Exploiting portability to build an RBMT prototype for a new source language

Nora Aranberri, Gorka Labaka, Arantza Díaz de Ilarraza and Kepa Sarasola
IXA Group
University of the Basque Country
Manuel Lardizabal 1, 20018 Donostia, Spain
{nora.aranberri,gorka.labaka,a.diazdeilarraza,kepa.sarasola}@ehu.eus

Abstract

This paper presents the work done to port a deep-transfer rule-based machine translation system to translate from a different source language by maximizing the exploitation of existing resources and by limiting the development work. Specifically, we report the changes and effort required in each of the system's modules to obtain an English-Basque translator, ENEUS, starting from the Spanish-Basque Matxin system. We run a human pairwise comparison for the new prototype and two statistical systems and see that ENEUS is preferred in over 30% of the test sentences.

1 Introduction

Building a corpus-based system is undeniably quicker than building a rule-based machine translation (RBMT) system, given the availability of large quantities of parallel text. However, this is often not the case for many language pairs, which makes building a mainstream statistical system suboptimal. Usually, lesser-resourced languages opt for RBMT systems, where language-specific NLP tools and resources are crafted.

Heavy investment and long development periods have been attributed to RBMT systems but (Surcin et al., 2013) pointed out that a large part of the systems' code is reusable. They state that 80% of Systran's code belongs to the analysis module, whereas the remaining 20% is equally divided into transfer and generation. Transfer is language-pair specific, but analysis and generation are built with information about one language only and they are therefore reusable for systems that use those languages. Rapid development of new language pairs benefits from existing resources but also from modular, stable infrastructures where new pairs can be developed by modifying the linguistic data.

An example of RBMT portability attempts for lesser-resourced languages is the Apertium project (Forcada et al., 2011). Apertium is a free/open-source shallow-transfer MT platform. Researchers have been active in porting the system to different language pairs (Peradin et al., 2014; Otte and Tyers, 2011). The system specializes in translation between related languages where shallow transfer suffices to produce good quality translations.

Shallow parsing is sometimes too limited for dissimilar language pairs. Unrelated languages often require a richer and more flexible deeper transfer architecture to tackle differing linguistic features. Examples are (Gasser, 2012) and Matxin (Mayor et al., 2011). In this work we present an attempt to port the deep-transfer RBMT Matxin[1], designed to cope with dissimilar languages.

The remaining work is organized as follows: Section 2 gives a brief overview of the architecture of the Matxin system. Section 3 describes the work done in each of the system's modules. Section 4 provides the results of the new Matxin ENEUS prototype's evaluation. Finally, Section 5 presents the conclusions and future work.

2 General system features

Matxin is a modular RBMT system originally developed to translate from Spanish into Basque (Mayor et al., 2011). It follows the standard three-step architecture, consisting of separate modules

[1]Matxin: https://matxin.sourceforge.net

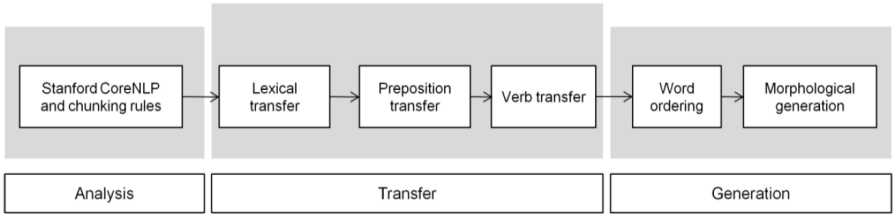


Figure 1: The general Matxin architecture.

for analysis, transfer and generation (Figure 1). It was devised to translate between dissimilar languages, that is, pairs that require deep analysis to enable translation and to do so, it works on dependency trees and chunks, and includes a module for reordering. Because it was developed with the Spanish-Basque pair in mind, the architecture can handle translation from analytic to agglutinative languages, thus dealing with rich morphology.

The portability exercise we present aims to examine the strengths and limitations of the Matxin architecture, by measuring the flexibility of the infrastructure and by specifying the language resource development needed for a new language pair. In particular, we examine the work effort required to change the source language and obtain English to Basque translations.

3 Portability exercise

Given the three-step architecture of Matxin, when modifying the system to translate from a different source language, we first need a completely new analysis module. Next, the transfer rules need to be updated to synchronize the new source with the target language. The generation module is mostly reusable and remains intact. In what follows, we describe the work done in each of the modules.

