A generalised alignment template formalism and its application to the inference of shallow-transfer machine translation rules from scarce bilingual corpora

Víctor M. Sánchez-Cartagena[a,b,*], Juan Antonio Pérez-Ortiz[a], Felipe Sánchez-Martínez[a]

[a] *Dep. de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, E-03071, Alacant, Spain*
[b] *Prompsit Language Engineering, Av. Universitat s/n. Edifici Quorum III. E-03202 Elx, Spain*

Abstract

Statistical and rule-based methods are complementary approaches to machine translation (MT) that have different strengths and weaknesses. This complementarity has, over the last few years, resulted in the consolidation of a growing interest in hybrid systems that combine both data-driven and linguistic approaches. In this paper we address the situation in which the amount of bilingual resources that is available for a particular language pair is not sufficiently large to train a competitive statistical MT system, but the cost and slow development cycles of rule-based MT systems cannot be afforded either. In this context, we formalise a new method that uses scarce parallel corpora to automatically infer a set of shallow-transfer rules to be integrated into a rule-based MT system, thus avoiding the need for human experts to handcraft these rules.

Our work is based on the alignment template approach to phrase-based statistical MT, but the definition of the alignment template is extended to encompass different generalisation levels. It is also greatly inspired by the work of Sánchez-Martínez and Forcada published in 2009 (Journal of Artificial Intelligence Research 34) in which alignment templates were also considered for shallow-transfer rule inference. However, our approach overcomes many relevant limitations of that work, principally those related to the inability to find the correct generalisation level for the alignment templates, and to select the subset of alignment templates that ensures an adequate segmentation of the input sentences by the rules eventually obtained. Unlike previous approaches in literature, our formalism does not require linguistic knowledge about the languages involved in the translation. Moreover, it is the first time that conflicts between rules are resolved by choosing the most appropriate ones according to a global minimisation function rather than proceeding in a pairwise greedy fashion.

Experiments conducted using five different language pairs with the free/open-source rule-based MT platform Apertium show that translation quality significantly improves when compared to the method proposed by Sánchez-Martínez and Forcada (2009), and is close to that obtained using handcrafted rules. For some language pairs, our approach is even able to outperform them. Moreover, the resulting number of rules is considerably smaller, which eases human revision and maintenance.

Keywords: machine translation, transfer rule inference, hybrid machine translation

1. Introduction

Machine translation (MT) can be defined as the process carried out by a computer to translate a text in a natural language, the *source language* (SL), into another language, the *target language* (TL). According

*Corresponding author
Email addresses: vmsanchez@prompsit.com (Víctor M. Sánchez-Cartagena), japerez@dlsi.ua.es (Juan Antonio Pérez-Ortiz), fsanchez@dlsi.ua.es (Felipe Sánchez-Martínez)

Accepted author manuscript.
Published in Computer Speech and Language 32 (2015) 46–90
http://dx.doi.org/10.1016/j.csl.2014.10.003

to the kind of knowledge used in their development, MT systems may be said to be corpus based or rule based, although hybrid approaches (Thurmair, 2009; Costa-Jussà and Farrús, 2014) are also possible.

On the one hand, corpus-based approaches use large collections of parallel texts as the source of knowledge. A *parallel text* is a text that is placed alongside its translation into another language; a collection of parallel texts is usually referred to as a *parallel corpus*. Statistical machine translation (SMT) (Koehn, 2010) is currently the leading paradigm in corpus-based MT. SMT systems can be built with little human effort, provided that a parallel corpus of sufficient size, of the order of tens of millions of words in each language, is available (Och, 2005).

Rule-based MT (RBMT) systems (Hutchins and Somers, 1992) meanwhile use linguistic resources, such as morphological dictionaries, bilingual dictionaries and structural transfer rules,[1] to describe the translation process. Building an RBMT system usually implies a considerable investment in the development of these resources, some of which can only be developed by trained experts. As the availability of a parallel corpus is not necessary for RBMT systems, the RBMT approach is the approach of choice when building MT systems for the translation between under-resourced language pairs (e.g. Breton–French, Icelandic–English, Kazakh–Tatar) for which large parallel corpora are not readily available.

RBMT systems usually work by analysing —at the morphological, syntactic or semantic level— the SL text in order to build an intermediate representation (IR), which is the basis for the generation of the translation into the TL. Depending on the nature of this IR, RBMT systems can be said to be *interlingua* based or *transfer* based. Interlingua-based RBMT systems use language-independent IR; this makes analysis and generation difficult —and almost impossible for broad domains— but avoids the need for transfer. Transfer-based RBMT systems use language-dependent IRs and include a transfer stage which transforms the SL IR into a TL IR by applying lexical and structural transfer rules. Since they are language-dependent, the IRs used in transfer-based RBMT are much easier to develop than those used by interlingua-based systems, thus making transfer-based RBMT the leading approach in RBMT.

Transfer-based RBMT systems can in turn be classified according to the complexity of the IR used into *shallow-transfer*, *syntactic-transfer*, and *semantic-transfer* RBMT systems. In this paper, we focus on *shallow-transfer* RBMT systems, which are those that perform a shallow-syntactic analysis of the SL, i.e. they do not perform full syntactic parsing and do not build a parse tree. This signifies that the IR they use is as simple as a sequence of *lexical forms* (lemma, lexical category and morphological inflection information) of the words to be translated; transfer rules usually split this sequence into *chunks* (groups) of lexical forms whose elements are processed together.

Shallow-transfer RBMT systems use bilingual dictionaries for lexical transfer, and shallow-transfer rules for structural transfer. These rules match chunks of lexical forms in the SL and produce TL lexical forms as their output. They operate on the lexical forms they have matched, regardless of the SL lexical forms matched by other rules, often with no interaction between the rules. The Apertium shallow-transfer MT engine (Forcada et al., 2011) has recently been used for the development of several language pairs, such as Breton–French (Tyers, 2010), Italian–Catalan (Toral et al., 2011) or Icelandic–English (Brandt et al., 2011), to name but a few.

This paper presents a new approach with which to automatically learn shallow-transfer MT rules from small parallel corpora,[2] which, inspired by the work by Sánchez-Martínez and Forcada (2009), uses *alignment templates* (AT),[3] like those initially used in SMT (Och and Ney, 2004), and overcomes the main limitations of their method (see Section 2.1 for a thorough description of these limitations): (i) its conservative approach which prevents the appropriate generalisation level for the ATs from which rules are generated from being found; (ii) as a particular case of the latter, its inability to perform context-dependent lexicalisations to give a different treatment to those words that are incorrectly translated by more general ATs; and (iii) the deficient selection of the ATs to eventually be used for the generation of rules which finally results in

[1] Structural transfer rules define the transformations needed to convert SL groups of words into their TL counterparts.

[2] Barely a few hundreds or thousands of sentences: a small size compared to the amount of parallel corpora required to train competitive SMT systems.

[3] Alignment templates are a generalised version of the phrase pairs used in phrase-based SMT (Koehn, 2010) which use word classes rather than words and also include word alignment information.

rules that prevent the application of other, more convenient rules. The inferred rules are compatible with the formalism used by Apertium (Forcada et al., 2011) to code shallow-transfer rules, may be modified (post-edited) by human experts and can co-exist with hand-written rules.

The remainder of the paper is organised as follows. The following section presents a brief description of the approach by Sánchez-Martínez and Forcada (2009), stressing its main limitations and summarising how they are overcome in our approach. Sections 3 and 4, respectively, introduce the AT formalism used in our approach and the method employed to extract ATs using this formalism. Section 5 then provides a review of the related approaches found in literature. The experiments conducted to test our approach are presented in Section 6, whereas the results obtained are reported and discussed in Section 7. The paper ends with some concluding remarks, two appendices containing low-level details to ensure the reproducibility of the experiments and a table of acronyms for the benefit of the reader.

2. Previous Approach

The approach by Sánchez-Martínez and Forcada (2009) is based on the alignment template (AT) approach (Och, 2002; Och and Ney, 2004) initially proposed in the context of SMT. An AT performs a generalisation of bilingual phrase[4] pairs (pairs of segments which are mutual translations) by using word classes rather than the words themselves. Sánchez-Martínez and Forcada (2009) adapted the AT approach for their application in RBMT by extending the ATs with a set of restrictions in order to control their application as shallow-transfer rules. An *extended* AT (henceforth, EAT) is defined as a tuple $z = (S, T, A, R)$, consisting of a sequence S of SL word classes, the corresponding sequence T of TL word classes, a set A of pairs of word class indexes (i, j) with the alignment information between the word classes in the two sequences, and a set R of restrictions over the TL inflection information that the words to translate need to meet. These restrictions prevent, for example, an AT producing a TL masculine noun from an SL feminine noun from being applied to an SL noun whose translation, according to the bilingual dictionary of the system, does not have a feminine gender.

The method by Sánchez-Martínez and Forcada (2009) needs a human-designed set of *lexicalised units*. This set is made up of both the SL and TL lexical forms (usually corresponding to closed lexical categories) involved in lexical changes and which should not be generalised. EATs are then learnt from a parallel corpus by using the following steps:

1. Analyse and convert both sides of the parallel corpus into the IR used by the RBMT system to be used; in the case of Apertium, sequences of lexical forms.

2. Apply classical statistical, word-translation models (Brown et al., 1993; Vogel et al., 1996) in order to obtain word alignments in both translation directions, and then symmetrise the alignments obtained using the refined intersection method proposed by Och and Ney (2003).

3. Extract bilingual phrase pairs that are compatible with the set of alignments (Koehn, 2010, Sec. 5.2.3).

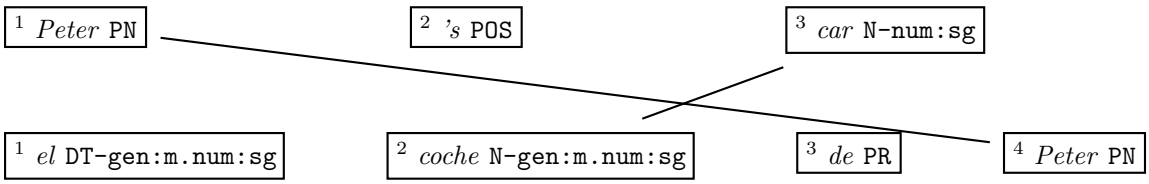
4. Remove those bilingual phrase pairs that cannot be reproduced by the RBMT system in which the transfer rules will be used because, according to the bilingual dictionary, the translation equivalent of at least one lexical form not present in the set of lexicalised units differs from that observed in the bilingual phrase.

5. Replace lexical forms with word classes. Word classes represent the lexical category and morphological inflection information of the corresponding words. They are obtained by removing the lemma from each lexical form that is not present in the set of lexicalised units provided by the user.

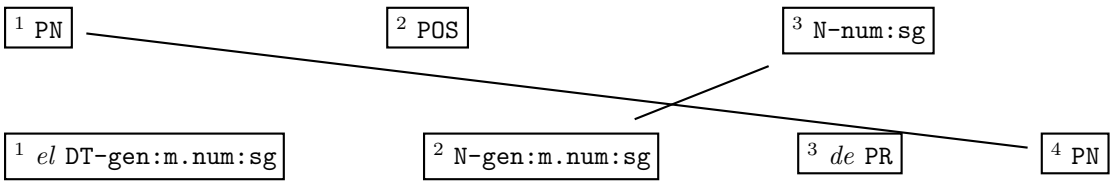
6. Infer the set of restrictions R by looking up in the bilingual dictionary the lexical forms that do not belong to the set of lexicalised units.

[4] In order to use the same terminology as that used in SMT (Koehn, 2010) we refer to segments as phrases, despite the fact that they are not necessarily syntactic constituents.

Bilingual phrase pair p:



EAT z:



$r_1 = \{\}; r_2 = \{\}; r_3 = \{\texttt{gen:m}\}$

Figure 1: English–Spanish bilingual phrase pair p and EAT z obtained with the method devised by Sánchez-Martínez and Forcada (2009). To obtain z, the lexical forms in p are replaced with word classes. These word classes are obtained by removing the lemma from the lexical forms, with the exception of those in the set of lexicalised units provided by the user (in the example, prepositions and determiners). Restrictions r_1 and r_2 are empty, whereas r_3 forces the EAT to be applied only to those SL nouns that are masculine in the TL. `PN`, `POS`, `N`, `DT` and `PR` stand for proper noun, possessive ending, noun, determiner and preposition, respectively. `gen:m` indicates that the gender of the word is masculine, and `num:sg` that its number is singular. Lines between word classes or lexical forms represent alignments; lemmas appear in italics. With this EAT, the translation into Spanish of the English phrase *Fran's pen* would be *el bolígrafo de Fran*.

The resulting set of EATs is then used to generate structural shallow-transfer rules after removing those EATs whose frequency is below a threshold that is empirically determined on a development parallel corpus.

During translation, the actions that need to be performed in order to build each TL lexical form depend on the type of word class:

- if the TL word class includes a lemma (e.g. *de* `PR`, the Spanish preposition *de*), because the corresponding lexical form belongs to the set of lexicalised units, it is introduced unchanged.
- if the TL word class does not include a lemma (e.g. `N-gen:m.num:sg`; noun, masculine, singular), the lemma to be included in the TL lexical form is obtained by looking up the SL lexical form that is matched by the SL word class to which the TL word class is aligned in the bilingual dictionary. The TL lemma is then accompanied by the morphological inflection attributes in the TL word class.

Figure 1 shows an English–Spanish bilingual phrase pair and the EAT obtained from it. This EAT matches any proper noun (`PN`) followed by a possessive ending (`POS`), and a singular (`num:sg`) noun (`N`). As an output, this EAT produces a masculine (`gen:m`) singular determiner (`DT`) with the lemma *el*, a masculine singular noun whose lemma is obtained by looking up the lemma of the noun that is matched in the SL in the bilingual dictionary, a preposition (`PR`) with the lemma *de*, and a proper noun whose lemma is retrieved from the bilingual dictionary by looking up the lemma of the proper noun that is matched in the SL. Restriction r_3 prevents the EAT from being applied when the noun is not masculine in the TL, which would produce a TL translation with no gender agreement between the determiner and the noun. With this EAT, the translation into Spanish of the English phrase *Fran's pen*, with SL IR w_1 =*Fran* `PN`, w_2 ='*s* `POS`, w_3 =*pen* `N-num:sg`, would be *el bolígrafo de Fran*, with TL IR w'_1 =*el* `DT-gen:m.num:sg`, w'_2 =*bolígrafo* `N-gen:m.num:sg`, w'_3 =*de* `PR`, w'_4 =*Fran* `NP`.

2.1. Main limitations

Although this method infers shallow-transfer rules capable of producing translations that are close to those produced with hand-written rules and provides better results than SMT systems trained on the same parallel corpus extended with the bilingual dictionary of the RBMT system (Sánchez-Martínez and Forcada, 2009, Sec. 5.2.1), it has three main limitations which we describe below. The first two limitations are inherent to the expressiveness of the EATs they use, whereas the third is a limitation of the aforementioned authors' learning algorithm.

First limitation: partial generalisation. The definition of word classes is not sufficiently flexible to permit EATs with different generalisation levels. The most general word classes are obtained by removing the lemma from the lexical forms and they therefore take into account the lexical category and all the inflection information (e.g. gender, number, verb tense, person, case, etc.). This often results in having to learn several EATs in order to describe the translation of the same linguistic phenomenon. For instance, adjectives in English are placed before the noun, whereas in Spanish they are usually placed after the noun. In order to properly translate an adjective followed by a noun into Spanish, Sánchez-Martínez and Forcada (2009) need to learn the four EATs shown in Figure 2; these EATs only differ in the morphological inflection information (gender and number) of the lexical forms they match. Note that if more general word classes were used, so that all adjectives (or nouns) were assigned to the same word class regardless of the inflection information, the adjective–noun reordering could be described with the EAT shown in Figure 3. In that Figure, the morphological inflection attributes `gen:*` and `num:*` in the SL word classes mean that they match any gender and number, respectively. The morphological inflection attribute `num:`$\$_{\mathrm{s}}^{1}$ in a TL word class means that the TL lexical form produced as a translation takes the value of the attribute with the same name in the first SL lexical form matched (more information on the word classes used is provided in Section 3).

In general, we solve the partial generalisation limitation by using word classes with different levels of generalisation and exploiting the information contained in the bilingual phrase pairs to decide, in a context-dependent manner, the generalisation level of the EATs, that is, the morphological inflection attributes that contain the wildcard value (*) that matches any possible value. In our approach, multiple EATs, with different levels of generalisation, are generated from each bilingual phrase pair. The set of EATs to be used —and therefore the appropriate generalisation level to be used to describe the translation of the different linguistic phenomena found in the training corpus— is then automatically determined by selecting the minimum number of EATs that are needed to reproduce the bilingual phrase pairs from which the EATs are obtained. In order to deal with the complexity of choosing that minimum set of EATs when working with all the bilingual phrase pairs extracted from the corpus, the problem is posed as an optimisation problem by defining a set of inequations which are solved using integer linear programming methods (Garfinkel and Nemhauser, 1972). Our approach is the first in literature (see Section 5) in which the problem of automatically inferring transfer rules is reduced to finding the optimal value of a minimisation problem.

