Rule-Based Translation of Spanish Verb+Noun Combinations into Basque

Uxoa Iñurrieta, Itziar Aduriz*, Arantza Díaz de Ilarraza, Gorka Labaka, Kepa Sarasola
IXA NLP group, University of the Basque Country
University of Barcelona
usoa.inurrieta@ehu.eus, itziar.aduriz@ub.edu
a.diazdeilarraza|gorka.labaka|kepa.sarasola@ehu.eus

Abstract

This paper presents a method to improve the translation of Verb-Noun Combinations (VNCs) in a rule-based Machine Translation (MT) system for Spanish-Basque. Linguistic information about a set of VNCs is gathered from the public database Konbitzul, and it is integrated into the MT system, leading to an improvement in BLEU, NIST and TER scores, as well as the results being significantly better according to human evaluators.

1 Introduction

Multiword Expressions (MWEs) constitute a challenging phraseological phenomenon for Natural Language Processing (NLP). They are formed by more than one word, but the whole expression has to be taken into account in order to understand its meaning (Sag et al., 2002). They are very frequent in natural language, but their processing is not straightforward, especially due to their morphosyntactic variability. Furthermore, difficulties multiply when it comes to Machine Translation (MT), since MWEs are not usually translated word for word and, hence, sophisticated processing methods are needed.

In this paper, we will deal with Verb-Noun Combinations (VNCs), and we will explain how MWE-specific linguistic information can be used to improve a rule-based MT system which translates Spanish into Basque, namely Matxin (Mayor et al., 2011). After discussing some related work (Section 2), a brief explanation about Matxin and the way it handles MWEs will be given (Section 3). Then, the experimental setup will be presented (Section 4), and results will be shown (Section 5).

2 Related Work

MWEs are word combinations that need to be treated as a whole in order to get good results in lexically-sensitive NLP tasks (Sag et al., 2002). Not all MWEs are morphosyntactically fixed –there are also semi-fixed and flexible combinations–, which makes their processing a complex task. Some kinds of MWEs, like VNCs, are specially tricky, as they are more likely to have multiple morphosyntactic variants.

Over the last decades, quite a lot of research has been done on MWE identification and extraction (Gurrutxaga and Alegria, 2011; Ramisch, 2015), which is relevant not only for NLP applications but also for other disciplines like Lexicography (Vincze et al., 2011). MWE-specific resources are being developed in a number of languages, as reported by Losnegaard *et al.* (2016) in a survey carried out within the PARSEME COST Action (IC1207).

However, not so much work has been undertaken concerning the multilingual aspects of this phraseological phenomenon, although challenges get bigger when multiple languages are involved. One of the reasons why this happens is that MWEs are not usually translated word for word from one language to another, especially when these languages are from very different typologies (Baldwin and Kim, 2010; Simova and Kordoni, 2013), as with Basque and Spanish[1].

Joint efforts are also being made towards improving Machine Translation systems, for example, within the european QTLeap project (Agirre et al., 2015). Although statistical MT systems already integrate some phraseological knowledge as a consequence of training their models on large

[1]Whereas Spanish is a romance language, Basque is a non-indoeuropean language which belongs to no known family. More details about the main differences between both languages are given in Section 3.

corpora (Ren et al., 2009; Bouamor et al., 2012; Kordoni and Simova, 2014), rule-based systems often get bad results when MWEs are involved, as they tend to translate each word separately. Thus, this kind of expression being so frequent in natural language, MT systems benefit greatly from including phraseological knowledge, and several studies have shown that even the simplest method to process MWEs makes a difference in the system's translation quality (Wehrli et al., 2009; Seretan, 2014).

3 Matxin: Rule-based MT from Spanish into Basque

Matxin (Mayor et al., 2011) is an MT system which translates Spanish into Basque, two long-distance families. As opposed to Spanish, which uses prepositions, Basque is a morphologically rich language where postpositions and cases are used and word order is free. The system is rule-based, mainly because of the scarcity of parallel corpora available in these languages.

Matxin's general architecture is divided into three phases:

1. **Analysis**. The source text is analysed using the FreeLing parser (Padró and Stanilovsky, 2012), which gives morphological information, chunking information, and determines the dependency relationship between words.

2. **Transfer**. The deep syntactic representation of the Spanish sentence is transferred into an equivalent representation in Basque. During this phase, on the one hand, the lexical components in the source language are replaced with their corresponding elements in the target language, and, on the other hand, the structure is also transferred. Specific modules for Spanish-Basque translation are included in this phase, like the one to change prepositions into postpositional information.

3. **Generation**. Firstly, the nodes in each chunk and the chunks themselves are reordered in the sentence from scratch, and postpositional information is added to the chunks when needed. Then, the forms of the words in Basque are created from the labelled lexical elements. The morphological processor used for this purpose is *Morfeus* (Alegría et al., 1996).