3.1 Analysis

Packages that analyze text at different levels are available, even more so for mainstream languages such as English. Therefore, what needs to be considered when selecting a package is whether it extracts the relevant information that the generation module will require. The information Matxin needs to translate into Basque is word forms, lemmas, part-of-speech categories, chunks, and dependency trees with named relations.

Note that chunks and dependency trees are different ways of representing sentence structure and both are necessary when translating into Basque. Chunks identify word groupings whereas dependency trees specify the relations between words. In Basque, postpositions are attached to the last word of the chunk they modify.[2] Therefore, chunks allow us to easily identify the word that needs to be flexed. Dependency relations provide the MT system with predicate-argument structures.

Two main contenders were found: Freeling, a rule-based analyzer developed at the Universitat Politécnica de Catalunya (Carreras et al., 2004) and the statistical analysis package developed at Stanford University (de Marneffe et al., 2006). The original architecture uses Freeling for Spanish analysis, and using their English package would make the integration easier, as tags are already known by the system. Yet we carried out a small comparison to opt for the best performing system.

We analyzed 50 sentences with both systems, 25 regular sentences and 25 news headlines. We included both simple and complex sentences showing a wide variety of features and structures. A sentence was to be correctly analyzed if all the lemmas, POS categories and the dependency tree were correctly annotated. 28% of the sentences were correctly analyzed by Freeling and 38% by Stanford. The remaining sentences show one or more errors, which would have varying impact on the translation process. Overall, the number of errors made by Freeling was higher compared to Stanford, 48 and 27 errors respectively. Freeling inserted 18 POS errors whereas Stanford inserted 17 (12 in headings). Dependency tree analyses include errors at different levels. One of the most severe error is the incorrect identification of the root (typically the main verb), which usually leads to the whole translation being wrong. Freeling failed to identify the root in 6 occasions. Stanford, in turn, did not make this type of error.

Overall, we saw that Stanford made fewer errors

[2] We include subject and object case-markers within this class because they are processed equally.

compared to Freeling. The popularity and development activity of this system at the time (Bach, 2012; Sagodkar and Damani, 2012) made us opt for the second package. The initial Spanish analysis component in Matxin was ported to English by integrating a new analysis package and by updating tag equivalences to allow for interoperability.

3.2 Transfer

The most labor-intensive component is the transfer module. In what follows, we examine the dictionaries and grammars that need to be updated in the order in which the architecture applies them.

Lexical transfer

Bilingual dictionaries are the basis for translation and these had to be compiled to include English-Basque equivalences. We used two main sources to build the new dictionaries. First, an English-Basque dictionary was made available for research purposes by Elhuyar, a Basque language technology company. We obtained 16,000 pairs and 1,047 multi-word units from this resource.

The words in the Elhuyar dictionary are probably enough to translate the most frequent English words and understand a general text. However, we decided to try to increase the coverage of Matxin ENEUS with a second resource, that is, WordNet (Miller, 1995). It is a lexical database that was initially built for English, where nouns, verbs, adjectives and adverbs are grouped around cognitive synonyms that refer to the same concept called synsets. Synsets are linked to each other through conceptual and lexical relations, making up a conceptual web of meaning-relations. Even if it was first built for English, WordNets have been developed for other languages, as is the case of Basque (Pociello et al., 2010). Words in different languages share synsets and therefore it is possible to extract equivalences, creating a bilingual pseudo-dictionary. The Basque WordNet has 33,442 synsets that are mapped to their English counterparts. We have paired the variants of each mapped synset in all possible combinations, obtaining over 82,000 pairs after discarding multiword units. These provide us with Basque equivalents for almost 32,000 English lemmas. Even if WordNet was not designed to be used as a dictionary and the equivalences have not been reviewed by an expert, we decided to include them in the system's dictionary even if priority was given to the Elhuyar data. The union of both resources accounts for around 35,000 entries.

```
plane + pos=[NN] → hegazkin + pos=[IZE][ARR] + num=[NUMS]
plane + pos=[NN] → plano + pos=[IZE][ARR] + num=[NUMS]
big + pos=[JJ] → handi + pos=[ADJ_IZO]
big + pos=[JJR] → handi + pos=[ADJ_IZO] + suf=[GRA][KONP]
big + pos=[JJS] → handi + pos=[ADJ_IZO] + suf=[GRA][SUP]
go + pos=[VB] → joan + pos=[ADI]
go + pos=[VBZ] → joan + pos=[ADI]
go + pos=[VBG] → joan + pos=[ADI]
```

Figure 2: Dummy examples of dictionary rules.