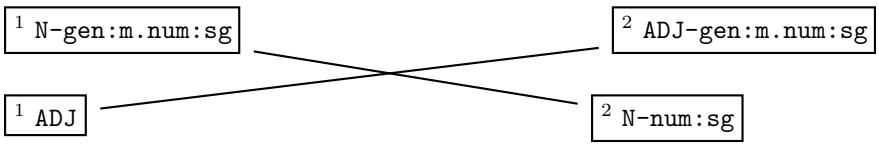
Being able to use word classes with different generalisation levels implies that our method needs fewer examples to learn common structural transformations between the SL and the TL. In addition, having more general EATs makes it easier for linguists to revise the inferred rules and for these rules to be combined with hand-written rules.

Second limitation: no context-dependent lexicalisations. The way in which word classes are defined by Sánchez-Martínez and Forcada (2009), that is, by using a set of lexicalised units, not only prevents better generalisations from being created, as explained above, but also prevents context-dependent lexicalisation from taking place. Context-dependent lexicalisation would permit a different treatment to be given to those words that, in a given context, are not properly translated by more general EATs. For instance, in Spanish some adjectives —called *prepositive* adjectives[5]— are usually placed before the noun, e.g. *gran hombre*,[6] instead of

[5] Although only a reduced set of adjectives are always prepositive in Spanish, all adjectives can be used prepositively in poetry and literature. Postpositive adjectives are unusual in English, but they can be found in phrases such as *body politic*, *queen consort* or *time immemorial*.

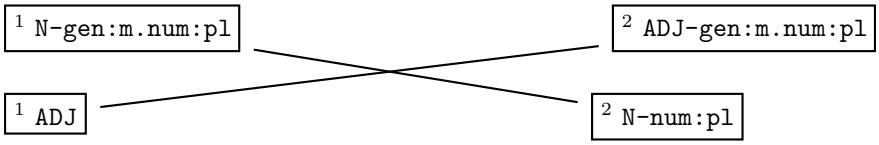
[6] Translated into English as *great man*.

z_1:



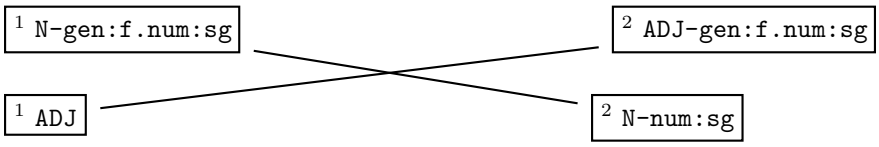
$r_1 = \{\}, r_2 = \{\}$

z_2:



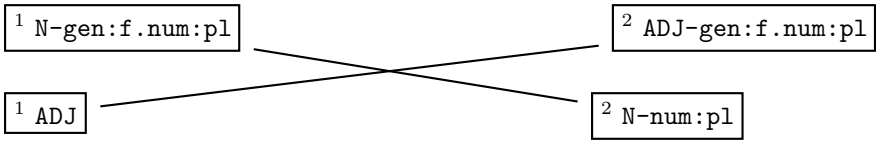
$r_1 = \{\}, r_2 = \{\}$

z_3:



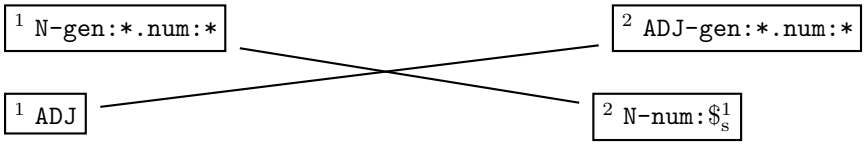
$r_1 = \{\}, r_2 = \{\}$

z_4:



$r_1 = \{\}, r_2 = \{\}$

Figure 2: Set of EATs needed by Sánchez-Martínez and Forcada (2009) to codify the noun–adjective reordering when translating Spanish into English. z_1 will be used to translate *tren viejo* into *old train*; z_2 will be used to translate *trenes viejos* into *old trains*; z_3 will be used to translate *locomotora vieja* into *old locomotive*; z_3 will be used to translate *locomotoras viejas* into *old locomotives*.



$r_1 = \{\}, r_2 = \{\}$

Figure 3: EAT learnt by our approach in order to codify the noun–adjective reordering when translating Spanish into English.

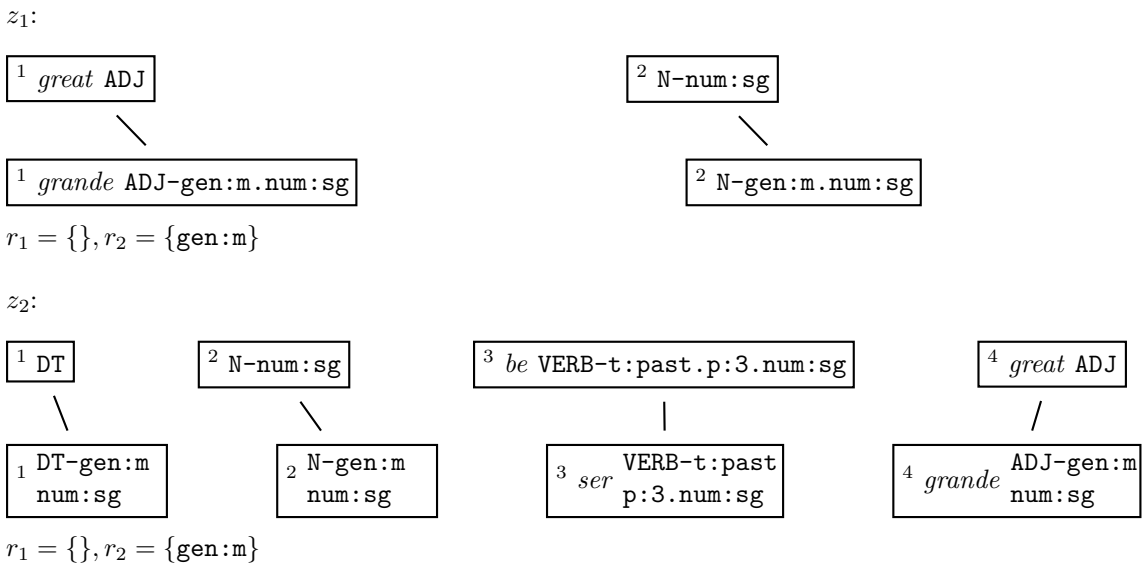


Figure 4: EATs learnt by Sánchez-Martínez and Forcada (2009) to translate the English adjective–noun construction into Spanish (z_1) when the adjective is *great*, and to translate this same adjective when it is preceded by a determiner, followed by a singular noun, and the verb to *be* in the past tense, 3rd person, singular (z_2). Note that this requires the adjective *great* to be added to the set of lexicalised units which do not have to be generalised.

after the noun as usual. In order to properly translate the English adjective–noun construction into Spanish when the Spanish equivalent of the English adjective is a prepositive adjective, prepositive adjectives need to be lexicalised. In the approach by Sánchez-Martínez and Forcada (2009) this would require knowing the set of the most frequent prepositive adjectives in Spanish in advance, adding them to the set of lexicalised units, and learning, in addition to the EATs in Figure 2, EATs like those shown in Figure 4 for the adjective *great*.[7] Note that z_1 from Figure 4 is an exception to the general rule used to translate the adjective–noun constructions because it does not perform any reordering, as opposed to the EATs in Figure 2, and that the translation rule encoded in z_2 from Figure 4 is equivalent to the general rule used to translate a determiner, followed by a singular noun, the verb to *be* in the past tense, 3rd person, singular, and a (predicative) adjective. It is therefore clear that the lexicalisation in z_2 is not needed and performing such a lexicalisation leads Sánchez-Martínez and Forcada (2009) to generate more EATs than are really necessary, some of which may be useless.

Our approach overcomes this limitation because the different generalisation levels explored for each word class include EATs in which the lemma of the lexical forms is kept unchanged. We then follow the approach outlined above to select the minimum number of EATs that are needed to reproduce the bilingual phrase pairs from which EATs are obtained. Consequently, these lexicalisations are only used when they are needed to encode an exception to a more general translation rule.

Third limitation: rules preventing the application of more convenient rules. Finally, Sánchez-Martínez and Forcada (2009) do not apply any method with which to discard those EATs that force SL lexical forms that should be processed together by the same rule —because they are involved in the same linguistic phenomenon— to be dealt with by different rules. This is a common situation when the bilingual phrase pairs, from which the EATs are obtained, are extracted by following the standard method in SMT (Koehn, 2010, Sec. 5.2.3), which is likely to separate words that should be processed together into different phrases. This is a problem in shallow-transfer RBMT because an SL lexical form can only be translated by a single

[7] We assume that if several EATs match the same sequence of SL lexical categories, the most specific EAT is applied. We also assume that the dictionaries of the RBMT system do not contain any information indicating whether or not an adjective is prepositive.

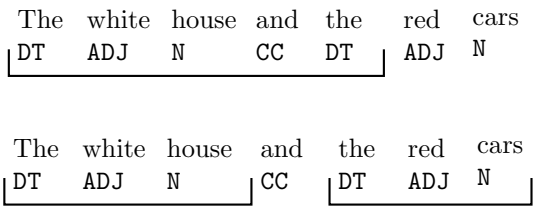


Figure 5: Example of the application of shallow-transfer rules in Apertium. The rule that matches a determiner–adjective–noun–conjunction–determiner construction is applied at the top, and, as a result, the last two words of the sentence are translated in isolation. The resulting translation into Spanish is *La casa blanca y el rojo coches*. Note that *el* is a masculine singular definite determiner that should be plural (i.e. *los*) in order to agree with the noun *coches*, and that the adjective *rojo* and the noun *coches* should get reordered and agree in gender and number. At the bottom, the last three words of the sentence are translated by a rule that matches a determiner–adjective–noun construction which performs the reordering and the gender and number agreement between the matched words and the resulting translation is *La casa blanca y los coches rojos*.

rule. In the specific case of Apertium, the SL sentence to be translated is divided into chunks so that each chunk is matched and translated by a single rule in a left-to-right, longest-match fashion. Apertium starts from the first SL lexical form in the sentence, selects the longest applicable rule, applies it to the matched chunk, prints the result, and starts the process again from the next (unmatched) SL lexical form in the sentence. If no rule can be matched, the corresponding SL lexical form is translated in isolation and the process starts again with the next one.

Not discarding the EATs that perform a segmentation of the SL sentence that is unsuitable for a shallow-transfer RBMT system may result in not applying other EATs that would perform a correct translation, despite having been learnt from the parallel corpus. This is illustrated in Figure 5 in which an EAT that matches a determiner, an adjective, a noun, a conjunction (`CC`) and another determiner is applied (top), rather than applying an EAT that matches a determiner, an adjective and a noun twice (bottom). In the first case, the translation of the determiner after the conjunction is not processed together with the noun and the adjective it has to agree with in gender and number, which may result in an incorrect translation into the TL.

We have tackled this last problem by retaining in the final set of EATs only those that make the translation of the SL side of the training parallel corpus sufficiently close to its TL side when the EATs to be used are selected in a left-to-right longest match manner, as the RBMT engine will do. This is done by following a greedy approach in order to identify the set of sequences of lexical categories that should be translated by the same rule, and by removing those EATs that produce the same translation as a set of shorter EATs would produce.

3. Generalised Alignment Templates

This section describes the notation that will be used in the remainder of the paper along with the improvement we have made to the EAT formalism used by Sánchez-Martínez and Forcada (2009) in order to be able to learn more general EATs which we shall refer to as *generalised alignment templates* (henceforth, GAT).

As stated in the introduction, shallow-transfer RBMT systems use as IR sequences of lexical forms in both languages. Recall that the translation process in shallow-transfer RBMT is as follows: first the SL IR is obtained from the SL text, usually with the help of a monolingual dictionary and a part-of-speech tagger; then the SL IR is converted into a TL IR by applying shallow-transfer rules (in our case encoded as GATs) and using a bilingual dictionary; finally the TL text is generated from the TL IR with the help of a TL monolingual dictionary.

A lexical form w, e.g. *car* `N-gen:`ϵ`.num:sg`, consists of:

- a lemma $\lambda(w)$, e.g. $\lambda(w) =$*car*,

- a lexical category[8] $\rho(w)$, e.g. $\rho(w)$ =N (noun),
- a set of morphological inflection attributes $\alpha(w)$, e.g. $\alpha(w) = \{\texttt{gen},\texttt{num}\}$ (gender and number), and
- their values $\upsilon(w, a)$, e.g. $\upsilon(w,\texttt{num})$ =sg (singular).

Some morphological inflection attributes may be assigned an empty value (ϵ) because they do not apply to that language; in the example above, $\upsilon(w,\texttt{gen}) = \epsilon$ because nouns do not have a gender in English. This is done for convenience so that lexical forms have the same morphological inflection attributes in the two languages involved in the translation. It is worth noting that the functions described above can equally be applied to lexical forms and words classes; what is more, $\alpha(\cdot)$ and $\upsilon(\cdot)$ can also be applied to restrictions.[9]

SL lexical forms are translated into TL lexical forms by looking them up in a bilingual dictionary. An SL lexical form may have more than one equivalent in the TL; in these cases, a lexical selection module (Tyers et al., 2012; Tyers, 2013) is responsible for selecting the most appropriate translation given the SL context prior to the execution of the structural transfer module. We shall refer to the result of translating an SL lexical form w into a TL lexical form as $\tau(w)$ throughout this document.

In order to be able to learn GATs we have introduced the following special values for the morphological inflection attributes:

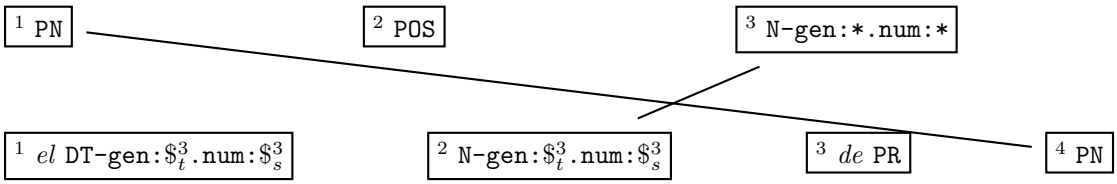
- The wildcard $*$ value in the morphological inflection attribute of an SL word class signifies that it matches any value. Hence, a GAT $z = (S, T, A, R)$,[10] with $S = (s_1, s_2, \cdots, s_n)$, and $R = (r_1, r_2, \cdots, r_n)$, *matches* a sequence of SL lexical forms $W = (w_1, w_2, \cdots, w_n)$ only if W and S have the same length and every SL lexical form $w_i \in W$ meets the following conditions:
  - either its lemma equals the lemma in the SL word class s_i, or s_i has no lemma (because it has been generalised):
    $$\lambda(w_i) = \lambda(s_i) \ \vee \ \lambda(s_i) = \epsilon;$$
  - its lexical category equals the lexical category in the SL word class s_i:
    $$\rho(w_i) = \rho(s_i);$$
  - either the value of the morphological inflection attributes of w_i equal those in s_i, or the value of the corresponding morphological inflection attributes in s_i contain wildcards:
    $$\forall a \in \alpha(s_i) : \upsilon(w_i, a) = \upsilon(s_i, a) \ \vee \ \upsilon(s_i, a) = *;$$
  - the value of the morphological inflection attributes specified in the restrictions r_i are equal to those in the TL lexical form obtained by looking up the SL lexical form w_i in the bilingual dictionary:[11]
    $$\forall a \in \alpha(r_i), \upsilon(\tau(w_i), a) = \upsilon(r_i, a).$$
- The SL reference $\$_s^j$ as the value of an attribute a of a TL word class t_i means that the TL lexical form w'_i produced as a translation takes the value of the corresponding attribute from the j-th SL lexical form matched by the GAT:
  $$\upsilon(w'_i, a) \leftarrow \upsilon(w_j, a).$$

[8] Without loss of generality, ρ could also be used to represent lexical subcategories.

[9] We abuse the notation slightly in this case; note, however, that a restriction merely consists of a set of restricted morphological inflection attributes and their values.

[10] The elements (S, T, A, R) have the same meaning as in EATs: S is a sequence of SL word classes, T is a sequence of TL word classes, A is a set of pairs of word class indexes (i, j) with the alignment information between the word classes in S and T, and R is a set of restrictions over the TL morphological inflection information of the lexical forms matching the GAT.

[11] The bilingual dictionary provides the translation of the lemma and also of the lexical category and morphological inflection attributes of the lexical form. For instance, the bilingual dictionary provides the gender that a noun must have when it is translated into Spanish.



$r_1 = \{\}, r_2 = \{\}, r_3 = \{\}$

Figure 6: GAT for the translation of the English Saxon genitive construction into Spanish. Compare with the EAT learnt by Sánchez-Martínez and Forcada (2009) (see Figure 1).