3.1 Current MWE handling

At the moment, Matxin uses a very simple method to process MWEs. When an entry in the system's bilingual dictionary is formed by more than one word, the whole expression is treated as a fixed sequence, that is, as if it was a single word. During the transfer phase, the Spanish MWE is replaced by its corresponding Basque word(s), as shown in example (1)[2].

(1) 'A vacancy was filled.'
ES: Se *cubrió_una_plaza*.
Refl covered a vacancy
MT: *Plaza_bat_bete* zen.
vacancy a fill AuxV

In the case of verbal MWEs (including VNCs), verb inflection is taken into account, but the rest of the words have to follow the verb exactly like they appear in the entry. This means that morphosyntactic variation is not processed correctly, neither when identifying the MWE in the source language, nor when translating it into the target language. More details about this are given in Sections 4.1 and 4.2.

(2) 'They filled all vacancies.'
ES: *Cubrieron* todas las *plazas*.
they-covered all the vacancies
MT: *Plaza* guztiak *estali* zituzten.
vacancy all.abs cover AuxV
CT: *Plaza* guztiak *bete* zituzten.
vacancy all.abs fill AuxV

(3) 'He doesn't pay me attention.'
ES: No me *hace_caso*.
not me.IndObj he-does attention
MT: Ez nau *kasu_egiten*.
not AuxV.DObj attention do
CT: Ez dit *kasu(rik) egiten*.
not AuxV.IndObj attention.part do

In example (2), the VNC *cubrir plazas* is not identified as a MWE and, as a consequence, the wrong lexical choice is done when translating it into Basque. In example (3), on the other hand, the VNC is identified well, but the grammatical information of its Basque translation is incorrect, because the system ignores that the Basque VNC needs an indirect object instead of a direct one.

[2] In examples, we use ES for the Spanish text to be translated, MT for the result of the MT system, and CT for the correct Basque translation.

4 Experimental setup

The VNC set used for the experiment consisted of 92 combinations taken from the Konbitzul database[3], where a number of Spanish VNCs and their Basque translations are collected along with linguistic data. The combinations in Konbitzul were gathered from several sources; the set we used here originally came from the Elhuyar Spanish-Basque dictionary[4] and was then analysed and tailored to meet the requirements of the database. According to the information in Konbitzul, 57 out of the 92 combinations were morphosyntactically semi-fixed, while the resting 26 were completely flexible.

Concerning the corpus, 4,991 sentences were selected from a bigger parallel corpus made of cross-domain texts collected by web-crawling and automatically aligned between Spanish and Basque. It was expressly crafted for this experiment, meaning that it did not consist of random sentences but of selected sentences containing: either instances of the Spanish VNCs in our set (Example 4), or both the verb and the noun of a given VNC in our set, but not being part of the VNC in this context (Example 5). This allowed us to test the performance of the MT system both when the VNC needed to be processed as a whole and when the verb and the noun needed to be translated separately.

(4) Iban *dando voces* por la calle.
they-went giving voices on the street
'They were shouting on the street.'

(5) Aquellas *voces* le *dieron* una pista.
those voices her.IndObj gave a clue
'Those voices gave her a clue.'

The information in Konbitzul was first used to help to identify instances of the VNCs when analysing the source text (Section 4.1), and then to transfer the source sentence into the target language (Section 4.2). Therefore, the identification of VNCs was done within the Analysis phase of the translation procedure, and their translation was done within the Transfer phase, the Generation phase not needing any special adaptation for MWE handling (Section 3).

[3] http://ixa2.si.ehu.eus/konbitzul
[4] http://hiztegiak.elhuyar.eus/

4.1 Identifying the Spanish VNCs

In Konbitzul, comprehensive linguistic information is specified for the VNC set we use here, including some features specifically analysed for NLP purposes. The morphosyntactic classification is first used, according to which the VNCs can be of three types: fixed, semi-fixed or flexible.

When a given VNC is classified as flexible, it means that, concerning morphosyntax, the noun and the verb work as any other noun and verb in the sentence, that is, they can have as many variants as any non-phraseological VNC.

(6) Me *da* muchísimo *miedo*.
me.IndObj gives very-much fear
'It scares me very much.'
¡Qué *miedo* me *da*!
what fear me.IndObj gives
'How scary (I find it)!'

On the other hand, when the VNC is classified as semi-fixed, some restrictions are needed in order to distinguish occurrences of the VNC from other sentences where the verb and the noun are present but should not be treated as an MWE.

(7) *Estoy* muy *de acuerdo*.
I-am very of agreement
'I agree very much.'
Estoy harta *del acuerdo*.
I-am fed-up of-the agreement
'I'm fed up with the agreement.'