The dictionary lists the source lemma and its POS tag and points to the equivalent target lemma together with its POS and morphological information (Figure 2). The information for both languages is the same, but the tag set used is different and generator-dependent. The information in the English tag is itemized into one or more Basque tags. For example, the English *NN* tag referring to common singular nouns is broken down into three separate tags, *IZE*, *ARR* and *NUMS* referring to noun, common and singular, respectively.[3]

The dictionary lists all the possible equivalences gathered from the bilingual resources. Yet, Matxin ENEUS selects the first available equivalent regardless of the context of use. The order in which alternatives are coded in the dictionary is based on frequency in the case of the Elhuyar dictionary and therefore, this already introduces some sort of selection rule. The architecture allows creating context-specific selection rules and other word sense disambiguation (WSD) techniques can be integrated but this is out of the scope of this work.

After the information from the bilingual dictionary is collected, the selected target word is searched for in a semantic dictionary (Díaz et al., 2002) and features added if available.

Preposition transfer

English prepositions are translated into Basque mainly through postpositions. As previously mentioned, these postpositions are attached to the last word of the postpositional phrase (chunk) and the information about it must be moved to the relevant word. To allow for this, prepositions are processed differently, using a purposely-built dictionary. It consists of English prepositions and their Basque postposition equivalents, where the lemmas and morphological tags are specified.

[3]Note that verbs are handled separately, and therefore, all forms carry the same neutral target tag in the dictionary.

[4]Statistics for work in progress when only 20 prepositions have been addressed. The level of ambiguity tends to increase as detailed disambiguation work is done.

	Simple preposition	Unique equivalent	Multiple [4] equivalents	Average ambiguity
English	66	20	46	3.8
Spanish	20	7	13	3.9

Table 1: Statistics for the preposition dictionary.

We have worked with a list of 66 English simple prepositions. We have identified 20 with a unique translation. The remaining 46 have an average of 3.8 translations (ranging from 2 to 10) (Table 1).

Equivalence rule	Example
by → ergative	written by Wilde → Wilde**ek** idatzia
by → instrumental	travel by plane → hegazkin**ez** bidaiatu
by → genitive	a book by Shelly → Shelly**ren** liburu bat
by → genitive + ondoan	by the door → atea**ren ondoan**
by → inessive	by candlelight → kandelaren argipe**an**
by → ablative	hold by the hand → esku**tik** heldu
by → genitive + arabera	by the barometer → barometroa**ren arabera**
by → adlative + time-location genitive	by now → hon**ezkero**
by → + bider	3 multiplied by 2 → 3 **bider** 2
by → + aurretik	drive by your house → zure etxe **aurretik**

Figure 3: Basque equivalences for *by*.

The linguistic work has to identify the different uses for the multiple equivalences, define contexts and write rules that will allow for the appropriate equivalent to be selected (Figure 3). Rules can include different types of knowledge. By default, the design of Matxin allows including elements that are in direct dependency (lemma, POS, morphological, syntactic and semantic features). At the time of write-up, 27 selection rules have been created and further effort is envisaged. If we compare the effort required for the English-Basque pair with the existing work for the Spanish-Basque system, we observe that the list includes 20 simple prepositions, that is, about a third, out of which 7 have a single translation and the ambiguous ones have an average of 3.9 translation options (ranging from 2 to 11). This reveals that the linguistic work necessary to set up the preposition transfer for the new pair is more labor-intensive. Rules are given full priority during selection, and translation equivalences which do not have a selection rule assigned to them are listed by frequency of appearance.

In addition to the equivalence table, Matxin avails of two other sources of information, which are used when no selection rules apply: lexicalized syntactic dependency triplets and verb subcategorisation, both automatically extracted from a monolingual corpus (Agirre et al., 2009).