- The TL reference $\$_t^j$ as the value of an attribute a of a TL word class t_i means that a takes the value from the corresponding morphological inflection attribute in the TL lexical form obtained after translating the j-th SL lexical form by looking it up in the bilingual dictionary:

$$\upsilon(w_i', a) \leftarrow \upsilon(\tau(w_j), a).$$

It is worth noting that, even though the translation of most linguistic phenomena can be encoded using only TL references ($\$_t^j$), there are situations, such as that described below, in which SL references ($\$_s^j$) are needed. Consider the translation into English of the Spanish phrase *es guapa*,[12] with SL IR $w_1 =$ *ser* `VERB-t:pres.p:3.num:sg`, $w_2 =$ *guapa* `ADJ-gen:f.num:sg`. As the Spanish phrase contains no personal pronoun, the GAT that must be applied for its translation has to resort to the gender of the adjective in the SL to determine which pronoun, *she* or *he*, needs to be used, and an SL reference therefore needs to be used.

Apart from the changes explained above, GATs are applied to translation in the same way in which the EATs of Sánchez-Martínez and Forcada (2009) are applied. The following example illustrates how the new attribute values $\$_s^j$ and $\$_t^j$ are used during translation. The GAT shown in Figure 6 encodes the translation of the English Saxon genitive construction —proper noun + possessive ending + noun— into Spanish. The wildcard attribute in the number makes it match both singular and plural nouns; the SL and TL reference values propagate the gender and number of the noun to the determiner. When translating the SL (English) phrase *Mary's family*, with the SL IR $w_1 =$ *Mary* `PN`, $w_2 =$ *'s* `POS`, $w_3 =$ *family* `N-gen:`ϵ`.num:sg` with the GAT in Figure 6, the four TL lexical forms $w_1' \cdots w_4'$ produced as output are obtained as follows. The lemmas of the first and third lexical forms are taken from the GAT: $\lambda(w_1') = \lambda(t_1) =$ *el* and $\lambda(w_3') = \lambda(t_3) =$ *de*. The lemmas of the other two lexical forms are obtained by looking up the SL lexical forms aligned with them in the bilingual dictionary: $\lambda(w_2') = \lambda(\tau(w_3)) =$ *familia*, $\lambda(w_4') = \lambda(\tau(w_1)) =$ *Mary*. The lexical categories are taken from the TL word classes: $\rho(w_1') = \rho(t_1) =$ `DT`, $\rho(w_2') = \rho(t_2) =$ `N`, $\rho(w_3') = \rho(t_3) =$ `PR`, and $\rho(w_4') = \rho(t_4) =$ `PN`. The morphological inflection attributes `gen` (gender) of the first and second TL lexical forms take their values from the corresponding attribute in the translation of the third SL lexical form (TL reference): $\upsilon(w_1', \texttt{gen}) = \upsilon(w_2', \texttt{gen}) = \upsilon(\tau(w_3), \texttt{gen}) =$ `f`. The morphological inflection attributes `num` (number) of these same TL lexical forms take their value from the corresponding attribute in the third SL lexical form (SL reference): $\upsilon(w_1', \texttt{num}) = \upsilon(w_2', \texttt{num}) = \upsilon(w_3, \texttt{num}) =$ `sg`. The resulting sequence of TL (Spanish) lexical forms is $w_1' =$ *el* `DT-gen:f.num:sg`, $w_2' =$ *familia* `N-gen:f.num:sg`, $w_3' =$ *de* `PR`, $w_4' =$ *Mary* `PN`, which after morphological generation leads to *la familia de Mary*.

4. Inference of Shallow-Transfer Rules

The complete process used to obtain shallow-transfer rules from a parallel corpus consists of the steps described in the remainder of this section and summarised in Figure 7. First, word alignments and bilingual phrase pairs are obtained from the parallel corpus (Section 4.1). Multiple GATs, each one with a different

[12] Translated into English as *She is beautiful*.

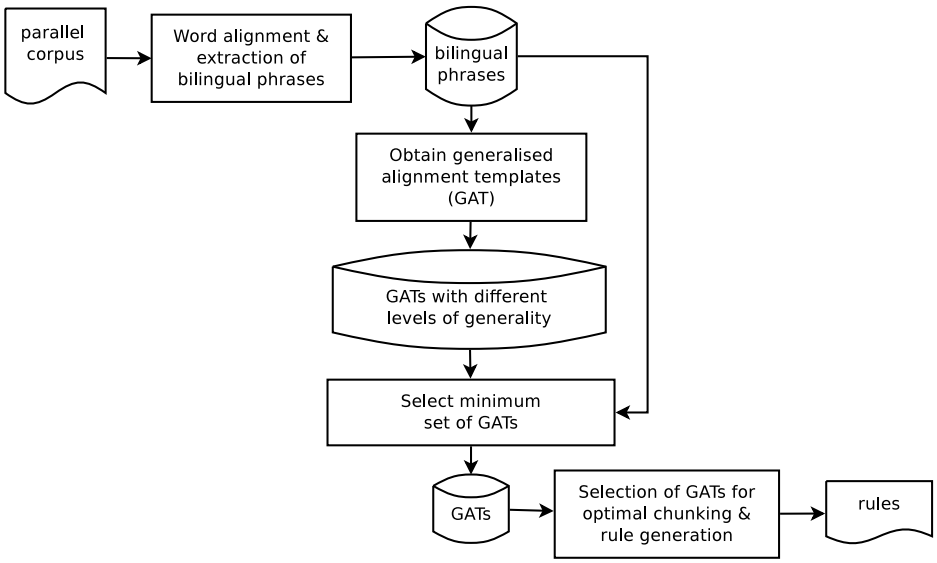


Figure 7: Steps followed to obtain a set of generalised alignment templates (GAT) from a parallel corpus.

level of generalisation, are then inferred from each of the bilingual phrase pairs obtained. This is done by using different sets of wildcard and reference attributes, and also with different lexicalised words (Section 4.2); these GATs, encoded with the formalism described in Section 3 do not suffer from the partial generalisation issue described in Section 2.1. After filtering certain GATs to deal with the noise present in the corpus and to prevent overgeneralisations (Section 4.3), the GATs with the most appropriate lexicalised words, generalisation level, and wildcard and reference attributes are automatically selected by finding the minimum set of GATs needed to correctly reproduce all the bilingual phrase pairs obtained from the corpus (Section 4.4). With this minimisation process, conflicts between GATs are removed and GATs with lexicalised word classes are selected only when they are strictly necessary in the context in which they appear; the second limitation described in Section 2.1 is therefore overcome. Any GATs that cause deficient chunking of the input are then discarded (Section 4.5) in order to get over the third limitation described in Section 2.1. Finally, the GATs selected are converted into the Apertium rule format, although they could be converted into the format used by any other shallow-transfer RBMT system.

4.1. Obtaining Word Alignments and Bilingual Phrase Pairs

Word alignments and bilingual phrase pairs are obtained using the state-of-the-art method in order to obtain bilingual phrases pairs for their use in SMT (Koehn, 2010). This method, which was also followed by Sánchez-Martínez and Forcada (2009), consists of the following steps:

1. Morphologically analyse both sides of the parallel corpus and solve the part-of-speech ambiguities in order to obtain sequences of lexical forms in both languages.
2. Train IBM models 1, 3 and 4 (Brown et al., 1993), and the HMM alignment model (Vogel et al., 1996), for 5 iterations by means of GIZA++ for both translations directions (Och and Ney, 2003).[13]
3. Compute the Viterbi alignment according to these models for both translation directions.
4. Symmetrise the two sets of Viterbi alignments using the refined intersection method proposed by Och and Ney (2003) to obtain word-aligned sentence pairs.
5. Extract bilingual phrase pairs that are consistent with the alignments (Koehn, 2010, Sec. 5.2.3).

[13] `http://code.google.com/p/giza-pp/`

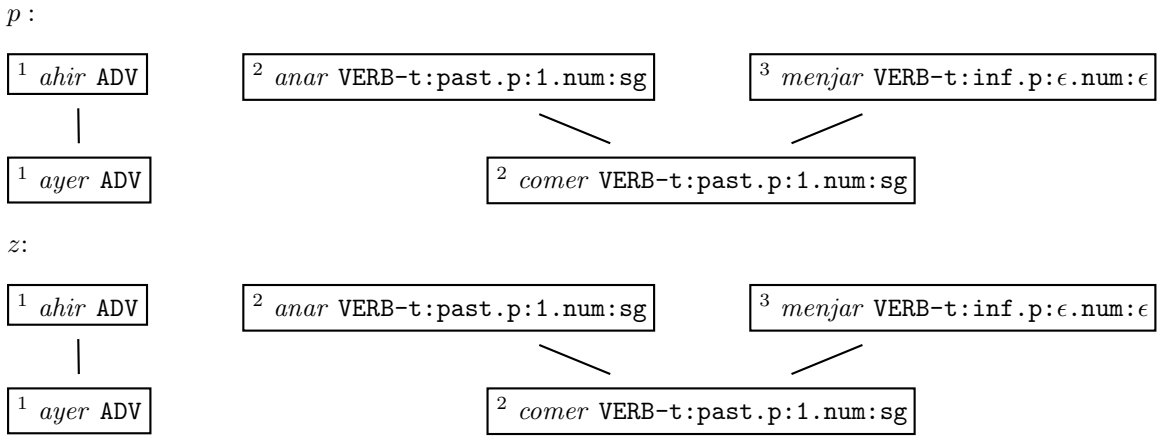


Figure 8: Catalan–Spanish bilingual phrase pair p and initial GAT z obtained from it with function β (see Section 4.2.1).

4.2. Extracting Generalised Alignment Templates from Bilingual Phrase Pairs

From each bilingual phrase p, many different GATs that correctly *reproduce* it —when applied to the SL phrase in p, the corresponding TL phrase is obtained— are generated, although not all of them will eventually be used for rule generation. The selection of GATs to be used for rule generation is described in sections 4.4 and 4.5.

Given a bilingual phrase pair p, the generation of GATs from it can be described as the initial generation of the most specific GAT, $\beta(p)$ (Section 4.2.1), and the chained application of 3 different functions (σ_1, σ_2 and σ_3), each one of which takes the set of GATs produced by the previous one as input and generates a new set of GATs from each GAT in the input set. Function σ_1 removes lexicalised words (Section 4.2.2), function σ_2 introduces wildcards, SL and TL references and removes restrictions (Section 4.2.3), and function σ_3 ensures that each non-lexicalised TL word class is aligned with at most one SL word class (Section 4.2.4).

4.2.1. Obtaining the Initial Generalised Alignment Template (β)

The initial GAT $z = \beta(p) = (S, T, A, R)$ created from a bilingual phrase pair p is the most specific GAT that can be obtained from it, and therefore only matches the SL phrase in p.

Let $p = (W, W', A')$ be a bilingual phrase pair with an SL sequence of lexical forms $W = (w_1, w_2, \cdots, w_n)$, a TL sequence of lexical forms $W' = (w'_1, w'_2, \cdots, w'_m)$ and alignment information $A' = \{(i, j) : i \in [1, n] \ \wedge \ j \in [1, m]\}$. Each SL word class $s_i \in S$ in GAT z has the same lemma, lexical category and morphological inflection attribute values as the corresponding SL lexical form w_i in p, i.e. $\forall i \in [1, n], s_i \leftarrow w_i$. The same applies to the TL word classes: $\forall i \in [1, m], t_i \leftarrow w'_i$. The alignment information A in z is also copied from the bilingual phrase pair: $A \leftarrow A'$. Finally, restrictions R are obtained by looking up each SL lexical form in the bilingual phrase pair in the bilingual dictionary as follows:

$$\forall w_i \in W, \alpha(r_i) \leftarrow \alpha(w_i), \text{ and}$$

$$\forall r_i, \forall a \in \alpha(r_i), \upsilon(r_i, a) \leftarrow \upsilon(\tau(w_i), a).$$

Figure 8 shows a bilingual phrase pair p and the initial GAT z obtained from it. Note that the restrictions limit the morphological attribute values to those in the bilingual dictionary.

4.2.2. Removing Lemmas (σ_1)

The next step as regards obtaining more general GATs is to remove from each initial GAT the lemma from some of the SL and TL lexical forms that are related according to the bilingual dictionary. Recall that, during translation, when a TL word class does not contain a lemma, the lemma of the TL lexical form produced is obtained by looking up the SL lexical form to which it is aligned in the bilingual dictionary.

Function σ_1 generates a new GAT for each of the possible subsets of the set E with the positions of the SL word classes from which the lemma can be removed. Given an input GAT $z = (S, T, A, R)$, with $S = (s_1, s_2, \cdots, s_n)$ and $T = (t_1, t_2, \cdots, t_m)$, the set E is obtained by first computing for each SL word class s_i the set D_i of the TL word classes aligned to it whose lemmas are related according to the bilingual dictionary:

$$D_i = \{t_j : (i, j) \in A \ \wedge \ \lambda(\tau(s_i)) = \lambda(t_j)\};$$

and then including in E the positions of the SL word classes whose lemmas, according to the bilingual dictionary, are related to at least one TL word class:

$$E = \{i : D_i \neq \emptyset\}.$$

For each possible subset $F \in \mathcal{P}(E)$,[14] σ_1 generates a new GAT z'. Each new GAT $z' = (S', T', A, R)$ is a copy of z in which the lemmas have been removed from the SL word classes whose positions are specified in F, and from the TL word classes aligned with them whose lemmas are related according to the bilingual dictionary:

$$\forall i \in F, \lambda(s'_i) \leftarrow \epsilon \text{ and}$$

$$\forall i \in F, \forall j : (i, j) \in A \ \wedge \ \lambda(\tau(s_i)) = \lambda(t_j), \lambda(t'_j) \leftarrow \epsilon.$$

As the empty set $\emptyset$ is always contained in $\mathcal{P}(E)$, the initial (non-generalised) GAT is always contained in the output of σ_1 (identity transformation).

Figure 9 shows the result of applying σ_1 to the GAT shown in Figure 8 (z_0). In this example, the number of GATs to be generated is 4 and $E = \{1, 3\}$ because, according to the bilingual dictionary, the first SL lemma is translated as the first TL lemma, and the third SL lemma is translated as the second TL lemma.

4.2.3. Introducing Wildcards and References in the Morphological Inflection Attributes (σ_2)

The use of wildcards and SL and TL references in the morphological inflection attributes allows the translation rules to be generalised to words with different values in their morphological attributes. This allows, for example, general reordering rules, like that presented in Figure 3, to be learnt, which are usually independent of the gender and number of the words involved.

Function σ_2 generates a set of GATs for each input GAT z by introducing wildcards in some of the morphological inflection attributes of the SL word classes and references in the counterpart morphological attributes in the TL word classes. It also removes the restrictions associated with the attributes of the SL word classes whose values have been replaced with a wildcard.

For each input GAT $z = (S, T, A, R)$, it is first necessary to obtain the set of candidate attributes C which are allowed to contain wildcards in the SL and references in the TL, and then the sets $M_{j,a}$ of possible SL references and TL references for each TL word class t_j and morphological inflection attribute $a \in C$.

A morphological attribute a is present in C only if for each TL word class $t_j \in T$ with $a \in \alpha(t_j)$ it contains an empty value ($\upsilon(t_j, a) = \epsilon$) or the non-empty value it contains can be obtained with an SL reference ($\exists i : \upsilon(s_i, a) = \upsilon(t_j, a)$) or with a TL reference ($\exists i : \upsilon(r_i, a) = \upsilon(t_j, a)$):

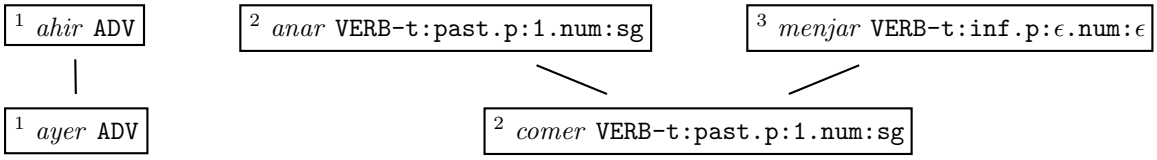
$$C = \{a : \upsilon(t_j, a) = \epsilon \ \vee \ (\exists i : \upsilon(s_i, a) = \upsilon(t_j, a)) \ \vee \ (\exists i : \upsilon(r_i, a) = \upsilon(t_j, a)) \ \forall t_j \in T : a \in \alpha(t_j)\};$$

Note that the restrictions are used to check whether an attribute value can be obtained with a TL reference, since their values have been obtained from the bilingual dictionary.

The sets $M_{j,a}$ of possible SL references and TL references for each TL word class t_j and morphological inflection attribute $a \in \alpha(t_j) \cap C$ are computed using Algorithm 1. This algorithm proceeds as follows. If attribute a can be obtained with a reference to an SL word class to which t_j is aligned, the corresponding reference is added to $M_{j,a}$. If not, the algorithm adds references to other SL word classes from which attribute a can be obtained to $M_{j,a}$. In either case, the SL references are only included in $M_{j,a}$ if TL

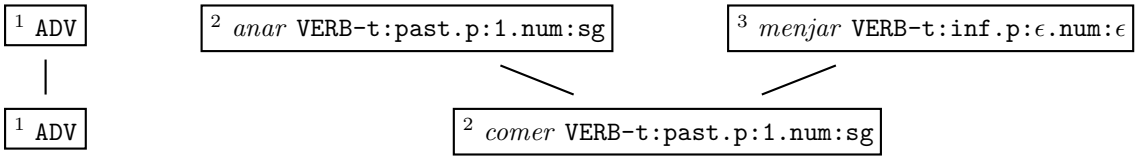
[14] $\mathcal{P}(E)$ is the power set of E.

