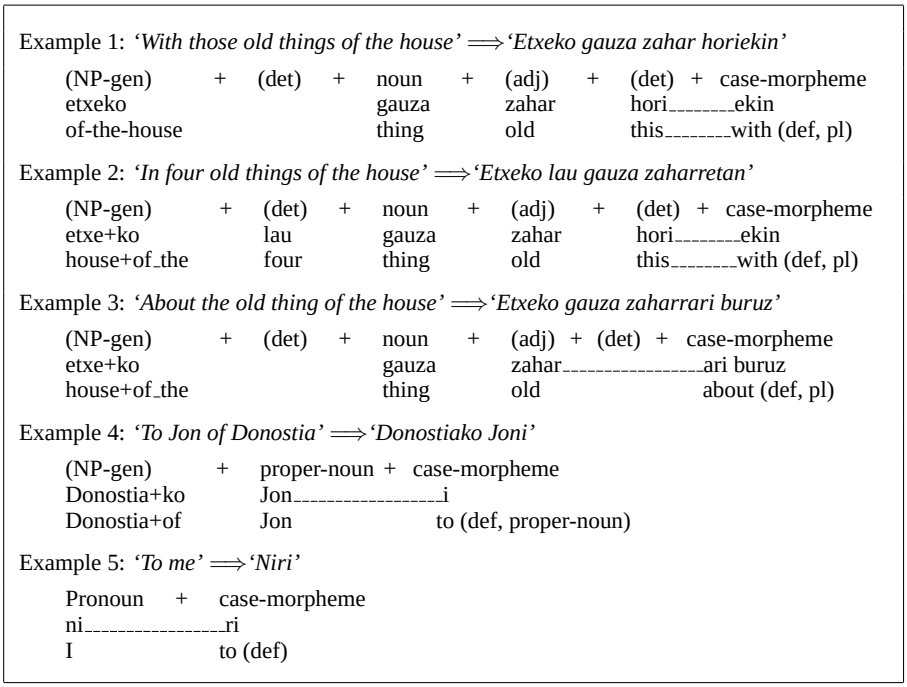

Figure 7: Examples of three main types of NPs (or PPs) in Basque.

5 Experiments

5.1 Data

The corpus used in our experiments was constructed from a translation memory of software manuals, generously supplied by Elhuyar Fundazioa. In total the corpus contains $320,000$ translation units; with 3.2 million words of English and 2.9 million words in the Basque section. The average number of words per unit is thus approximately 10 for English and 9 for Basque.[3] This corpus was filtered based on sentence length and relative sentence length resulting in $276,000$ entries.

5.2 Methodology

From our filtered corpus, we randomly extracted $3,000$ sentences for testing, using the remainder for training. We performed Basque to English translation and plan to do the reverse direction in further experiments. For English, we used the marker-based chunker as described in Section 3.2, and used the EUSMG chunker for Basque. The alignment strategy we employed made use of a (uniform) combination of cognate, word and chunk label information to create the aligned source–target chunks. It should be noted that in these preliminary experiments, we did not perform any kind of specific pre- or post-processing such as named-entity recognition, idiom detection, numbers or date processing.

We evaluated the translations produced by the system using a number of automatic metrics, namely WER (Word-Error Rate), BLEU, PER (Position-Independent WER), Precision and Recall, using only one reference translation. In addition we compare the performance of our MaTrEx system against that of two baseline systems: a word-based SMT system, making use of only the word-level alignments, and a phrase-based SMT system, which uses both words and phrases induced using the method of (Och and Ney, 2004). In both cases, the word alignments are extracted from the training data using the GIZA++ software (Och and Ney, 2003). Decoding is performed with the PHARAOH decoder (Koehn, 2004) with default settings. The English (trigram) language model is trained on the English portion of the training data, using the SRI Language Modeling

[3]The facts that verbs in Basque are usually composed of more than one word (because of inflection and because of the frequent use of periphrastic verbs) and that lexical compound terms are more frequent in Basque means the average sentence length for Basque is not as low as would be expected.

	PER	WER	BLEU	Prec.	Rec.
word-based SMT	69.14	97.35	10.42	43.14	52.17
phrase-based SMT	68.86	89.91	17.31	49.63	52.37
MaTrEx	49.36	68.25	22.23	59.99	55.67

Table 1: Translation results (in %)

Toolkit (Stolcke, 2002) with modified Kneser-Ney smoothing.