Lexicalized triplets are groupings of verbs, lemmas and argument cases with which each verb appears in the corpus (Figure 4). In the cases where selection rules are not sufficient, the verb is identified and the lemma of the word to which the postposition needs to be attached is searched for. If the verb-lemma combination is present, the candidate argument cases from the dictionary are checked against the triplets and the first matching selected.

Verb	Lemma	Argument case
eman	unibertsitate	inessive
	Paul	ergative dative
	amore	absolutive partitive
	...	...

Figure 4: Examples of triplets for *eman* (give).

The information contained in lexicalized triplets is often too precise and restrictive. If triplets do not cover the verb-lemma combination, we turn to verb subcategorisation. This resource includes, ordered by frequency, a list of the most common argument case combinations for each verb (Figure 5). The possible postpositions for each of the prepositions that depend on a verb are collected from the dictionary and matched against the subcategorisation information until the combination that suits best is selected.

Verb	Paradigm	Subject case	Arg case	Arg case
suntsitu	subj-dObj	ergative	absolutive	-
	subj	absolutive	-	-
	subj-dObj	ergative	absolutive	instrumental
	...	...	...	...

Figure 5: Examples of verb subcategorization for *suntsitu* (destroy).

Because both Spanish and English use prepositions, the design of Matxin has been adequate for our goal. The preposition dictionary and selection rules were replaced, but verb subcategorisation and lexical triplets were reused, as they are Basque-specific and source-language-independent.

Verb transfer

Basque verbs carry considerable information, such as, person and number of the subject and objects, tense, aspect and mood. In Spanish, information about the objects is not present. In English, the verbs carry even less information: tense, aspect and mood are present, but it is only in the case of present tense third person singular that we know about the subject thanks to the *s* mark attached to

the verb. No reference to the subject (exception above) or objects is made explicit in the verb.

Before applying verb transfer rules, therefore, a set of movement rules needs to collect all the relevant information for Basque verbs from the dependency tree. This difference was partially addressed during the Spanish-Basque implementation. In the case of English, movement rules were modified to include the person and number of the subject and objects, if they explicitly appeared in the text to be translated, as well as the paradigm information obtained from the preposition selection step. Thus, the developer availed of all the source text information required to work on transfer rules. Given the information of subject and objects, the rules are written to identify tense, aspect and mood information from the source verb and replacement rules gather up information to generate an equivalent Basque verb (Figure 6).

I **drive** *my car to university every morning*
input pattern to verb transfer drive[VBP]+[subj1s][dObj3s][iObj00]+[paradigm2]+gidatu
target pattern assigned by grammar gidatu{Asp}{Mod+Asp}{Aux}{Tense}{Subj}{dObj}{iObj}
transformed pattern gidatu{IMPERF}{}{edun}{A1}{subj1s}{dObj3s}
Nik nire autoa **gidatzen dut** *unibertsitatera goizero.*

Figure 6: Dummy example of verb transfer steps.

Verb transfer in the Matxin architecture is carried out using finite-state transducers (Alegria et al., 2005; Mayor et al., 2012). In short, the transducers take the source verb phrase as input, perform a number of replacements and create the final output which is ready for the syntactic and morphological generators to interpret.

We kept the three-step organization of the grammar used in the original language pair.

1. Identification of the Basque verbal schema corresponding to the source verbal chunk.

   We use 21 patterns that we then unify into 5 general schemes corresponding to simple tenses (*works, worked*), compound tenses (*have worked, will work*), continuous tenses (*is working, had been working*), simple tenses preceded by a modal (*should work*), and compound or continuous tenses preceded by a modal (*must have worked*).

2. Resolution of the values for the attributes in each of the Basque schemes.

   A total of 222 replacement rules were written to transfer verbal information into the target language in a format that is interpreted by the generators (Table 2).

3. Elimination of unnecessary information (4 rules in total).

Type	Number of rules
auxiliary verb selection	20
aspect of main verb or auxiliary	65
modal-specific	2
negation	4
paradigm selection and feature assignment	107
tense	24
Total	222

Table 2: Verb transfer rules by type.

When building the prototype, considerable effort was made to ensure wide verb coverage. Most of the tenses in the indicative have been covered, for all four paradigms in Basque (subj, subj-dObj, subj-dObj-iObj, subj-iObj) in the affirmative, negative and questions, for active and passive voices. The imperative was also included.