$z_0 = z_1$:



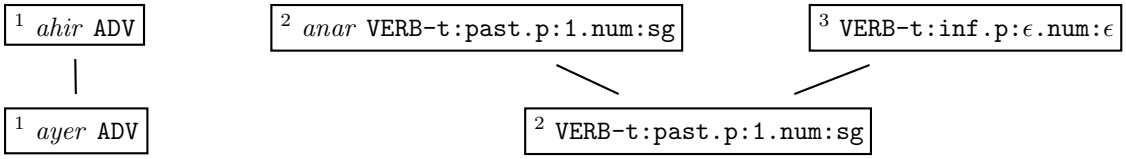
$r_1 = \{\}, r_2 = \{\texttt{t} : \texttt{past}, \texttt{p} : 1, \texttt{num} : \texttt{sg}\}, r_3 = \{\texttt{t} : \texttt{inf}, \texttt{p} : \epsilon, \texttt{num} : \epsilon\}$; $F = \{\}$

z_2:



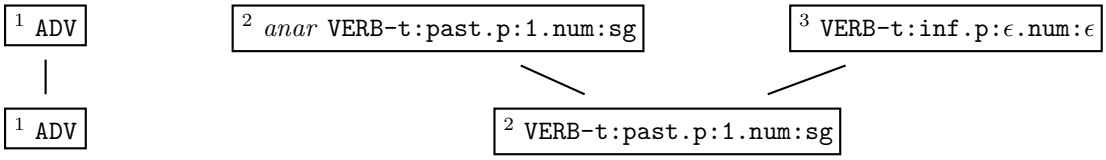
$r_1 = \{\}, r_2 = \{\texttt{t} : \texttt{past}, \texttt{p} : 1, \texttt{num} : \texttt{sg}\}, r_3 = \{\texttt{t} : \texttt{inf}, \texttt{p} : \epsilon, \texttt{num} : \epsilon\}$; $F = \{1\}$

z_3:



$r_1 = \{\}, r_2 = \{\texttt{t} : \texttt{past}, \texttt{p} : 1, \texttt{num} : \texttt{sg}\}, r_3 = \{\texttt{t} : \texttt{inf}, \texttt{p} : \epsilon, \texttt{num} : \epsilon\}$; $F = \{3\}$

z_4:



$r_1 = \{\}, r_2 = \{\texttt{t} : \texttt{past}, \texttt{p} : 1, \texttt{num} : \texttt{sg}\}, r_3 = \{\texttt{t} : \texttt{inf}, \texttt{p} : \epsilon, \texttt{num} : \epsilon\}$; $F = \{1, 3\}$

Figure 9: Set of GATs generated by σ_1 from the GAT in Figure 8 (z_0). For each GAT, the set $F \in \mathcal{P}(E)$ used to remove the lemmas is provided; $E = \{1, 3\}$ (see Section 4.2.2). Note that, according to the bilingual dictionary, the translation into Spanish of a lexical form whose lemma is *menjar* is a lexical form whose lemma is *comer*, while the translation of a lexical form whose lemma is *anar* is a lexical form whose lemma is *ir*, which is not part of any TL word class in z_0.

Algorithm 1 Algorithm that computes the set of possible SL and TL reference values that a given morphological inflection attribute a of a TL word class t_j can have.

$M_{j,a,1} \leftarrow \{\$_t^i : (i,j) \in A \ \wedge \ \upsilon(r_i,a) = \upsilon(t_j,a)\}$
$M_{j,a} \leftarrow M_{j,a,1} \cup \{\$_s^i : (i,j) \in A \ \wedge \ \upsilon(s_i,a) = \upsilon(t_j,a) \ \wedge \ \$_t^i \notin M_{j,a,1}\}$
if $M_{j,a} = \emptyset$ **then**
  $M_{j,a,2} \leftarrow \{\$_t^i : \upsilon(r_i,a) = \upsilon(t_j,a)\}$
  $M_{j,a} \leftarrow M_{j,a,2} \cup \{\$_s^i : \upsilon(s_i,a) = \upsilon(t_j,a) \ \wedge \ \$_t^i \notin M_{j,a,2}\}$
end if
return $M_{j,a}$

references cannot be used. $M_{j,a,1}$ represents the TL reference attributes to SL word classes to which t_j is aligned that give the value of the attribute a in t_j as a result, while $M_{j,a,2}$ contains the TL reference attributes to SL word classes to which t_j is not aligned and give the value of that attribute a as a result.

Finally, a set of GATs G_L is then obtained for each possible set of attributes $L \in \mathcal{P}(C)$, thus permitting GATs with different generalisation levels to be built: the more attributes in L, the more general the resulting GATs. As occurs with σ_1, the empty set $\emptyset$ is always contained in L, and every input GAT is therefore also part of the result of applying σ_2 to it.

All the GATs in G_L share the same sequence of SL word classes S', set of restrictions R' and alignment information A'; they only differ in the sequence of TL word classes. S' is a copy of the original sequence of SL word classes S in which the value of the morphological inflection attributes in L has been replaced with a wildcard:

$$\forall s_i \in S, \forall a \in \alpha(s_i) : a \in L, \upsilon(s'_i, a) \leftarrow *;$$

R' is a copy of the original sets of restrictions R in which the attributes in S whose values have been replaced with a wildcard in S' have been removed:

$$\forall s_i \in S, \forall a \in \alpha(s_i) : a \in L, r_i \leftarrow r_i - \{a\};$$

and A' is a copy of the original alignment information A.

The different sequences of TL word classes to be generated, one for each GAT in G_L, differ as regards the attribute values that need to be used. These values are obtained as the Cartesian product $N = \prod_{t_j \in T} \prod_{a \in \alpha(t_j)} \omega(t_j, a)$, where $\omega(t_j, a)$ equals a set with the original attribute value if attribute a will not be assigned a reference, or otherwise a set with the references to be used:

$$\omega(t_j, a) = \begin{cases} M_{j,a} & \text{if } a \in L \wedge |M_{j,a}| > 0 \\ \{\upsilon(t_j, a)\} & \text{otherwise.} \end{cases}$$

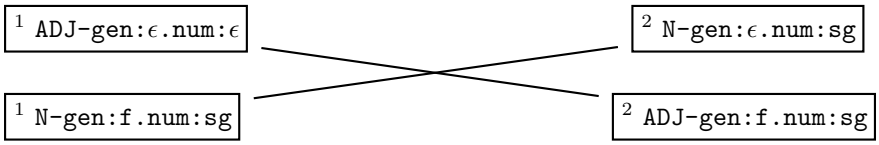
Finally, a GAT is created for each element $n \in N$. The sequence of TL word classes T' of each new GAT is a copy of the original sequence of TL word classes T in which the values of the attributes have been replaced with those in n.

Figure 10 shows the four GATs (z_1–z_4) generated by σ_2 for the input GAT z_0 from the same figure. These GATs codify the reordering and gender and number agreement rule that must be applied for the English–Spanish translation of an adjective followed by a noun. The set of morphological inflection attributes that can be assigned a wildcard in the SL, and a reference in the TL is $C = \{\texttt{gen}, \texttt{num}\}$; wildcards are permitted in the `num` (number) attribute because its value can be obtained by using an SL reference or a TL reference (in this case using both types of references) for both TL word classes; wildcards are permitted in the `gen` (gender) attributes because its value can be obtained using a TL reference. The sets of possible reference values to be used are $M_{1,\texttt{gen}} = \{\$_t^2\}$, $M_{1,\texttt{num}} = \{\$_t^2\}$, $M_{2,\texttt{gen}} = \{\$_t^2\}$ and $M_{2,\texttt{num}} = \{\$_t^2\}$, $\mathcal{P}(C) = \{\{\}, \{\texttt{gen}\}, \{\texttt{num}\}, \{\texttt{gen}, \texttt{num}\}\}$.

4.2.4. Removing Alignments (σ_3)

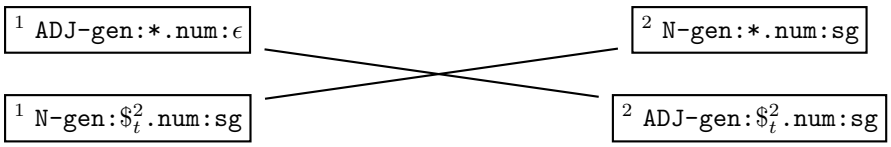
For a GAT to be useful in shallow-transfer RBMT, every non-lexicalised TL word class must be aligned with at most one SL word class: that from which, at translation time, the TL lemma will be obtained by

$z_0 = z_1$:



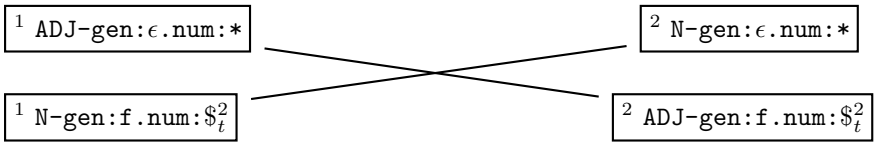
$r_1 = \{\texttt{gen} : \epsilon, \texttt{num} : \epsilon\}, r_2 = \{\texttt{gen} : \texttt{f}, \texttt{num} : \texttt{sg}\}$; $L = \{\}$

z_2:



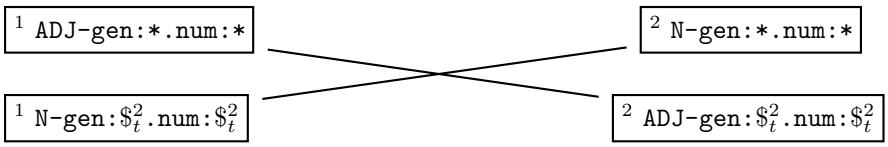
$r_1 = \{\texttt{num} : \epsilon\}, r_2 = \{\texttt{num} : \texttt{sg}\}$; $L = \{\texttt{gen}\}$

z_3:



$r_1 = \{\texttt{gen} : \epsilon\}, r_2 = \{\texttt{gen} : \texttt{f}\}$; $L = \{\texttt{num}\}$

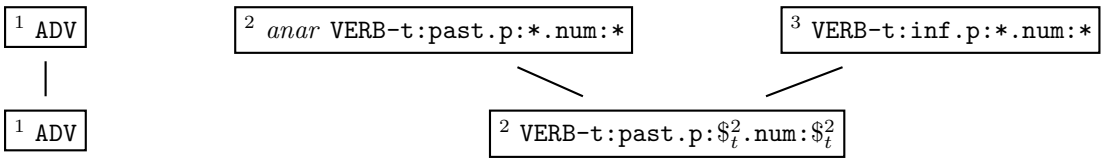
z_4:



$r_1 = \{\}, r_2 = \{\}$; $L = \{\text{gen}, \text{num}\}$

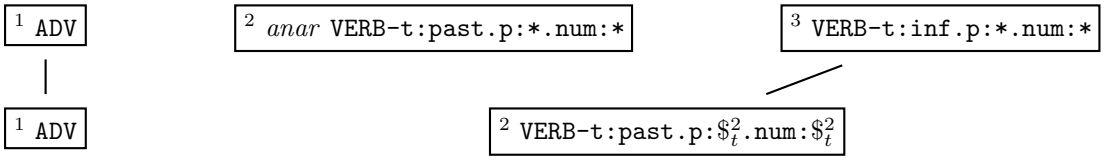
Figure 10: GAT codifying the reordering and gender and number agreement rule when translating a singular noun preceded by an adjective from English to Spanish (z_0). The noun is feminine in Spanish. The set of GATs (z_1–z_4) resulting from the application of σ_2 to z_0 are shown. For each GAT, the set L used to introduce wildcards in the SL and references in the TL are provided; $C = \{\texttt{gen}, \texttt{num}\}$, $M_{1,\texttt{gen}} = \{\$_t^2\}$, $M_{1,\texttt{num}} = \{\$_t^2\}$, $M_{2,\texttt{gen}} = \{\$_t^2\}$, $M_{2,\texttt{num}} = \{\$_t^2\}$ (see Section 4.2.3). The minimisation process described in Section 4.4 will be responsible for removing the redundancy present in this set of rules.

z_0:



$r_1 = \{\}$, $r_2 = \{\texttt{t} : \texttt{past}\}$, $r_3 = \{\texttt{t} : \texttt{inf}\}$

z_1:



$r_1 = \{\}$, $r_2 = \{\texttt{t} : \texttt{past}\}$, $r_3 = \{\texttt{t} : \texttt{inf}\}$

Figure 11: One of the GATs obtained from the bilingual phrase pair p in Figure 8 (z_0) and the GAT obtained from it (z_1) by function σ_3 (see Section 4.2.4).

looking up the SL lexical form matched in the bilingual dictionary.

Function σ_3 removes those alignments that would render z not applicable in shallow-transfer RBMT from each input GAT $z = (S, T, A, R)$. This is done by first obtaining the set with the positions of the non-lexicalised TL word classes V:

$$V = \{j : t_j \in T \ \wedge \ \lambda(t_j) = \epsilon\}.$$

Then, for each TL word-class position $j \in V$, the set of possible alignments X_j is computed by considering the bilingual phrase pair (W, W') from which the GAT z was obtained and ensuring that it can be reproduced using the selected alignment points:

$$X_j = \{(i, j) : (i, j) \in A \ \wedge \ \lambda(\tau(w_i)) = \lambda(w'_j)\}.$$

Finally, all the possible subsets of A that ensure that the original bilingual phrase pair can be reproduced from z are calculated as the Cartesian product $Y = \prod_{j \in V} X_j$, and σ_3 generates an alternative GAT $z_y = (S, T, A_y, R)$ for each element $y \in Y$, where A_y stands for the subset of A that contains exactly the elements from the tuple y.

Figure 11 shows an input GAT z_0 and the GAT z_1 produced from it by σ_3. The valid alignment points for each TL word class are $X_1 = \{(1, 1)\}$ and $X_2 = \{(3, 2)\}$; the Cartesian product $Y = \{((1, 1), (3, 2))\}$ consists of a single element, which generates GAT z_1.

4.3. Filtering Unreliable Generalised Alignment Templates

Once a set of GATs has been generated from each bilingual phrase pair, a filtering of the GATs obtained must be carried out in order to discard those that are very infrequent or are not able to reproduce a large proportion of the bilingual phrase pairs they match. This may occur as a result of either the noise present in the training parallel corpus or overgeneralisations.

Given the set of bilingual phrase pairs P extracted from the parallel corpus (see Section 4.1) and a GAT $z \in Z$, the set of GATs obtained from P (see Section 4.2), we refer to $\mathcal{M}(z) \subseteq P$ as the set of bilingual phrase pairs that are matched by z.[15] Some of these bilingual phrase pairs, $\mathcal{G}(z) \subseteq \mathcal{M}(z)$, are correctly translated

[15] A GAT matches a bilingual phrase pair if the SL word classes match the sequence of SL lexical forms and all restrictions are met (see Section 3 for more details).

by z —when applied to their SL side, their TL side is obtained— while others, $\mathcal{B}(z) = \mathcal{M}(z) - \mathcal{G}(z)$, are not.

The filtering consists of discarding, on the one hand, those GATs z whose number of correctly reproduced bilingual phrase pairs $\mathcal{G}(z)$ is below a threshold θ; and on the other, those GATs for which the ratio of bilingual phrase pairs correctly reproduced and matched to the total number of bilingual phrase pairs matched is below a threshold δ. Any GATs that encode very infrequent linguistic transformations, along with those that overgeneralise, are thus avoided. The number of correctly reproduced bilingual phrase pairs is calculated by considering the frequency in the training parallel corpus of each bilingual phrase pair. A GAT z is thus discarded if

$$\mathcal{Q}(\mathcal{G}(z)) < \theta \ \vee \ \frac{\mathcal{Q}(\mathcal{G}(z))}{\mathcal{Q}(\mathcal{M}(z))} < \delta,$$

where $\mathcal{Q}$ is the aggregated frequency of a set of bilingual phrase pairs:

$$\mathcal{Q}(P) = \sum_{p \in P} \text{freq}(p),$$

and $\text{freq}(p)$ is the absolute frequency in the parallel corpus of the bilingual phrase pair p.

4.4. Choosing the Most Appropriate Generalised Alignment Templates

The objective of our approach is to obtain a set of GATs that is able to correctly translate at least the set of bilingual phrase pairs extracted from the training parallel corpus. What is more, the GATs in that set must be as general as possible in order to extend the linguistic knowledge obtained from the corpus to unseen input texts. This objective is achieved by selecting the minimum amount of GATs needed to correctly reproduce all the bilingual phrase pairs. Since the more general the GATs, the higher the amount of bilingual phrase pairs they match and (hopefully) reproduce, if the amount of GATs is minimised, the most general ones that are able to reproduce the bilingual phrase pairs in the training corpus are selected.

Unlike the other approaches used to automatically learn shallow-transfer rules from parallel corpora (Sánchez-Martínez and Forcada, 2009; Caseli et al., 2006; Probst et al., 2002), here all the bilingual phrase pairs are considered together when checking their reproducibility by the set of GATs obtained. We thus treat conflicting rules at a global level, while previous approaches treat them locally.