In example (7), two sentences are shown, both of which contain the verb *estar* and the noun *acuerdo* preceded by the preposition *de*. In the first sentence, those words constitute a MWE (*estar de acuerdo*, 'agree'), but not in the second one, where the noun phrase (NP) has a determiner. By restricting determiners from the NP in the VNC, the system identifies a MWE in the first sentence but not in the second one[5].

For the identification task, we followed the same procedure as the one used in (Inurrieta et al., 2016). First of all, the method currently used by Matxin is run, that is: word sequences are searched for against entries in the database, taking verb infletion into acount, but not considering the potential variability of the rest of the elements.

[5] All restrictions are collected and explained in (Inurrieta et al., 2016).

Then, automatically-produced chunking information and syntactic dependencies are used, and morphosyntactic restrictions specified in Konbitzul are applied (Example 7).

4.2 Translating the VNCs into Basque

Concerning translation, Konbitzul classifies the Spanish VNCs according to what needs to be changed when translating them into Basque: lexicon, grammar, or both lexicon and grammar.

For the VNCs needing lexical treatment, Basque equivalents are specified for the verb and the noun in Spanish. This information is integrated into Matxin, so that, when a VNC is identified, the system does not translate it regularly (Example 8).

(8) 'The topic aroused interest.'
ES: El tema *despertó interés*.
the topic awakened interest
MT: Gaiak *interesa esnatu* zuen.
topic.$_{erg}$ interest awaken $_{AuxV}$
CT: Gaiak *interesa piztu* zuen.
topic.$_{erg}$ interest turn-on $_{AuxV}$

On the other hand, for the VNCs needing special grammatical treatment, the features that need to be taken into account are specified. For those cases, exceptional rules are added within the Transfer phase, so that the specified feature(s) is/are not translated regularly.

The features specified in the database are:

- Cases or postposition marks of the NPs
- Determiner irregularities
- Number and definiteness of the NPs
- Syntactic relations of the verbs and the NPs
- Postpositions of open slots

In example (9), for instance, the Basque NP needs a postposition other than the one automatically given as a translation of the Spanish preposition. Furthermore, it needs to be indefinite, but it would be translated as definite if no special rule was applied.

(9) 'She treats me with respect.'
ES: Me *trata con respeto*.
she-me.$_{DObj}$ treats with respect
MT: *Errespetuarekin tratatzen* nau.
respect.$_{soc}$ treat $_{AuxV}$
CT: *Errespetuz tratatzen* nau.
respect.$_{ins}$ treat $_{AuxV}$

When it comes to example (10), the noun in the Spanish VNC is preceded by a preposition, and this prepositional phrase works as a modifier of the verb. On the other hand, the combination has an object which works as an open slot, that is, an element which is always present but can be filled with any NP. In the Basque translation, the object of the verb in the VNC is actually the noun in the VNC, and the open slot is a postpositional phrase which works as a modifier. Therefore, both the syntactic relation and the postposition of the open slot need special rules to be processed correctly.

(10) 'They miss him.'
ES: Lo *echan en falta*.
him.$_{IndObj}$ throw in lack
MT: *Faltan botatzen* dute.
lack.$_{ine}$ throw $_{AuxV}$
CT: Haren *falta sumatzen* dute.
his lack.$_{abs}$ feel $_{AuxV}$

5 Results

After integrating all the linguistic information into Matxin, the system was evaluated using three automatic evaluation metrics: BLEU (Papineni et al., 2002), NIST (Doddington, 2002) and TER (Snover et al., 2006). Evaluation was carried out without casing, and two systems were compared: (a) the original one, Matxin, and (b) the same system with VNC-specific information.

System	BLEU	NIST	TER
Matxin	7.28	3.88	84.36
Matxin-VNC	7.50	3.90	84.27

Table 1: BLEU, NIST and TER scores obtained by Matxin with and without VNC-specific information

As shown in Table 1, all scores improve when VNC-specific information is used. The greatest improvement is obtained in BLEU score (0.22 points), and results are statistically significant according to paired bootstrap resampling ($p>0.05$). It must be noted that BLEU scores are low for Spanish-Basque, and this result means a relative increase of 3.02%.

5.1 Human evaluation

Apart from using automatic evaluation metrics, three human evaluators were also given a representative sample of the sentences translated differently by both systems and were asked to compare

them. All evaluators were Spanish and Basque native speakers: two of them (A and B) were linguists, whereas the third one (C) had no linguistic background.

System	A	B	C
Matxin-VNC	77.50%	77.50%	46.50%
Matxin	6.50%	8%	40.50%
No preference	16%	14.50%	13%

Table 2: Scores by three human evaluators

Although scores clearly show that the system with VNC-specific information gets better results, they also suggest that improvements are much more evident for linguists than for native speakers with no linguistic background (Table 2). In fact, 43.52% of the evaluation set led to disagreements among annotators, but 78.57% of these (33% of the whole set) were cases in which both linguists said the new system performed better while annotator C chose the other translation.