Work was also done for modals, even if to a more limited extent. Matxin ENEUS identifies the most common modals: ability (*can, could, would*), permission and prohibition (*must, mustnt, can, have to*), advice (*should*) and probability (*may, might, will*) for affirmative and negative cases. Depending on the context, modals acquire a slightly different meaning. At the time of writing, only one sense per modal was covered by the system.

Complex sentences

The modifications mentioned so far describe how simple sentences and their components are treated. However, complex sentences require a more intricate approach. The transfer rules that so far handled finite verbs now need to consider the varying translations of non-finite verbs as well as the permutations subordinate markers require. Also, information movements are directed by more elaborate rules. For Matxin ENEUS, we addressed, in their simplest forms, relative clauses, completives, conditionals and a number of adverbial clauses (time, place and reason).

3.3 Movements

It is the flexibility to move information along the dependency tree-nodes that provides the Matxin architecture with the capacity to tackle dissimilar

languages (Mayor et al., 2011). In this first portability exercise few changes were introduced to the movement rule-sets as basic structures in Spanish and English required similar basic movements. Generally, Basque chunks (verbs aside) consist of a number of lemmas and a last word to which flexion information is attached. Therefore, the basic information movements for both Spanish and English have been (1) preposition information moved to the last word of the chunk, and (2) number and definiteness information of the source chunk moved to the last word of the target chunk.

Additionally, the movement rule-set preceding verb transfer was modified to address certain English-specific structures. For example, English *verb+to* and *verb+ing* structures, e.g. *want to eat*, *intend to go* and similar, require that the second verb is treated differently to how main verbs are treated. This needs to be noted before the verbs arrive in the verb transfer component. In order to do that, a special attribute needs to be passed on to the verb phrase. We tested these two cases and saw that Matxin's design can be appropriate for language-specific structures.

3.4 Generation

The generation component of an RBMT system is usually developed using target-language knowledge only to increase reuse possibilities. In Matxin, the three modules included in the generation component avail of Basque knowledge only (with the exception of the rule-set to address non-canonical source language word order). First, the sentence-level ordering rules in the generation component establish the canonical word order given the elements in the dependency tree.

Secondly, the chunk-level information stored at the chunk-level node is passed on to the word that needs to be flexed. Again, this set of rules avails of target language knowledge only. The rule-set is used as is for different source languages.

Finally, the information collected over the translation process (lemmas and corresponding tag sequences) is passed on to the word generation module, a morphological generator specifically developed for Basque, which was fully reused.

4 System evaluation

We used human evaluation as the main indicator for the prototype's performance. Also, we ran automatic metrics to compare their scores against the human evaluation even when it is known that automatic scores tend to favor SMT systems over RBMT systems because they do not consider the correctness of the output but rather compare the difference between the output and the reference translations (Callison-Burch et al., 2006). And the use of a single reference accentuates this.

To get a perspective on the overall performance, we ran the evaluation for two additional systems, an in-house statistical system, SMTs, and Google Translate, as well as Matxin ENEUS. Our SMT system was trained on a parallel corpus of 12 million Basque words and 14 million English words comprising user manuals, academic books and web data. We implemented a phrase-based system using Moses (Koehn et al., 2007). To better deal with the agglutinative nature of Basque, we trained the system on morpheme-level segmented data (Labaka, 2010). As a result, we need a generation postprocess to obtain real word forms for the decoder. We incorporated a second language model (LM) based on real word forms to be used after the morphological postprocess. We implemented the word form-based LM by using an n-best list following (Olafzer and El-Kahlout, 2007). We first generate a candidate ranking based on the segmented training. Next, these candidates are postprocessed. We then recalculate the total cost of each candidate by including the cost assigned by the new word form-based LM in the models used during decoding. Finally, the candidate list is re-ranked according to the new total cost. This revises the candidate list to promote those that are more likely to be real word form sequences. The weight for the word form-based LM was optimized with minimum error rate training together with the weights for the rest of the models.

We used the same evaluation set for both the human evaluation and the automatic metrics. It is a set of 500 sentences consisting of 250 sentences set aside from the training corpus and 250 out-of-domain sentences from online news sites and magazines. All sentences contain at least one verb, are self-contained and have 5 to 20 tokens.