To define the minimisation problem, GATs need to be ordered according to their level of specificity. A GAT $z = (S, T, A, R)$ is said to be more specific than another GAT $z' = (S', T', A', R')$ if it has any component —either a lemma, a morphological inflection attribute or a restriction— that takes into account more fine-grained information than z':[16]

$$\begin{aligned}\text{more\_specific}(z, z') \iff \quad & |S| = |S'| \ \wedge \ \forall s_i \in S, \\ & (\rho(s_i) = \rho(s'_i) \ \wedge \\ & (\lambda(s_i) = \lambda(s'_i) \ \vee \ \lambda(s'_i) = \epsilon) \ \wedge \\ & \forall a \in \alpha(s_i), (\upsilon(s_i, a) = \upsilon(s'_i, a) \ \vee \ \upsilon(s'_i, a) = *) \ \wedge \\ & \forall a \in \alpha(r_i), (\upsilon(r_i, a) = \upsilon(r'_i, a) \ \vee \ a \notin r'_i)) \ \wedge \\ & (\exists s_i \in S : s_i \neq s'_i \vee \exists r_i \in R : r_i \neq r'_i).\end{aligned}$$

On the basis of the set of bilingual phrase pairs P, the set of GATs Z and their relation of specificity defined by the function more_specific$(\cdot)$, the minimum set of GAT $O \subseteq Z$ is chosen subject to the following constraints:

[16] Another option would be to compare the sets of bilingual phrase pairs matched by each GAT and consider z as more specific than z' if $\mathcal{M}(z) \subset \mathcal{M}(z')$. However, when the training corpus is small (only a few hundred sentences), it may occur that z and z' match the same set of bilingual phrase pairs in spite of z' being more general than z because it has the potential to match more sequences of lexical forms when translating new texts.

$\mathcal{C}_1$**:** Each bilingual phrase pair is correctly reproduced by at least one GAT that is part of the solution:

$$\bigcup_{z_i \in O} \mathcal{G}(z_i) = P$$

$\mathcal{C}_2$**:** If a GAT z_i that is part of the solution incorrectly reproduces the TL part of a bilingual phrase pair p, there is another GAT z_j that is part of the solution, is more specific than z_i and correctly reproduces the TL part of p:

$$\forall z_i \in O, \forall p \in \mathcal{B}(z_i), \exists z_j \in O : \text{more\_specific}(z_j, z_i) \ \wedge \ p \in \mathcal{G}(z_j)$$

In practice, constraint $\mathcal{C}_1$ needs to be relaxed because, as a result of the filtering method described above (see Section 4.3), there may not be a subset $O \subset Z$ satisfying it, i.e., the minimisation problem may not have a solution because it is impossible to reproduce all the bilingual phrase pairs regardless of the set of GATs chosen. This occurs when the highly lexicalised GATs that would be needed to reproduce certain bilingual phrase pairs have been removed and there is a conflict between the less specific GATs that are able to reproduce them. When this happens, we find the set of bilingual phrases $P_O \subset P$ that maximises $\sum_{p \in P_O} \text{freq}(p)$ and makes the minimisation problem solvable, i.e., that permits finding a set of GATs that meets the constraints $\mathcal{C}_1$ and $\mathcal{C}_2$.

There may also be multiple solutions to the minimisation problem, i.e., different sets of GATs with the same (minimum) size may satisfy the two constraints above. In this case, we choose the set of GATs containing the most general GATs. This is done by defining a function spec_level(z)[17] that accounts for the level of specificity of a GAT z (see below), computing the aggregated level of specificity of the possible solutions to the minimisation problem, $\sum_{z \in O} \text{spec\_level}(z)$, and choosing the set with the smaller aggregated level of specificity as the solution. The level of specificity of a GAT z is simply obtained by counting the number of lexicalised words and the number of morphological inflection attributes in the SL word classes with non-wildcard values:

$$\text{spec\_level}(z) = \quad \gamma_1 |\{s_i : s_i \in S \wedge \lambda(s_i) \neq \epsilon\}| + \gamma_2 \sum_{s_i \in S} |\{a : \upsilon(s_i, a) \neq *\}| + 1.$$

The first two terms in the equation above are assigned a weight so that lexicalised word classes have a higher impact on the final result than the morphological inflection attributes with non-wildcard values. This is achieved by making $\gamma_2 = 1$ and γ_1 higher than the highest possible value of the second term, that is, $\gamma_1 = \sum_{s_i \in S} |\alpha(s_i)| + 1$, since, in practice, a different minimisation subproblem is solved for each sequence of SL lexical categories (see below). The third term is added for convenience, to prevent spec_level(z) from returning a null value.

The minimisation problem we have defined is similar to the well-known *set covering* problem (Garey and Johnson, 1979), which is NP-hard (Korte and Vygen, 2012, Sec. 15.7). Despite its complexity, it can be solved in a reasonable amount of time when the quantity of bilingual phrase pairs and GATs is relatively small —a common situation when the amount of training parallel corpora is scarce— by splitting the problem into independent sub-problems: one for each different sequence of the SL lexical categories.

Each minimisation sub-problem is formulated as an integer linear programming problem (Garfinkel and Nemhauser, 1972). This kind of problems involves the optimisation of a linear objective function subject to linear inequality constraints. In our experiments we have used the state-of-the-art *branch and cut* approach (Xu et al., 2009). For a detailed description on how the minimisation problem is reformulated using linear inequations we refer the reader to AppendixA.

[17] Note the difference between more_specific(·) and spec_level(·). more_specific(·) defines a strict partial order in which two GATs are related if, and only if, the set of bilingual phrase pairs matched by one of them is a subset of the set of bilingual phrase pairs matched by the other. This makes the solution of the minimisation problem look like a hierarchy with general rules and more specific rules fixing the cases not correctly translated with the general ones. Contrarily, spec_level(·) simply allows our strategy to select from among different solutions with the same amount of GATs.

4.5. Optimising Rules for Chunking

The problem of selecting the minimum set of GATs that are needed to reproduce all the bilingual phrase pairs obtained from the training parallel corpus has been independently solved for each sequence of SL lexical categories. However, several GATs are used in the translation of a SL sentence, and each one translates a different sequence of SL lexical categories. The segmentation of the input SL sentences into chunks (sequences of SL lexical forms) is done by the GATs to be applied, which are chosen by the engine in a greedy, left-to-right, longest match fashion. It is therefore necessary to avoid the situation of having lexical forms that should be processed together —because they are involved in the same linguistic phenomenon— being assigned to different chunks.

This section describes the process carried out in order to select the subset of the set of GATs obtained after solving the minimisation problem that ensures that the text to be translated will be chunked in the most convenient way. We select the sequences of SL lexical categories that GATs must contain in order to be part of the final solution; to do this, we follow a greedy approach that attempts to maximise the similarity between the TL side of the training parallel corpus and the result of translating its SL side using GATs in the same way as the RBMT engine would do. The method first identifies the minimum set of SL text segments (*key segments*) in the training corpus which need to be translated by a rule to obtain the highest similarity. Afterwards, the sequences of SL categories that ensure that the maximum number of key segments get translated properly are selected.

Identifying key segments. Let $\mathcal{K}$ be the set containing all the possible sets of text segments in the SL sentences of the training corpus,[18] and $\mathcal{K}^\star \subseteq \mathcal{K}$ be the set of sets of text segments that maximise the *similarity* between the TL side of the training corpus and the translation obtained by translating each text segment in $K \in \mathcal{K}^\star$ with the most specific GAT available (as the RBMT engine would do) and the rest of the SL words in the training corpus word for word by looking them up in the bilingual dictionary. Here, similarity may be computed by using any standard MT evaluation measure:

The set of key text segments $\mathcal{I}$ is one of the sets in $\mathcal{K}^\star$. As $\mathcal{K}^\star$ may contain more than a single set, $\mathcal{I}$ is chosen so that it satisfies the following two conditions:

1. $\mathcal{I}$ is one of the sets with the fewest and shortest segments, i.e., with the minimum number of words covered by segments:

$$\mathcal{I} \in \arg\min_{K \in \mathcal{K}^\star} \sum_{k \in K} |k|,$$

   where $|x|$ denotes the number of words of text segment x.

2. $\mathcal{I}$ is one of the sets with the minimum average segment length:

$$\mathcal{I} \in \arg\min_{K \in \mathcal{K}^\star} \frac{\sum_{k \in K} |k|}{|K|}$$

These two conditions give priority to short text segments, and therefore to short GATs, over longer ones, in addition to the use of as few GATs as possible. If more than one set satisfies these two conditions, $\mathcal{I}$ is chosen at random from among them.

As exploring the whole set $\mathcal{K}$ is computationally unfeasible, in practice $\mathcal{I}$ is obtained by processing one parallel sentence at a time and following a dynamic programming approach similar to the *beam search* approach used for decoding in SMT (Koehn, 2004b).[19]

Note that when computing the set of key text segments $\mathcal{I}$, two text segments consisting of the same sequence of words are considered different if they appear in different positions in the corpus. This is also applicable to the description provided as follows.

[18] Note that the text segments in $K \in \mathcal{K}$ do not overlap and do not necessarily cover all the words in the corpus.

[19] Note that, despite the fact that the key text segments are computed independently for each sentence, it is highly unlikely that the addition of a new sentence substantially affects the solution because all the key segments are considered together when selecting the sequence of lexical categories for which rules will be generated.

Selecting the sequences of lexical categories. The sequences of lexical categories that GATs must contain in order to be part of the final solution are chosen from among the set $\mathcal{L}$ with the candidate sequences of lexical categories, which are in turn obtained from the words in the set of key text segments:

$$\mathcal{L} = \bigcup_{g \in \mathcal{I}} \{(\rho(w_i))_{i=1}^{|g|}\}.$$

For each sequence $l \in \mathcal{L}$, a score seq_qa(l) is computed. This score measures the impact on the translation quality of having rules matching the sequence of lexical categories l:

$$\text{seq\_qa}(l) = \frac{|\text{key\_seg\_ok}(l)|}{|\text{key\_seg\_ok}(l)| + |\text{key\_seg\_broken}(l)|},$$

where key_seg_ok(l) is the set of key text segments correctly translated by a rule matching the sequence of lexical categories l; and key_seg_broken(l) is the set of key text segments not correctly translated by a rule matching l plus the set of key text segments whose words are not translated together by the same rule as a consequence of having a rule matching l.

On the one hand, key_seg_ok(l) is defined as:

$$\begin{aligned}\text{key\_seg\_ok}(l) = \quad & \{g : g \in \mathcal{I} \ \wedge \ (((\rho(w_i))_{i=1}^{|g|} = l) \ \vee \\ & (\exists g' : g \in \text{seg}(g') \ \wedge \ (\rho(w_i))_{i=1}^{|g'|} = l \ \wedge \ \exists K \in \mathcal{K}^\star : g' \in K))\},\end{aligned}$$

where seg(x) is the set of all possible (sub)segments of text segment x. The text segments returned by key_seg_ok(l) are the key text segments ($g \in \mathcal{I}$) with a sequence of lexical categories l ($(\rho(w_i))_{i=1}^{|g|} = l$), and the key text segments contained in longer segments ($\exists g' : g \in \text{seg}(g')$) with a sequence of lexical categories l ($(\rho(w_i))_{i=1}^{|g'|} = l$) and correctly translated by a GAT ($\exists K \in \mathcal{K}^\star : g' \in K$).

On the other hand, key_seg_broken(l) is defined as:

$$\begin{aligned}\text{key\_seg\_broken}(l) = \quad & \{g : g \in \mathcal{I} \ \wedge \ ((\exists g' : g \in \text{seg}(g') \ \wedge \ (\rho(w_i))_{i=1}^{|g'|} = l \ \wedge \\ & \nexists K \in \mathcal{K}^\star : g' \in K \ \wedge \ \exists z \in O : \text{match}(z, g')) \ \vee \\ & (\exists g'' : (\rho(w_i))_{i=1}^{|g''|} = l \ \wedge \ \exists z \in O : \text{match}(z, g'') \ \wedge \\ & \text{start}(g'') < \text{start}(g) \ \wedge \ \text{end}(g'') < \text{end}(g) \ \wedge \ \text{end}(g'') \geq \text{start}(g)))\}\end{aligned}$$

where start(x) and end(y) refer to the position in the corpus of the first word of text segment x and the last word of text segment y, respectively; match(z, x) equals true if the GAT z matches the sequence of SL lexical forms of text segment x, otherwise zero. The text segments returned by key_seg_broken(l) are the key text segments ($g \in \mathcal{I}$) contained in longer segments ($\exists g' : g \in \text{seg}(g')$) with a sequence of lexical categories l ($(\rho(w_i))_{i=1}^{|g|} = l$), matched by at least one GAT ($\exists z \in O : \text{match}(z, g')$) and not correctly translated by any of the GATs matching it ($\nexists K \in \mathcal{K}^\star : g' \in K$). It also returns the key text segments which are intersected on the left by another text segment g'' with a sequence of lexical categories l ($\exists g'' : (\rho(w_i))_{i=1}^{|g''|} = l$) and matched by at least one GAT ($\exists z \in O : \text{match}(z, g'')$). Note that any text segment intersecting on the left with a key text segment g and matched by a GAT prevents the words in g from being translated together by the same GAT. This happens, for instance, in the example presented at the top of Figure 5 for the sentence *The white house and the red cars*: the GAT applied to the chunk *The white house and the* prevents the words in the chunk *the red cars* from being processed together by the GAT that would perform the gender and number agreement that is needed to produce a correct translation of that sentence into Spanish.

A subset of the set of GATs O obtained as a result of the minimisation step described in Section 4.4 is then selected as follows:

$$O_{\text{sel}} = \{z = (S, T, A, R) : z \in O \wedge (\rho(s))_{s \in S} \in \{l : l \in \mathcal{L} \wedge \text{seq\_qa}(l) \geq \mu\}\}$$

Where μ is a threshold whose value is automatically determined by trying all its possible values[20] and choosing that which maximises the *similarity* of the TL side of the training corpus and the translation obtained when its SL sentences are translated with the set of GAT O_{sel}. Note that not all GATs in O_{sel} will eventually be used to generate shallow-transfer rules, since some of them may be discarded as a result of the next step.

Removing Redundant Generalised Alignment Templates. The number of GATs can be further reduced without decreasing the translation performance by removing those GATs which produce the same translations that a set of shorter GATs would produce. Let us suppose that GAT z produces the translation W' when applied to the SL segment W. It often occurs that, when removing z from the set of GATs of the RBMT system, the engine still produces W' when translating W. This may occur because the RBMT system splits W into two or more chunks and the translation of these chunks by the matching GATs yields W', because the word for word translation of W produces W' as a result, or because of a combination of these two reasons. If this occurs for all the SL segments that match z, then z can be safely removed from the set of GATs from which rules will be generated because it is redundant, i.e., the information in z is already contained in other GATs. Removing these longer GATs has actually improved the translation performance. Since long GATs are learnt from fewer examples and the useless ones are removed, shorter, more reliable GATs are applied.

In order to detect and remove these redundant GATs, the following process is carried out. First, the GATs in O_{sel} are sorted in order of decreasing length, while GATs of the same length are sorted by increasing level of specificity.[21] For each GAT z, the bilingual phrase pairs correctly reproduced by it, $\mathcal{G}(z)$, are then collected, and each bilingual phrase pair $p \in \mathcal{G}(z)$ is checked in order to ascertain whether or not, when translating the SL side of p with the set of GAT $O_{\text{sel}} - \{z\}$, its TL side is obtained. If this requirement is met for all $p \in \mathcal{G}(z)$, z is definitively removed from O_{sel}, i.e., $O_{\text{sel}} \leftarrow O_{\text{sel}} - \{z\}$. It is therefore possible to guarantee that, after removing redundant GATs, the TL side of each bilingual phrase pair can be safely reproduced with the GATs that remain in O_{sel}.

For example, the Catalan–Spanish GAT z_1 in Figure 11 could be safely removed from the set of GATs O_{sel} if the GAT in Figure 12 is also part of O_{sel} because the presence of an adverb before the Catalan verb *anar* does not change the way in which the verb *anar* in the past tense followed by a verb in infinitive mood is translated. All the bilingual phrase pairs matching z_1 in Figure 12 can thus be reproduced by translating the adverb in isolation, i.e., by looking it up in the bilingual dictionary, and applying the GAT in Figure 12 to the other two lexical forms.[22]

4.6. Generation of Apertium Shallow-Transfer Rules

Finally, the GATs resulting from the application of all of the above steps are converted into the rule format of the Apertium RBMT engine so that they can be used in real-world translation tasks. A list containing all the GATs and compatible with the strict partial order defined by the function $more\_specific(\cdot)$ is built by means of a topological sorting algorithm (Kahn, 1962). This list contains the resulting GATs sorted in decreasing order of specificity and is used when generating the rules so that the most specific GAT is always applied when different GATs match the same input sequence of lexical forms. We refer the reader to AppendixB for the details of this conversion.

[20] Actually, all the possible vales of μ do not need to be tested since those that generate the same set O_{sel} will produce the same result.

[21] The sorting is based on the function more_specific, defined in Section 4.4. GATs of the same length are arranged in a list compatible with the strict partial order defined by the function $more\_specific(\cdot)$ by means of a topological sorting algorithm (Kahn, 1962).