Taking into account that only a few combinations were tested and the corpus used was specifically prepared based on those combinations, it can be foreseen that the overall improvement this method would produce on large corpora would not be as significant. However, as the kind of linguistic information we chose is proved to have a positive effect on the system's output, we conclude that this methodology is relevant and useful for further investigation.

6 Conclusion

In the experiment presented in this paper, linguistic information was used to improve the translation of VNCs in Matxin, a rule-based MT system for Spanish-Basque. MWE-specific linguistic information was gathered from Konbitzul, a database collecting data about a list of VNCs, and this information was then used both for the identification of idiomatic VNCs in Spanish and for their translation into Basque.

After integrating information about 92 VNCs into Matxin, the system was evaluated on a 4,991-sentence cross-domain corpus, using three automatic metrics: BLEU, NIST and TER. The score that raised the most was BLEU, with an increase of 0.22 points (3.02%). A human evaluation was also carried out, where the improvement became even more evident, even if it also suggested that linguists are more likely to notice improvements than native speakers with no linguistic background.

It must also be noted that the corpus we used here was specifically crafted for this experiment, which means that the improvement would probably not be as significant in a bigger general corpus. However, results are positive as a start, and we intend to keep investigating how this methodology can be enhanced. The next step will be to add more VNCs and test them in bigger corpora, so that conclusions can be drawn at a greater scale.

Acknowledgments

Uxoa Iñurrieta's doctoral research is funded by the Spanish Ministry of Economy and Competitiveness (BES-2013-066372). The work was carried out in the context of the SKATeR (TIN2012-38584-C06-02), EXTRECM (TIN2013-46616-C2-1-R) and TADEEP (TIN2015-70214-P) projects.

References

Eneko Agirre, Iñaki Alegría, Nora Aranberri, Mikel Artetxe, Ander Barrena, António Branco, Arantza Díaz de Ilarraza, Koldo Gojenola, Gorka Labaka, Arantxa Otegi, et al. 2015. Lexical semantics, Basque and Spanish in QTLeap: Quality Translation by Deep Language Engineering Approaches. *Procesamiento del Lenguaje Natural*, 55:169–172.

Iñaki Alegría, Xabier Artola, Kepa Sarasola, and Miriam Urkia. 1996. Automatic morphological analysis of basque. *Literary and Linguistic Computing*, 11(4):193–203.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing, Second Edition*, pages 267–292. Chapman and Hall/CRC.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Identifying bilingual multiword expressions for statistical machine translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 674–679.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

Antton Gurrutxaga and Iñaki Alegria. 2011. Automatic extraction of nv expressions in basque: basic issues on cooccurrence techniques. In *Proceedings*

of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World, pages 2–7. Association for Computational Linguistics.

Uxoa Inurrieta, Arantza Diaz de Ilarraza, Gorka Labaka, Kepa Sarasola, Itziar Aduriz, and John Carroll. 2016. Using linguistic data for english and spanish verb-noun combination identification. In *The 26th International Conference on Computational Linguistics (COLING 2016): Technical Papers*, pages 857–867.

Valia Kordoni and Iliana Simova. 2014. Multiword expressions in machine translation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1208–1211.

Gyri Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sacha Bargmann, and Johanna Monti. 2016. Parseme survey on MWE resources. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Paris, France. European Language Resources Association (ELRA)*.

Aingeru Mayor, Iñaki Alegría, Arantza Díaz De Ilarraza, Gorka Labaka, Mikel Lersundi, and Kepa Sarasola. 2011. Matxin, an open-source rule-based machine translation system for basque. *Machine translation*, 25(1):53–82.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2473–2479.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhug. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Carlos Ramisch. 2015. *Multiword Expressions Acquisition*. Springer.

Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (ACL 2009*, pages 47–54. Association for Computational Linguistics.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15. Springer.

Violeta Seretan. 2014. On collocations and their interaction with parsing and translation. In *Informatics*, volume 1, pages 11–31. Multidisciplinary Digital Publishing Institute.

Iliana Simova and Valia Kordoni. 2013. Improving English-Bulgarian statistical machine translation by phrasal verb treatment. In *Proceedings of MT Summit XIV Workshop on Multi-word Units in Machine Translation and Translation Technology*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

Orsolya Vincze, Estela Mosqueira, and Margarita Alonso Ramos. 2011. An online collocation dictionary of Spanish. In *Proceedings of the 5th International Conference on Meaning-Text Theory*, pages 275–286.

Eric Wehrli, Violeta Seretan, Luka Nerima, and Lorenza Russo. 2009. Collocations in a rule-based mt system: A case study evaluation of their translation adequacy.