4.1 Human evaluation

We performed a human evaluation for the three systems mentioned above as part of a wider evaluation campaign. We carried out a pairwise comparison evaluation with non-expert volunteer participants who accessed an evaluation platform on-

line. They were presented with a source sentence and two machine translations. They were asked to compare the translations and decide which was better. They were given the options *1st is better, 2nd is better* and *they are both of equal quality*. Over 551 participants provided responses in the campaign which allowed us to collect over 7,500 data points for the systems we show here. We collected at least 5 evaluations per source sentence for each system-pair (2,500 evaluations per pair).

We adopted the following strategy to decide on a winning system for each evaluated sentence in each system-pair comparison: if the difference in votes between two systems is larger than 2, the system with the highest number of votes is the undisputed winner (System X++). If the difference in votes is 1 or 2, the system scoring higher is the winner (System X+). If both systems score the same amount of votes, the result is a draw (equal).

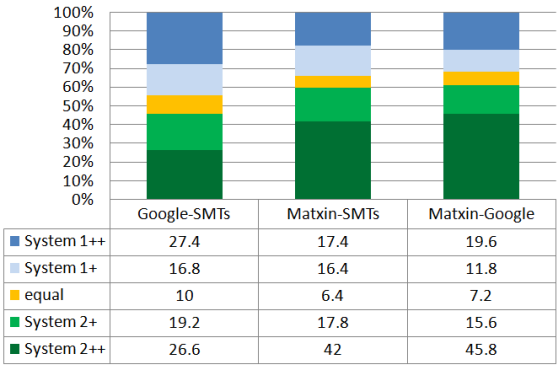
From the evaluations collected (Figure 7), we see that the output of Matxin ENEUS is considered better than its competitors 31-34% of the time, a significant proportion given the prototype's rapid development and limited coverage. This is particularly interesting for hybridization purposes. It would be invaluable to pinpoint the specific structures in which this system succeeds and its specific strengths to guide future hybridization attempts.

SMTs and Google are preferred over the prototype. When compared against each other, the difference in sentences allocated to each system is not significant, with only 8 additional sentences allocated to SMTs (229 vs 221, 50 equal).

4.2 Automatic scores

We provide BLEU and TER scores in Table 3. Low BLEU scores are common for agglutinative target languages when using word-based metrics. A unigram match in these languages can easily equate to a 3-gram match in analytic languages, i.e., a word in Basque often consists of a lemma and number, definiteness and postpositional suffixes.

The human comparison evaluation tells us which translation candidate is preferred over another but it does not capture the distance between their quality. On the other hand, BLEU tries to provide the difference in the overall quality of the systems. Our results seem to suggest that Google has a better overall quality whereas SMTs has more variability in terms of quality, and this leads to our system being preferred for over 40% of the sentences, despite having a lower BLEU score.



	Google-SMTs	Matxin-SMTs	Matxin-Google
System 1++	27.4	17.4	19.6
System 1+	16.8	16.4	11.8
equal	10	6.4	7.2
System 2+	19.2	17.8	15.6
System 2++	26.6	42	45.8

Figure 7: Human comparison results.

System	BLEU	TER
SMTs	8.37	75.893
Google	11.64	72.997
Matxin ENEUS	4.27	83.940

Table 3: Automatic scores.

In the case of Matxin ENEUS, the overall quality seems to be lower, but it still surpasses the statistical systems in over 30% of the sentences, which is not captured by BLEU.

5 Conclusions

We have ported the Matxin deep-transfer rule-based system to work with a different source language and described the requirements and effort involved in the process. More precisely, we have replaced the analysis module with an existing English package which provided us with the necessary lemma, morphological, chunk and dependency information. Most of the work was devoted to the transfer module: we compiled a new bilingual dictionary from an existing electronic version and WordNet; we wrote a preposition-specific dictionary with several disambiguation rules; we wrote the verb transfer grammar and we specified a number of information movements across the dependency tree to address complex sentences and non-finite structures. The generation module was fully reused as the target language remained the same. We estimate that this process required about 8 person month full-time work for a linguist and 1 person month full-time work for a computer scientist, although this estimates will vary depending on each professional's skills and familiarity with the architecture and linguistic work.

Overall, we have gathered evidence that, thanks to its modularity, the use of trees and the flex-

ibility it offers to move information across tree-nodes, Matxin can be a suitable architecture to develop systems for dissimilar languages or those for which deep-transfer is necessary.