5.3 Results and discussion

The results for Basque-English MT are displayed in Table 1. We can see that the MaTrEx system achieves a BLEU score of 22.23%, a 123% relative improvement over the baseline word-based SMT system and a 28.42% relative increase over the baseline phrase-based SMT system. We also see a significant drop in PER and WER (19.5% and 21.66% absolute, respectively, compared with the phrase-based SMT system) and an increase in Precision and Recall. This indicates that the EBMT chunks contribute positively to overall translation quality, enabling the MaTrEx system to outperform both word-based and phrase-based baseline SMT systems. It also indicates that the EBMT chunks allow the system to correctly translate many more of the input words correctly, reflected particularly in the drops in PER and WER.

In addition to our translation experiments, from the set of SMT chunks with frequency greater than 10, we randomly selected a subset consisting of 100 examples and manually evaluated them in terms of quality. The alignments were classified by hand as either being correct (semantically correct), quasi-correct or incorrect (semantically incorrect). The results of this manual evaluation is given in Table 2.

Correct	Quasi-Correct	Incorrect
63.45%	31.10%	5.45%

Table 2: Alignment results

The results in Table 2 we can see that over 94% of the chunk alignments are either correct or quasi-correct, indicating that precision is very high, at least with very frequent chunks.

The quasi-correct chunks are those which are correct in restricted situations; i.e. they are not *a priori* exact equivalents, but in many cases are possible and can even be optimal translations. These quasi-correct alignments cannot be applied everywhere, but in many cases they reflect interesting phenomena, which otherwise, for example when using rule-based approaches, would not be described accurately. For example, if looking at the translation pair *'your computer'* $\Longrightarrow$ *'zure ordenadore ++gel++ms'* out of context it would appear to be incorrect as a genitive case exists in Basque (equivalent to *'of the computer'*)), that does not occur in the English chunk. However, when translating the chunk *'your computer screen'* into Basque, the genitive case information encapsulated within the chunk is necessary for accurate translation.

In order to better understand the contribution the EBMT chunks make to the overall increase in performance of the MaTrEx system configuration, we then performed a similar evaluation setup, this time looking at those chunks which are extracted by both the EBMT chunking and alignment methods and the SMT phrasal extraction methods. The results for this evaluation are given in Table 3.

Correct	Quasi-Correct	Incorrect
84.27%	12.36%	3.37%

Table 3: Alignment Evaluation results for the Intersection of the SMT phrase and EBMT chunk sets

From Table 3, we can see that those chunks that are found by both methods actually are of higher quality than those found by SMT methods alone. Out of 100 of the most frequent chunks occurring in the intersection of the two sets of chunks, 84.27% can be considered as correct translations of each other, 12.36% as quasi-correct and only 3.37% as completely erroneous. This indicates that these higher quality chunks extracted by SMT methods are given a boost in probability when merged with the EBMT data, resulting in the significant improvements in translation quality observed in Table 1.

6 Related work

In this section, we describe some related work carried out within Corpus-Based MT with Basque or related highly inflected languages.

(Nevado et al., 2004) apply two automatic bilingual segmentation techniques (GIATI and Recursive Bilingual Segmentation) based on SMT methods to create new, shorter bilingual segments to be included in a TM database. They have performed this task for the Basque-Spanish pair and for other languages pairs such as Catalan-Spanish or English-Spanish and found that the task is much more difficult for Basque.

Several other studies have been carried out with (somehow) related languages such as German, Czech or Turkish. However, it has to be noted that Basque is more complex morphologically than German or Czech. In German there are four grammar cases represented by means of inflection, seven in Czech and seventeen for Basque.

In (Niessen and Ney, 2001), the authors present work on German-English SMT, where they investigate various type of morphosyntactic restructuring of sentences to harmonize word order in both languages. In particular, they merge German verbs with their detached prefixes, and undo question inversion in both German and English. The results reveal a better exploitation of the bilingual training dataset. Later on (Niessen and Ney, 2004), they annotate a handful of frequent ambiguous German multi word expressions with POS tags, combine idiomatic multi-word expressions into single words, decompose German words into a hierarchical representation using lemmas and morphological tags, and use a MaxEnt model to combine the different levels of representation in the translation model. They conclude that the restructuring operations yielded a large improvement in translation quality, but the morphological decomposition provided only a slight additional benefit.