[22] The proportion of GATs discarded because they can be replaced by shorter ones varies across language pairs and training corpus sizes. Generally, larger training corpora and more distant language pairs involve fewer GATs discarded. For instance, in the experiments described in Section 6, 75% of the Catalan–Spanish GATs with 5 SL lexical forms were discarded when the training corpus contained 250 sentences, while the proportion dropped to 41% for the training corpus with 5 000 sentences. When Spanish–English rules were inferred from the training corpus with 5 000 sentences, only 14% of the GATs with 5 SL lexical forms were discarded.

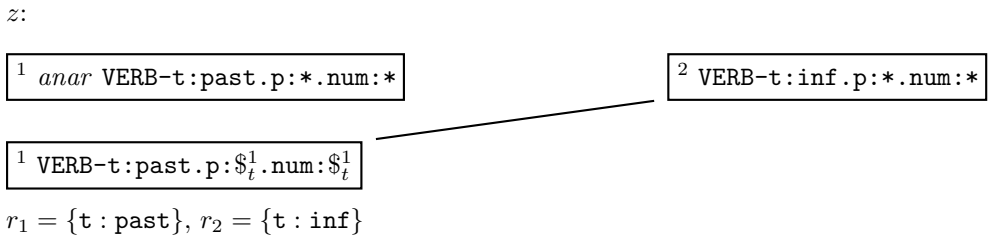


Figure 12: GAT encoding the translation from Catalan into Spanish of the verb *anar* in the past tense followed by a verb in infinitive mood.

5. Related Work

There have been other attempts to automatically learn structural transfer rules for RBMT. Probst (2005) developed a method with which to learn transfer rules from a reduced set of bilingual segments. These segments are obtained by asking a group of bilingual annotators to translate a controlled, parsed corpus containing examples of all the relevant grammatical structures in the SL. Word alignments are also provided by the bilingual annotators. The transfer rules learnt follow a hierarchical formalism similar to that used in the early METAL system (Hutchins and Somers, 1992).

The main differences between the approach by Probst (2005) and that presented in this paper are the following. First, their method learns hierarchical syntactic rules, whereas ours learns flat, shallow-transfer rules. Second, their method uses a corpus whose SL side needs to be parsed, whereas ours does not use information about the syntactic constituents; in addition to this, our approach learns how to automatically segment the text into chunks for their translation. Third, the TL side of the training corpus used by Probst (2005) is provided by bilingual annotators; in our approach, however, the TL side of the training corpus originates from a parallel corpus and there is no guarantee that the TL side will be obtained by directly translating the SL side. Fourth, the alignments between the words their approach uses are human-annotated, whereas ours obtains the alignments automatically through the use of statistical methods; our approach consequently has to tolerate alignment errors, especially when the training corpus is very small. Fifth, the strategy applied in order to generalise bilingual phrase pairs to rules is clearly different. Their initial approach (Probst et al., 2002) consists of selecting the minimum set of rules which correctly translates the set of bilingual phrase pairs by following a greedy strategy based on merging pairs of rules, while ours selects the minimum set of rules by using a global strategy based on integer linear programming that is able to find the optimal solution. In her latest approach (Probst, 2005), a two-step procedure is followed for the generalisation problem. Their system first learns the context-free backbone of the rules, that is, how the terminal symbols of the grammar (which represent lexical categories) are grouped together and with non-terminal symbols to generate other non-terminal symbols. Value and agreement constraints are then obtained. Rules initially contain only value constraints, i.e., they are only applied to words with the same morphological feature values as the examples from which the rules have been extracted. Agreement constraints, which replace value constraints and generalise the values of the morphological attributes in the learning examples, are then inferred by considering the frequency of the different values that each morphological inflection attribute happens to have in the examples used for learning. Font-Llitjós (2007) approached the automatic inference of the same kind of hierarchical rules from a completely different source of bilingual information: post-editings performed by users of the MT system. In addition, the rule inference/refinement is performed incrementally.

Caseli et al. (2006) present a method in which shallow-transfer rules and bilingual dictionaries are learnt from a parallel corpus. With regard to the shallow-transfer rule inference, these rules are learnt from a set of bilingual phrase pairs obtained by aligning the words in the source and target sentences by means of statistical methods in a way similar to that used by the approach described in this paper. After obtaining the bilingual phrase pairs, rules are inferred from them and those containing complementary information are joined in order to reduce their number. For instance, if rule a is applied to masculine nouns and rule b is applied to feminine nouns, a new rule c is created, which is applied to both masculine and feminine

nouns. In a final step, conflicts between rules are avoided in a greedy fashion by specialising the rules, either by including more morphological inflection attributes as a condition for their application, or by lexicalising some of the lexical forms they match. If a rule cannot be further specialised, the most frequent one is retained.

The approach by Caseli et al. (2006) principally differs from that presented in this paper as regards the way in which bilingual phrase pairs are generalised to obtain rules. On the one hand, their approach does not generalise unseen linguistic features, that is, if a rule is learnt from bilingual phrase pairs containing only masculine nouns, it will never be applied to feminine nouns. Our method, meanwhile, generates rules that generalise morphological inflection values not seen in the training set thanks to our more powerful formalism, provided that these rules are able to correctly reproduce the bilingual phrase pairs from which they are learnt. This is a great advantage when the size of the training corpus is very small. On the other hand, our minimisation approach considers all the possible alternatives when dealing with conflicts between rules matching the same sequence of lexical categories, rather than doing so in a greedy manner. With regard to the way in which the rules learnt affect the segmentation into chunks of the SL sentences to be translated, Caseli et al. (2006) do not confront the problem that long rules may prevent the application of shorter and more accurate ones; they merely select the sequences of SL and TL lexical categories for which rules will be generated based on their frequencies in the parallel corpus.

There have also been attempts to learn linguistic resources which are not used in RBMT, but are in fact similar to structural transfer rules. For instance, in the example-based MT (EBMT) framework (Carl and Way, 2003), some researchers have dealt with the problem of inferring a kind of translation rules called *translation templates* (Kaji et al., 1992; Brown, 1999; Cicekli and Güvenir, 2001). A translation template can be defined as a bilingual pair of sentences in which corresponding units (words or phrases) are coupled and replaced with variables. Liu and Zong (2004) summarise different translation template acquisition methods. Other approaches with which to learn structural transformations in the EBMT framework include, among others, the acquisition of transfer mappings from bilingual corpora (Menezes and Richardson, 2003) and the induction of probabilistic translation grammars from syntactically-parsed parallel sentences (Carl, 2001).

There are multiple differences between our approach and those applied in the EBMT framework. For instance, our approach is mainly based on lexical forms consisting of lemma, lexical category and morphological inflection information, while EBMT translation templates use variables which, even though they may be linguistically motivated (e.g. NOUN, VERB, NP, PP), do not include lower level morphological inflection attributes (e.g. gender, number, person, case) whose values can be obtained through references to those in other variables. Another distinguishing feature is that EBMT translation templates can be hierarchical and consequently more than one translation template may be applied to a given SL segment. In addition, the way in which transfer rules are usually applied in shallow-transfer RBMT and in EBMT are also different: whereas in shallow-transfer RBMT rules are applied in a left-to-right longest match greedy manner and a SL lexical form can only be processed by a single transfer rule, in EBMT different translation templates, not necessarily nested ones, can match the same SL word, i.e. EBMT allows the overlapping matching of translation templates.

Finally, in the SMT framework, the use of ATs (Och, 2002; Och and Ney, 2004) can be seen as an integration of translation rules into statistical translation models, since an AT is a generalisation or an abstraction of the transformations to be applied when translating SL into TL by using word classes. Hierarchical SMT systems (Chiang, 2007), in which hierarchical statistical translation rules are learnt from parallel corpora, are also moderately similar to our approach, particularly when the rules have many different non-terminal symbols (Zollmann and Vogel, 2011). The differences are again that the shallow-transfer rules in our approach are flat, less structured and non-hierarchical. In addition, the application of shallow-transfer rules is not statistically driven.

6. Experimental settings

Our method has been evaluated by comparing the number of GATs extracted and the resulting translation quality with those obtained by (a) following the method proposed by Sánchez-Martínez and Forcada (2009), (b) using hand-written rules, and (c) using no rules at all (word-for-word translation). In order to assess

the contribution of the different methods that are used in our approach for the improvement in translation quality, we have also evaluated the translation quality obtained when: (d) wildcards and reference values are not used in word classes (i.e. the function σ_2 described in Section 4.2.3 returns the input set of GATs unchanged); and (e) the approach by Sánchez-Martínez and Forcada (2009) benefits from the method described in Section 4.5 when selecting the final set of rules, which ensures a convenient chunking of the input. We have also tested the translation performance of the combination of hand-written rules and rules inferred with our approach.

The evaluation covers a wide variety of language pairs: pairs in which the two languages involved in the translation belong to the same language family (Spanish↔Catalan; the arrows show the translation direction), in addition to pairs in which the languages belong to different language families (English↔Spanish and Breton→French).[23]

The Spanish–Catalan training corpus consists of parallel sentences extracted from the newspaper *El Periodico de Catalunya*,[24] which is published in both languages; the test corpus consists of sentences randomly chosen from the *Revista Consumer Eroski* parallel corpus (Alcázar, 2005), which contains product reviews. The English–Spanish rules have been inferred from the Europarl Parallel Corpus (Koehn, 2005) version 7, a collection of minutes from the European Parliament, and they have been evaluated with the *newstest2013* corpus, a set of parallel sentences extracted from pieces of news and released as part of the shared translation task of the eighth Workshop on Statistical Machine Translation (Bojar et al., 2013). The Breton–French training and test corpora have both been randomly extracted from the collection compiled by Tyers (2009) from a heterogeneous set of sources, including software localisation and tourism.

In order to evaluate the impact of the size of the training corpus on the quality of the resulting translations, subsets containing 100, 250, 500, 1 000, 2 500 and 5 000 sentences have been randomly extracted from each training corpus in a such a way that we have ensured that all the sentences in the smaller subsets are contained in the bigger ones.[25] For English↔Spanish and Breton→French, two additional subsets containing 10 000 and 25 000 sentences respectively have also been used to evaluate our approach when no generalisation of the morphological inflection attributes is performed (i.e. no wildcard and reference values are used). Each corpus subset has then been split into two parts: the largest one, containing $\frac{4}{5}$ of the sentences, has been used as the actual training subset from which GATs are extracted, whereas the remaining sentences have been used as the development set to determine the threshold values to be used with each method (see below).[26] Table 1 provides the number of sentences in the training and development corpora, the number of words and the size of the vocabulary for each language pair and corpus size; Table 2 provides these data for the different test sets used for evaluation.[27]

With regard to the threshold used by each method, Sánchez-Martínez and Forcada (2009) use a threshold to discard the EATs that reproduce a number of bilingual phrase pairs below its value; this threshold is obtained as the integer value between 1 and 10 which maximises the BLEU score (Papineni et al., 2002) in the development corpus. Our approach uses two different thresholds, θ and δ, as described in Section 4.3. The value of δ (used to discard those GATs with an inadequate ratio of bilingual phrase pairs correctly reproduced over the total number of bilingual phrase pairs matched) has been chosen by trying all the values in the range $[0, 1]$ at increments of 0.05 and selecting the value that maximises the BLEU score in the development set. With regard to θ (used to discard GATs that reproduce a small number of bilingual phrase pairs), we have used different values, one for each different minimisation subproblem (one subproblem per sequence of SL lexical categories), to ensure that the number of input GATs to each of the different

[23] Although languages with grammatical case are not covered in the experimental set-up, dealing with them would not require any modification in the algorithms described in this paper: the grammatical case is just another morphological feature. It remains to be studied how the presence of grammatical case affects the data requirements of the rule inference algorithm.

[24] `http://www.elperiodico.com/`

[25] As it is usually done in SMT, only sentences containing at most 45 words have been chosen in order to prevent GIZA++ from truncating long sentences.

[26] For the subsets containing 25 000 sentences, the training part contains 23 000 sentences, while the development section contains the remaining 2 000 sentences.

[27] For a given language pair, the same test set has been used to evaluate the systems built with the different sizes of the training corpus.

(a) Spanish↔Catalan

training + development # sentences	Spanish		Catalan	
	# words	# vocabulary	# words	#vocabulary
100	1,539	789	1,597	798
250	3,830	1,684	3,969	1,685
500	7,697	2,985	7,939	2,946
1,000	15,136	5,062	15,576	4,959
2,500	37,301	9,783	38,470	9,580
5,000	73,637	15,315	75,981	14,933

(b) English↔Spanish

training + development # sentences	English		Spanish	
	# words	# vocabulary	# words	#vocabulary
100	2,145	913	2,151	945
250	5,460	1,868	5,672	1,992
500	11,228	3,016	11,704	3,342
1,000	22,447	4,653	23,292	5,209
2,500	56,003	7,756	57,961	8,984
5,000	113,290	11,045	117,051	13,197
10,000	227,088	15,329	234,854	18,723
25,000	571,364	22,703	589,400	28,927

(c) Breton→French

training + development # sentences	Breton		French	
	# words	# vocabulary	# words	#vocabulary
100	1,319	714	1,520	760
250	3,451	1,537	3,768	1,621
500	6,937	2,623	7,565	2,833
1,000	14,456	4,364	15,863	4,915
2,500	35,335	7,846	37,931	9,038
5,000	69,500	11,880	75,427	14,008
10,000	141,838	17,741	153,455	20,985
25,000	354,417	28,828	387,354	34,117

Table 1: Number of sentences, number of words, and vocabulary size of the training and development corpora for each language pair and corpus size. These corpora are divided into training (4/5 of the sentences) and development (1/5 of the sentences).

Language pair	Source			Target		
	#sentences	#words	#voc	#sentences	#words	#voc
English–Spanish	3,000	62,873	10,867	3,000	67,762	12,400
Spanish–Catalan	3,000	76,794	13,414	3,000	78,089	13,130
Breton–French	3,000	41,800	8,824	3,000	45,278	10,211

Table 2: Number of sentences, words, and size of the vocabulary of the test set used for evaluation for each language pair.

minimisation subproblems is below 1 000; in any case a minimum value of 2 has been established for θ to discard those GATs that are only able to reproduce a single bilingual phrase pair. This is done to make the minimisation problem computationally feasible. For the experiments where wildcards and reference values are not used, the value of δ has been optimised in the same way, while the value of θ has been always set to 2 because the computational complexity of the minimisation problem is much smaller.

With respect to the similarity measure used to optimise the rules for chunking and select the sequences of lexical categories for which rules will eventually be generated (see Section 4.5), we have used BLEU with the smoothing implemented by the National Institute of Standards and Technology's (NIST)[28] to avoid null values when it is used at the sentence level.

All the experiments have been carried out with the translation engine[29] and linguistic data[30] of the rule-based MT system Apertium (Forcada et al., 2011). We have released a software package[31] which implements the pipeline for the inference of shallow-transfer rules as described in Section 4. However, some external tools have also been used, namely, the minimisation subproblems have been solved with the integer linear programming Cbc solver,[32] while word alignment and bilingual phrase pair extraction were carried out by using the Giza++ toolkit (Och and Ney, 2003) and the phrase extraction implementation in the Moses statistical MT system (Koehn et al., 2007), respectively. It is worth noting that before word alignment we added the Apertium bilingual dictionary to the corpus and removed it afterwards. This has improved word alignment when the amount of parallel sentences is scarce.

Finally, we have added two heuristics to our method presented in Section 4, one to further reduce the number of input GATs to each minimisation subproblem (see Section 6.1), and another to discard bilingual phrase pairs before GATs are generated from them (see Section 6.2).

6.1. Reducing the Number of Input GATs to the Minimisation Subproblems

As explained in Section 4.2.3, in order to introduce wildcards and SL and TL references in the morphological inflection attributes of a GAT $z = (S, T, A, R)$, σ_2 considers the power set $\mathcal{P}(C)$ of the set C with the attributes that can be generalised in z. This could lead to a situation in which the minimisation subproblems are unsolvable in a reasonable amount of time as a result of the combinatorial explosion that occurs when generating GATs for the translation between highly inflected languages, such as some of those in our experimental settings (e.g. Spanish, Catalan). In order to reduce the amount of GATs generated by σ_2 in our experiments, only one subset $H_C \subset \mathcal{P}(C)$ has been considered for each GAT; this subset is defined as:

$$H_C = \{H : H \in \mathcal{P}(C) \ \wedge \ (\forall s_i \in S, \nexists a, a' : a \in H \ \wedge \ a' \notin H \ \wedge \ \mathrm{rank}(\rho(s_i), a) < \mathrm{rank}(\rho(s_i), a')\};$$

where $\mathrm{rank}(c, a)$ returns the position of the morphological inflection attribute a in the list, ordered in decreasing order of *specificity*, of the morphological inflection attributes associated with the lexical category c. An attribute a is considered to be more specific than another attribute a' if it is applicable to a smaller number of lexical categories, e.g. the attribute verb tense is more specific than the attribute number because it can only be applied to verbs, whereas number can be applied to verbs, nouns, pronouns and (in some languages) adjectives and determiners. Therefore, for a lexical category c (e.g. verb) an attribute a (e.g. verb tense) is generalised only if the more general attributes of c (e.g. number and person) are also generalised.[33]

[28] MTeval utility version 13; `ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13.pl`.