We have evaluated the new English-to-Basque prototype by a human pair-wise comparison together with two statistical systems. Although these systems are generally preferred, Matxin ENEUS surpasses statistical competitors in 30% of the cases. Apart from continuing with development work for the new language pair, we now aim to find out the characteristics of those cases, in particular, for hybridization opportunities.

Acknowledgements

The research leading to this work received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007/2013) under REA agreement 302038, FP7-ICT-2013-10-610516 (QTLeap) and Spanish MEC agreement TIN2012-38523-C02 (Tacardi) with FEDER funding.

References

Agirre, Eneko, Aitziber Atutxa, Gorka Labaka, Mikel Lersundi, Aingeru Mayor and Kepa Sarasola. 2009. Use of rich linguistic information to translate prepositions and grammar cases to Basque. *EAMT 2009*, Barcelona, Spain. 58–65.

Alegria, Iñaki, Arantza Díaz de Ilarraza, Gorka Labaka, Mikel Lersundi, Aingeru Mayor and Kepa Sarasola. 2005. An FST grammar for verb chain transfer in a Spanish-Basque MT System. *FSMNLP, Lecture Notes in Computer Science*, 4002:295–296.

Bach, Nguyen. 2012. Dependency structures for statistical machine translation. *SMNLP-2012*, Donostia, Spain. 65–69.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. *EACL-2006*, Trento, Italy. 249–256.

Carreras, Xavier, Isaac Chao and Lluís Padró and Muntsa Padró. 2012. FreeLing: An Open-Source Suite of Language Analyzers. *LREC-2004*, Lisbon.

Díaz de Ilarraza, Arantza, Aingeru Mayor and Kepa Sarasola. 2002. Semiautomatic labelling of semantic features. *COLING-2002*, Taipei, Taiwan.

Forcada, Mikel, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martnez, Gema Ramírez-Snchez and Francis Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation Journal*, 25(2):127–144.

Gasser, Michael. 2012. Toward a rule-based system for English-Amharic translation. *SALTMIL-AfLaT-2012*, Istanbul, Turkey.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondej Bojar, Alexandra Constantin and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. *ACL-2007, Interactive Poster and Demonstration Sessions*, Prague, Czech Republic.

Gorka Labaka. 2010. EUSMT: Incorporating Linguistic Information into SMT for a Morphologically Rich Language. *PhD*, University of the Basque Country.

de Marneffe, Marie-Catherine, Bill MacCartney and Christopher Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. *LREC-2006*, Genoa, Italy.

Mayor, Aingeru, Iñaki Alegria, Arantza Diaz de Ilarraza, Gorka Labaka, Mikel Lersundi and Kepa Sarasola. 2011. Matxin, an open-source rule-based machine translation system for Basque. *Machine Translation Journal*, 25(1):53–82.

Mayor, Aingeru, Mans Hulden and Gorka Labaka. 2012. Developing an Open-Source FST Grammar for Verb Chain Transfer in a Spanish-Basque MT System. *FSMNLP-2012*, Donostia, Spain.

Miller, George. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Oflazer, Kemal and Ilknur Durgar El-Kahlout. 2007. Exploring Different Representation Units in English-to-Turkish Statistical Machine Translation. *WMT-2007*, Prague, Czech Republic. 25–32.

Otte, Pim and Francis. Tyers. 2011. Rapid rule-based machine translation between Dutch and Afrikaans. *EAMT-2011*, Leuven, Belgium. 153–160.

Peradin, Hrvoje, Filip Petkovski and Francis. Tyers. 2014. Shallow-transfer rule-based machine translation for the Western group of South Slavic. *LREC-2012*, Reykjavík, Iceland. 25–30.

Pociello, Elisabete, Aitziber Atutxa and Izaskun Aldezabal. 2010. Methodology and construction of the Basque WordNet. *Language Resources and Evaluation*, 45:121–14.

Sangodkar, Amit and Om Damani. 2012. Re-ordering Source Sentences for SMT. *LREC-2012*, Istambul, Turkey. 2164–2171.

Surcin, Sylvain, Elke Lange and Jean Senellart. 2007. Rapid Development of New Language Pairs at SYSTRAN. *XI MT Summit*, Copenhagen, Denmark. 443–449.