(Lee, 2004) presents a system for Arabic-English translation, where the Arabic sentences have been presegmented into stems and affixes. Using a phrase-based translation model, Lee concludes that morphological analysis helped only for training corpora up to $350,000$ parallel sentences and that it was not so helpful with a larger corpus consisting of 3.3 million sentences. Note the corpus we used for Basque/English MT is smaller than $350,000$ sentences.

(Cicekli and Güvenir, 2003) learn translation templates from English-Turkish translation examples. They define a template as an example translation pair where some components (e.g. word stems and morphemes) are generalized by replacing them with variables in both sentences. The use of morphemes as units allows them to represent relevant templates for Turkish. The authors believe that their approach is applicable to many pairs of natural languages, assuming sets of morphologically tagged bilingual examples. There is currently no template implementation in our EBMT system, but we plan to integrate related techniques in the near future and to apply them to Basque.

(Al-Onaizan et al., 1999) relates that some transformations of Czech input text, performed with the aim of harmonizing words with equivalents in English, provided a small additional increase in translation quality over basic lemmatization.

(Goldwater and McClosky, 2005), also working with English and Czech, compare four different ways to use only morphological information to improve translation: lemmas, pseudowords, modified lemmas and morphemes. In the case of morphemes they propose a modification in the translation model itself to take advantage of morphological information, rather than simply transforming the output. Word truncation, which requires no morphological information at all, was effective but did not perform quite as well as lemmatization. They found that certain tags were more useful when they treated them as discrete words, while others provided a greater benefit when attached directly to their lemmas. The choice of the best method to use for each class of tags seems to closely correspond with how that particular class of information is expressed in English (either using function words or inflection). That is, for them the main goal of using morphological information is to harmonize Czech and English texts.

The method we used to align chunks (cf. Section 3.3) is able to handle differences in constituent order but does not rely on deep structures. In the future, we plan to additionally use deeper re-structuring techniques to further help the alignment process.

7 Conclusion and future work

In this paper, we presented MATREX, a large-scale modular Data-Driven MT system based on chunking and chunk alignment and applied it to the task of Basque to English Machine Translation. In this system, chunk alignment is performed thanks to a simple dynamic programming algorithm which exploits relationships between chunks. In this context, different kinds of relationships can be considered and even combined. Moreover, this system can be considered a hybrid MT system since it also makes use of aligned phrases extracted using classical SMT techniques.

Experimental evaluation has been performed on Basque to English translation of software manuals data. The results we obtained showed significant improvements on state-of-the-art phrase-based SMT (28% relative increase for BLEU and 21.66% absolute drop in WER). Additionally, we have manually assessed the quality of the chunks aligned by our method.

Future work will aim at completing our experiments in various directions. First, we will investigate the task of English to Basque translation, which is expected to be more difficult since lots of morphological information has to be generated. Then, we will apply our system to other types of data. We would also like to investigate different language pairs, especially Basque-Spanish. This would allow in particular a comparison with Rule-Based MT systems dedicated to the Basque-Spanish pair (Alegria et al., 2005).

At the methodological level, we will try to adapt our Marker-Based chunker to the case of Basque. We will also test the influence of restructuring Basque constituents at a deeper level in order to harmonize with the order of English constituents.

Acknowledgments

This work is partly supported by Science Foundation Ireland (grant number OS/IN/1732) and by an IRCET Ph.D. Fellowship award. Elhuyar Fundazioa is kindly acknowledged for generously providing us with the corpus and the translation memory used in our experiments.

References

J. Abaitua, I. Aduriz, E. Agirre, I. Alegria, X. Arregi, X. Artola, J.M. Arriola, A. Díaz de Ilarraza, N. Ezeiza, K. Gojenola, M. Maritxalar, K. Sarasola, M. Urkia, and J.R. Zubizarreta. 1992. Estudio comparativo de diferentes formalismos sintácticos para su aplicación al Euskara. Technical report, UPV/EHU/LSI, Donostia.

J. Abaitua. 1988. *Complex predicates in Basque: from lexical forms to functional structures*. Ph.D. thesis, University of Manchester.