[29] More specifically, revision 47871 of the Subversion repository at `https://svn.code.sf.net/p/apertium/svn/trunk/apertium`.

[30] Repository for English–Spanish: `https://svn.code.sf.net/p/apertium/svn/trunk/apertium-en-es`, revision 41294; Spanish–Catalan: `https://svn.code.sf.net/p/apertium/svn/trunk/apertium-es-ca`, revision 34111; Breton–French: `https://svn.code.sf.net/p/apertium/svn/trunk/apertium-br-fr`, revision 28674; Chinese–Spanish: `https://svn.code.sf.net/p/apertium/svn/incubator/apertium-zho-spa`, revision 49858.

[31] The latest version can be downloaded from the Subversion repository at `https://svn.code.sf.net/p/apertium/svn/trunk/apertium-transfer-tools-generalisation`.

[32] `https://projects.coin-or.org/Cbc`, version 2.7.

[33] In practice H_C does not need to be explicitly calculated because the ordering provided by $\mathrm{rank}(\cdot)$ matches that used to codify the morphological inflection attributes in the Apertium dictionaries.

6.2. Filtering Bilingual Phrase Pairs

When extracting the bilingual phrase pairs (see section 4.1) to be used both with our approach and with that by Sánchez-Martínez and Forcada (2009), the following criteria have been applied:

- The maximum length allowed for the SL and TL side of a bilingual phrase pair has been set to 5 in order to limit the number of minimisation subproblems to be solved.
- Bilingual phrase pairs containing either unknown words or punctuation marks have been removed. On the one hand, bilingual phrase pairs containing unknown words are not useful unless a morphological guesser is used; on the other hand, we assume that punctuation marks do not provide relevant information from the point of view of the structural transference.
- Bilingual phrase pairs whose first or last word on either side (SL and TL) are left unaligned have been discarded because there is no evidence that they are actually part of the translation of the segment in the opposite language, and using them could result in incorrect GATs.
- Bilingual phrase pairs that are not consistent with the bilingual dictionary have also been discarded to avoid unnecessary lexicalisations (see next section).

6.2.1. Bilingual Phrase Pairs Consistent with the Bilingual Dictionary

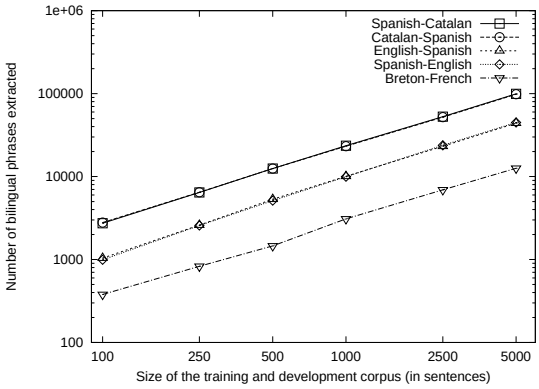
Recall that our approach obtains a set of GATs which correctly reproduces all the bilingual phrase pairs, and that, when the translation of a word in a bilingual phrase pair does not appear as an equivalent in the bilingual dictionary, the GATs obtained from it need to be lexicalised, i.e. its lemma cannot be removed from the corresponding word classes. If a bilingual phrase pair consists of a *free* translation or contains translation equivalents that are different to those in the Apertium dictionaries, an unnecessary lexicalisation may occur. To avoid these unnecessary lexicalisations while allowing the method to learn common lexical changes between the SL and the TL, we have filtered the set of bilingual phrase pairs obtained. Those bilingual phrase pairs for which one of the following conditions is not met for all SL and TL lexical forms have been discarded:

1. If the lexical form w is an *open-class* lexical form, i.e. it belongs to an open-class lexical category,[34] it must be either aligned with an open-class lexical form in the other language that appears in the bilingual dictionary as an equivalent for w, or otherwise not aligned with any open-class lexical form. If it is a *closed-class* lexical form it may be aligned to any lexical form. This filtering is based on the assumption that open-class words carry the meaning of the sentence while the role of closed-class words is to provide grammatical information.
2. If an (open-class) lexical form w does not meet the previous condition, it must be single-aligned to an open-class lexical form in the other language that meets the first condition. This second condition is based on the assumption that here the open-class lexical form that does not meet the first condition might be working as an auxiliary particle, and does not therefore convey any meaning.

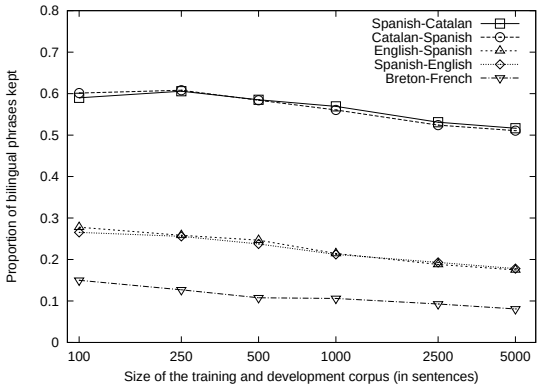
When inferring rules with the method by Sánchez-Martínez and Forcada (2009), the filtering is much simpler and consists of discarding those phrase pairs with at least one lexical form aligned with a lexical form in the other language that does not appear in the bilingual dictionary as its equivalent and that does not belong to the set of lexicalised units provided by the user. For the Spanish↔Catalan experiments we have used the set of lexicalised units originally defined by Sánchez-Martínez and Forcada (2009); for the rest of language pairs we have used the set of closed-class lexical forms instead.

Figure 13 shows the number of bilingual phrase pairs obtained from the different training corpora after applying the filtering criteria described above. These bilingual phrase pairs have then been used to infer

[34] Nouns, adjectives, adverbs, and verbs are among the set of open-class lexical categories; whereas determiners, pronouns and prepositions are considered to be closed-class lexical categories.



(a) Number of bilingual phrases obtained after filtering.



(b) Proportion of bilingual phrases kept after filtering.

Figure 13: Number of bilingual phrases obtained from the training corpora after applying the filtering criteria defined in Section 6.2 (left) and proportion of the bilingual phrases with length 5 or lower initially extracted from the parallel corpus that are kept after the filtering (right). The extraction of bilingual phrases is the first step of the rule inference algorithm described in Section 4.

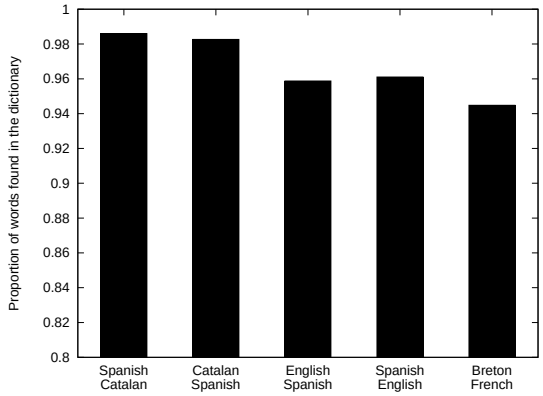


Figure 14: Proportion of words in the test set for which there is at least one analysis in the Apertium dictionary, for the different language pairs.

GATs with our approach. The figure also depicts the proportion of bilingual phrase pairs discarded as a result of the filtering. Note that the number of bilingual phrase pairs discarded for Spanish↔Catalan is much smaller than for the other language pairs. This is because Spanish and Catalan are closely-related languages with less lexical translation ambiguity, which signifies that more translation equivalents in the bilingual phrase pairs match those in the bilingual dictionary. In addition, Breton→French is the language pair with the highest proportion of discarded bilingual phrase pairs because its dictionaries have a low coverage, as shown in Figure 14.

7. Results and discussion

The translation quality achieved by the rules inferred when they are used with the Apertium RBMT engine, and the exact number of ATs obtained with each approach, are presented in figures 15–19. Translation quality has been estimated using the automatic evaluation metrics BLEU (Papineni et al., 2002), TER (Snover et al., 2006) (the figures represent 1-TER) and METEOR (Banerjee and Lavie, 2005). We have also tested whether our approach outperforms the approach proposed by Sánchez-Martínez and Forcada (2009) (henceforth, baseline approach) by a statistically significant margin through the use of paired bootstrap resampling (Koehn, 2004a) with each evaluation metric and test set ($p \leq 0.05$, $1,000$ iterations); if the difference between the two approaches is statistically significant, a diagonal cross is placed on top of the points that represent the results of the approach that performs best. The figures also show the coverage provided by the rules, i.e., the proportion of words in each test set that have been translated using an AT, and the time spent on the inference of the ATs from the bilingual phrase pairs.[35]

The results show that, overall, our approach (Sánchez-Cartagena et al.) outperforms the baseline approach (Sánchez-Martínez and Forcada) by a statistically significant margin ($p \leq 0.05$) for all language pairs and automatic evaluation metrics. As expected, the translation quality of both approaches lies between the translation quality achieved by a word-for-word translation and a translation performed using hand-written rules. Our approach achieves results close to those obtained with hand-written rules and, in the case of Breton→French, it even outperforms the use of hand-written rules when translation quality is evaluated using TER (Figure 19(b)). This may be explained by the fact that the Breton→French hand-written rules are less mature since less work seems to have been carried out for their development.[36]

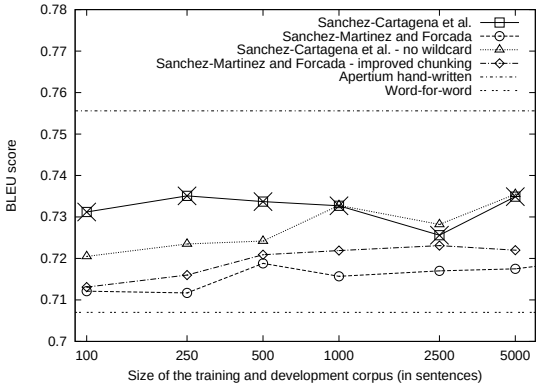
In general, the translation quality achieved by the rules inferred with our approach grows with the size of the training corpus for the language pairs which are not closely related, namely English↔Spanish and Breton→French. Systems built with the baseline approach also follow this pattern. In the case of closely-related languages, e.g. Spanish↔Catalan, the translation quality grows with the amount of corpora used for training at a slower pace and with some fluctuations (Catalan→Spanish) or does not grow at all (Spanish→Catalan). These results suggest that a few hundred parallel sentences are sufficient to infer useful shallow-transfer rules for closely-related language pairs, since it would appear that no clear improvement is obtained by increasing the size of the training corpus. The drop in translation quality detected by the three metrics for Spanish→Catalan when the training corpus contains $2,500$ sentence pairs is caused by an inadequate value of δ: the value which optimizes the BLEU score in the development set appears to cause a drop in performance in the test set.

The difference in performance between our approach and the baseline is reduced as the amount of corpora used for training grows, mainly because the effect of generalising the morphological inflection attributes is stronger when the corpus is very small (see below). However, the effect of the filtering based on the threshold θ performed to reduce the amount of input GATs to each minimisation subproblem should also

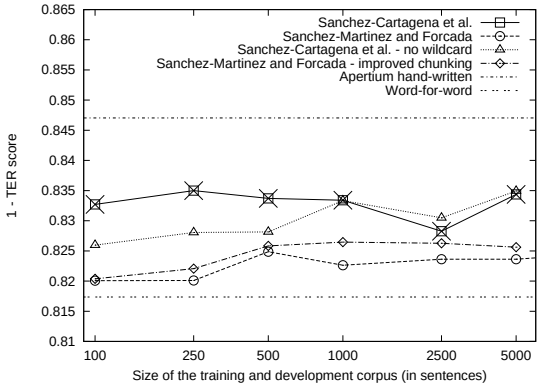
[35] For our approach, the time is computed as the sum of the processes described in Sections 4.2 and 4.4 for the best threshold δ, since they constitute the most time-consuming part of the rule inference pipeline. For the approach by Sánchez-Martínez and Forcada (2009), the time reported is that spent on the generation of the final set of EATs from the set of bilingual phrases for the best threshold θ. The experiments have been executed in a computing cluster with 26 computing nodes with a hexacore Intel Xeon X5660 CPU each one. Times displayed are the sum of the times of the different parallel jobs.

[36] This conclusion is drawn from the number of commits made in the Apertium Subversion repository affecting the files containing the rules in each language pair.

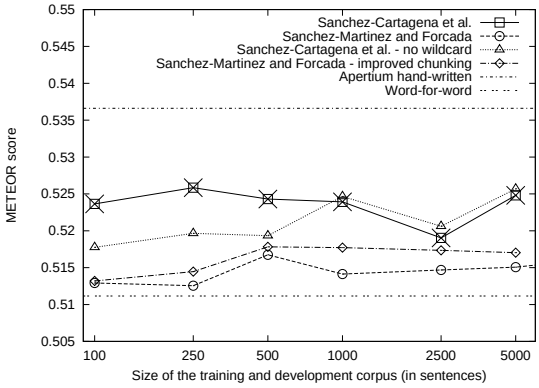
Spanish→Catalan



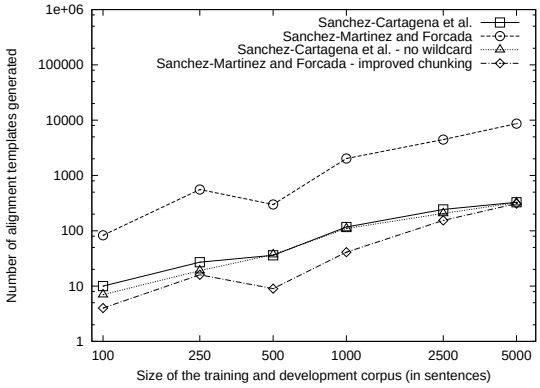
(a) Translation quality measured using BLEU.



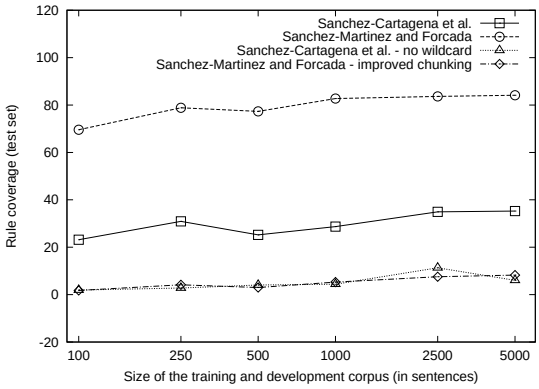
(b) Translation quality measured using TER.



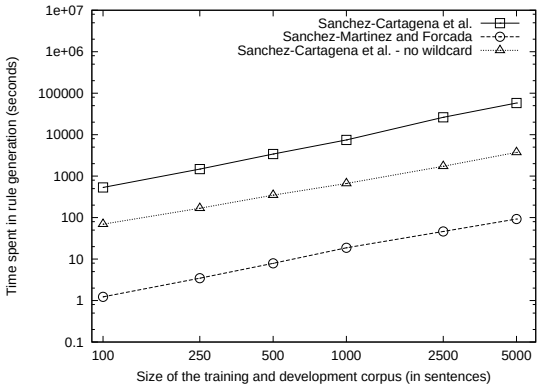
(c) Translation quality measured using METEOR.



(d) Number of alignment templates inferred.



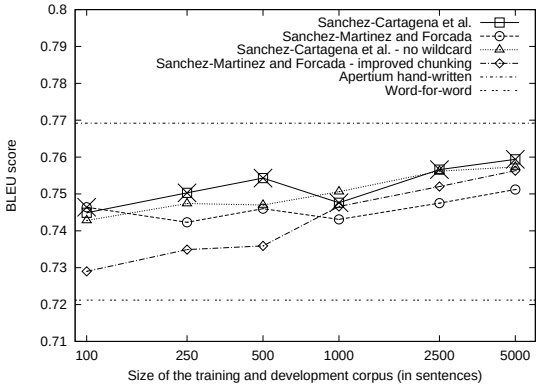
(e) Proportion of words from the test set translated by an alignment template.



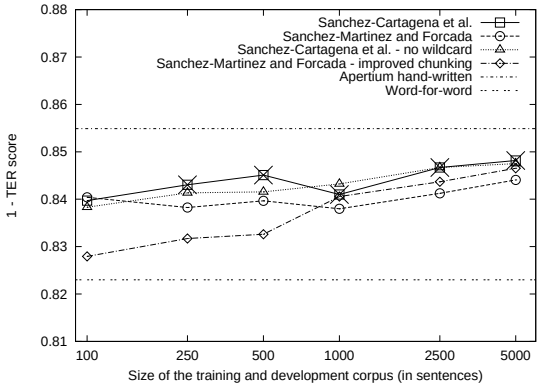
(f) Computing time required to infer alignment templates.

Figure 15: Translation quality, number of alignment templates inferred, coverage (proportion of words in the test set translated by an alignment template) and computing time required to infer alignment templates from the different systems evaluated for the Spanish→Catalan language pair. A diagonal cross over a square point indicates that our approach outperforms the baseline approach proposed by Sánchez-Martínez and Forcada (2009) by a statistically significant margin ($p \leq 0.05$). If the cross is over a circle, the baseline outperforms our approach.