I. Aduriz and A. Díaz de Ilarraza. 2003. Morphosyntactic disambiguation and shallow parsing in Computational Processing of Basque. *Inquiries into the lexicon-syntax relations in Basque*.

Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, D. Melamed, F. Och, D. Purdy, N. Smith, and D. Yarowsky. 1999. Statistical machine translation. Technical report, JHU Summer Workshop.

I. Aldezabal, M. Aranzabe, A.Atutxa, K.Gojenola, and K.Sarasola. 2003. Patrixa: A unification-based parser for basque and its application to the automatic analysis of verbs. *Inquiries into the lexicon-syntax relations in Basque*.

I. Alegria, A. Díaz de Ilarraza, G. Labaka, M. Lersundi, A. Mayor, K. Sarasola, M. Forcada, S. Ortiz, and L. Padró. 2005. An Open Architecture for Transfer-based Machine Translation between Spanish and Basque. In *Proceedings of the MT Summit X Workshop on Open-Source Machine Translation*, pages 7–14, Phuket, Thailand.

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

R. Brown. 1999. Adding Linguistic Knowledge to a Lexical Example-based Translation System. In *Proceedings of TMI-99*, pages 22–32, Chester, England.

I. Cicekli and H. A. Güvenir. 2003. Learning translation templates from bilingual translation examples. In M. and A. Way, editors, *Recent Advances in Example-Based Machine Translation*, pages 255–286. Kluwer Academic Publishers, Dordrecht, The Netherlands.

P. Goenaga, editor. 1980. *Gramatika Bideetan*. Erein.

S. Goldwater and D. McClosky. 2005. Improving Statistical MT Through Morphological Analysis. In *Proceedings of the EMNLP-05*, pages 676–683, Vancouver, Canada.

N. Gough and A. Way. 2004. Robust Large-Scale EBMT with Marker-Based Segmentation. In *Proceedings of TMI-04*, pages 95–104, Baltimore, Maryland.

T. Green. 1979. The Necessity of Syntax Markers. Two experiments with artificial languages. *Journal of Verbal Learning and Behavior*, 18:481–496.

D. Groves and A. Way. 2005. Hybrid Example-Based SMT: the Best of Both Worlds? In *Proceedings of the ACL-05 Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 183–190, Ann Arbor, Michigan.

D. Groves and A. Way. 2006. Hybrid Data-Driven Models of MT. *Machine Translation, Special Issue on EBMT*.

M. Hearne and A. Way. 2003. Seeing the Wood for the Trees: Data-Oriented Translation. In *Machine Translation Summit IX*, pages 165–172, New Orleans, Louisiana.

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

P. Koehn, F. Och, and D. Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of HLT-NAACL-03*, pages 48–54, Edmonton, Canada.

P. Koehn. 2004. Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In *Proceedings of AMTA-04*, pages 115–124, Washington, District of Columbia.

Y.-S. Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings of HLT-NAACL-04*, pages 57–60, Lisbon,Portugal.

G. Leusch, N. Ueffing, and H. Ney. 2006. CDER: Efficient MT evaluation using block movements. In *Proceedings of EACL-06*, pages 241–248, Trento, Italy.

M. Nagao. 1984. A framework of a mechanical translation between japanese and english by analogy principle. In A. Elithorn and R. Banerji, editors, *Artificial and Human Intelligence*, pages 173–180. North-Holland, Amsterdam, The Netherlands.

F. Nevado, F. Casacuberta, and J. Landa. 2004. Translation memories enrichment by statistical bilingual segmentation. In *Proceedings of LREC-04*, pages 335–338, Lisbon, Portugal.

S. Niessen and H. Ney. 2001. Morpho-syntactic analysis for reordering in statistical machine translation. In *Proceedings of the MT Summit VIII*, pages 247–252, Santiago de Compostela, Spain.

S. Niessen and H. Ney. 2004. Statistical machine translation with scarce resources using morphosyntactic analysis. *Computational Linguistics*, 30(2):181–204.

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

F. Och and H. Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 30(4):417–449.

A. Stolcke. 2002. SRILM – An extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904, Denver, Colorado.

H. Watanabe, S. Kurohashi, and E. Aramaki. 2003. Finding translation patterns from paired source and target dependency structures. In M. Carl and A. Way, editors, *Recent Advances in Example-Based Machine Translation*, pages 397–420. Kluwer Academic Publishers, Dordrecht, The Netherlands.

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