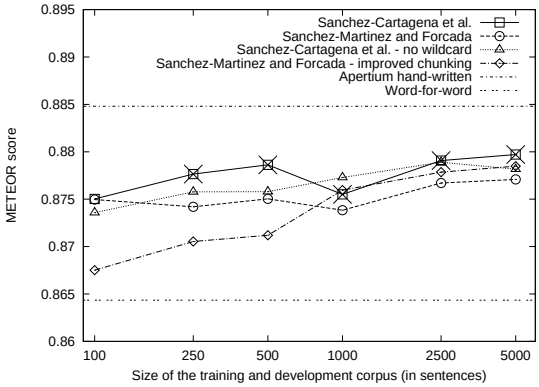
Catalan→Spanish



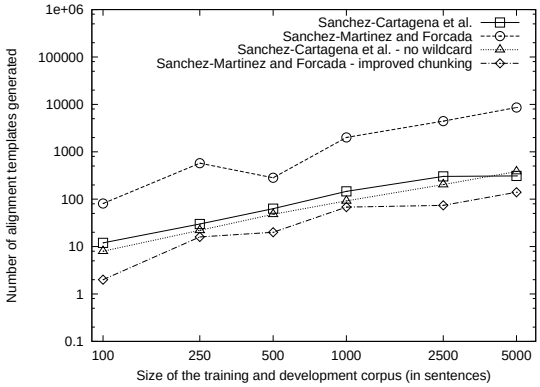
(a) Translation quality measured using BLEU.



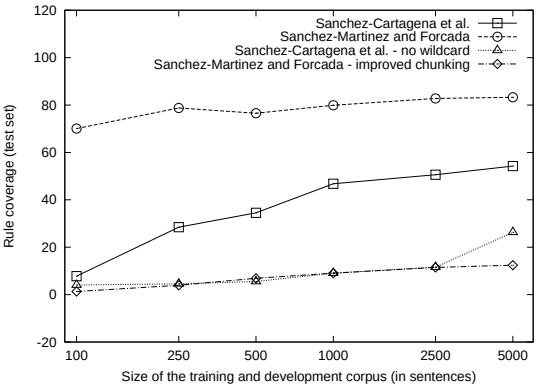
(b) Translation quality measured using TER.



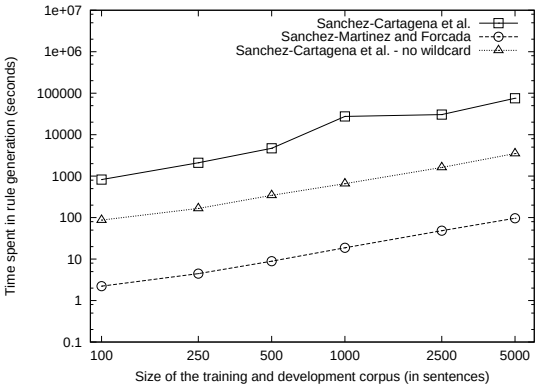
(c) Translation quality measured using METEOR.



(d) Number of alignment templates inferred.



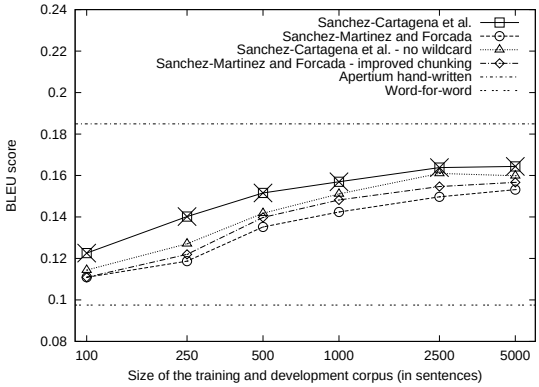
(e) Proportion of words from the test set translated by an alignment template.



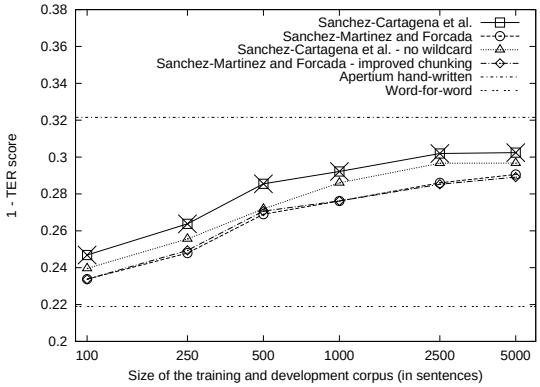
(f) Computing time required to infer alignment templates.

Figure 16: Translation quality, number of alignment templates inferred, coverage (proportion of words in the test set translated by an alignment template) and computing time required to infer alignment templates from the different systems evaluated for the Catalan→Spanish language pair. A diagonal cross over a square point indicates that our approach outperforms the baseline approach proposed by Sánchez-Martínez and Forcada (2009) by a statistically significant margin ($p \leq 0.05$). If the cross is over a circle, the baseline outperforms our approach.

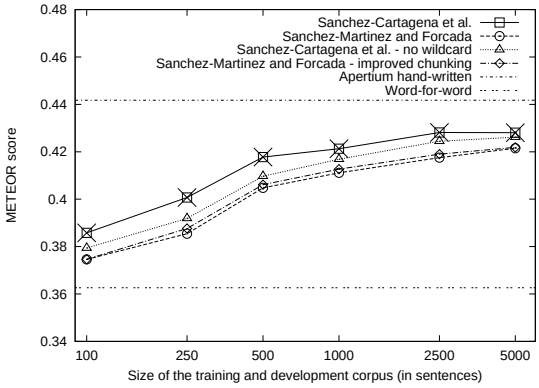
English→Spanish



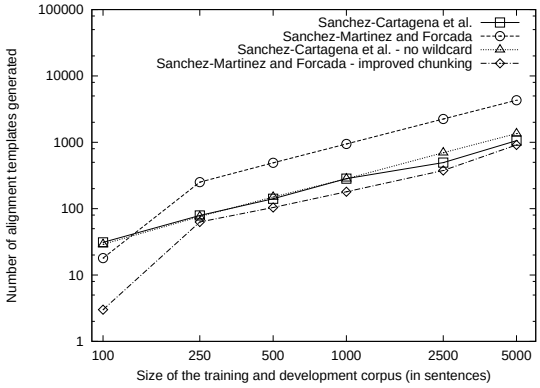
(a) Translation quality measured using BLEU.



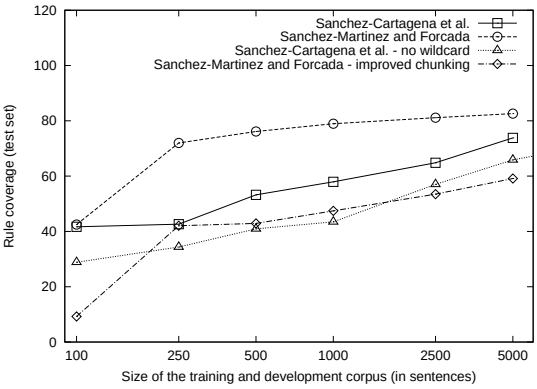
(b) Translation quality measured using TER.



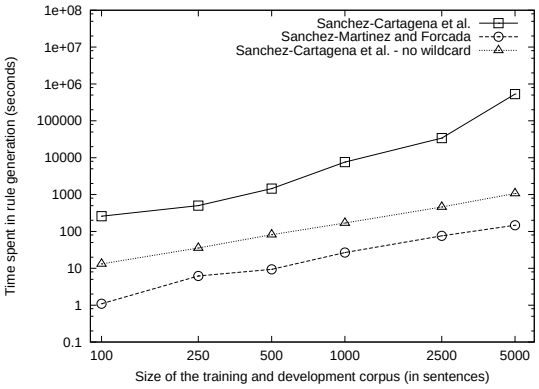
(c) Translation quality measured using METEOR.



(d) Number of alignment templates inferred.



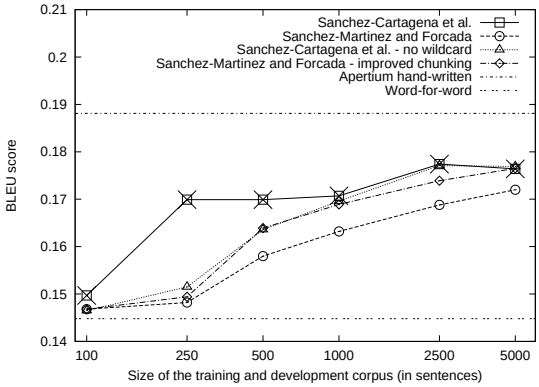
(e) Proportion of words from the test set translated by an alignment template.



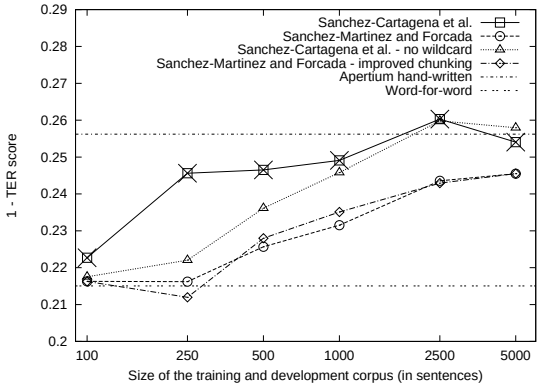
(f) Computing time required to infer alignment templates.

Figure 17: Translation quality, number of alignment templates inferred, coverage (proportion of words in the test set translated by an alignment template) and computing time required to infer alignment templates from the different systems evaluated for the English→Spanish language pair. A diagonal cross over a square point indicates that our approach outperforms the baseline approach proposed by Sánchez-Martínez and Forcada (2009) by a statistically significant margin ($p \leq 0.05$). If the cross is over a circle, the baseline outperforms our approach.

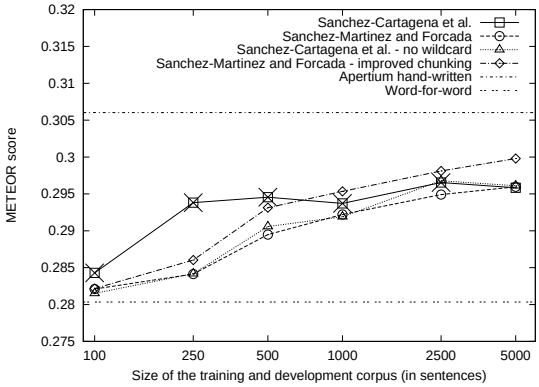
Spanish→English



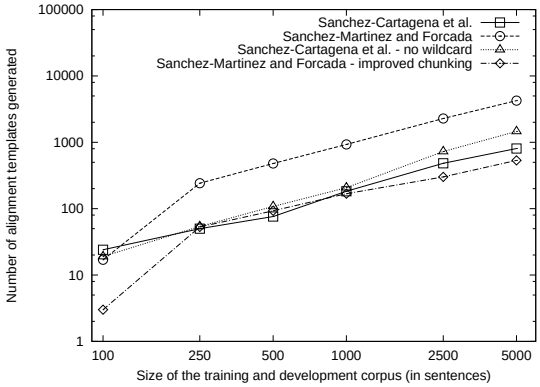
(a) Translation quality measured using BLEU.



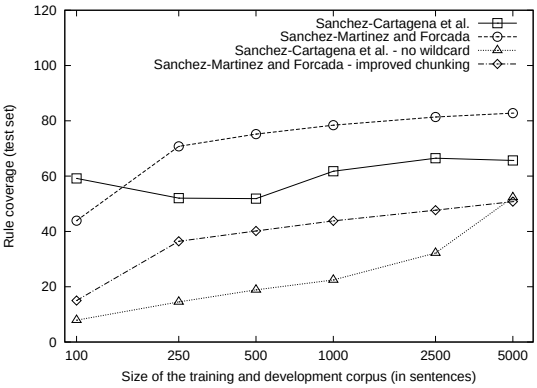
(b) Translation quality measured using TER.



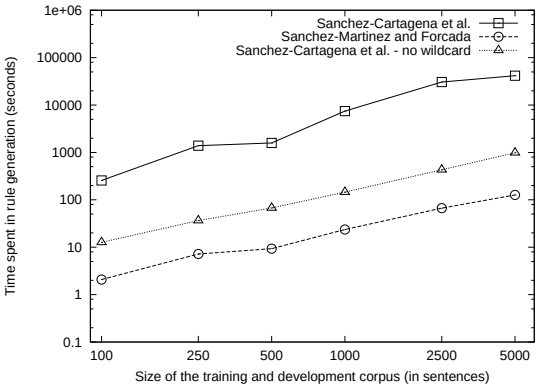
(c) Translation quality measured using METEOR.



(d) Number of alignment templates inferred.



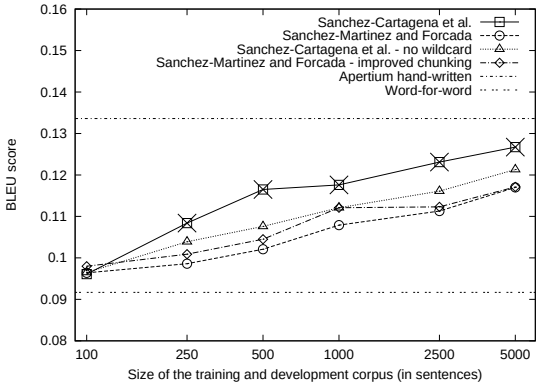
(e) Proportion of words from the test set translated by an alignment template.



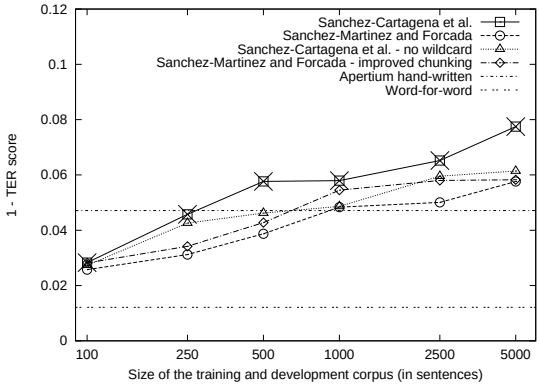
(f) Computing time required to infer alignment templates.

Figure 18: Translation quality, number of alignment templates inferred, coverage (proportion of words in the test set translated by an alignment template) and computing time required to infer alignment templates from the different systems evaluated for the Spanish→English language pair. A diagonal cross over a square point indicates that our approach outperforms the baseline approach proposed by Sánchez-Martínez and Forcada (2009) by a statistically significant margin ($p \leq 0.05$). If the cross is over a circle, the baseline outperforms our approach.

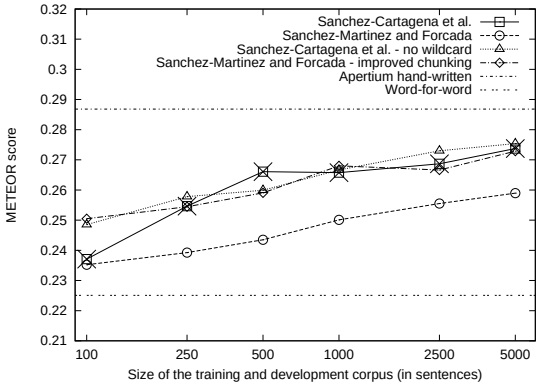
Breton→French



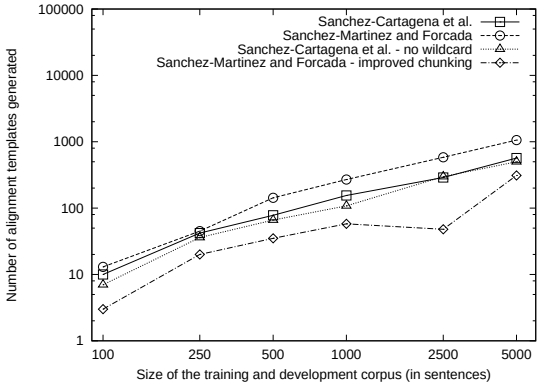
(a) Translation quality measured using BLEU.



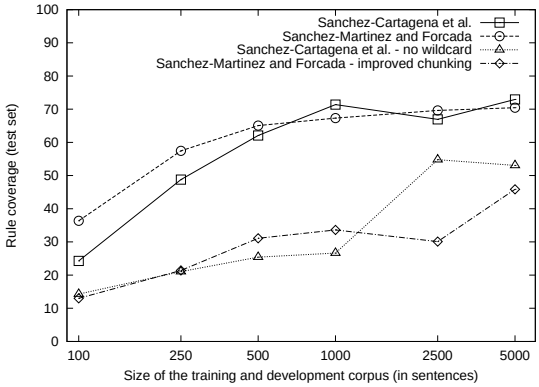
(b) Translation quality measured using TER.



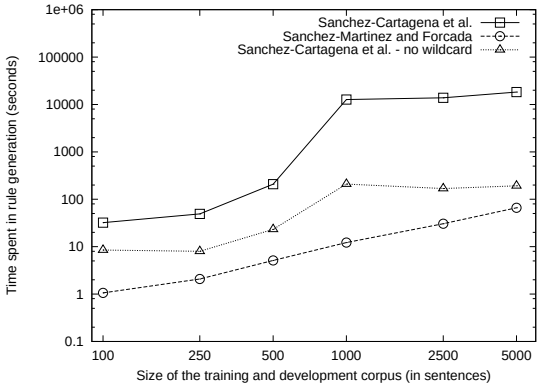
(c) Translation quality measured using METEOR.



(d) Number of alignment templates inferred.

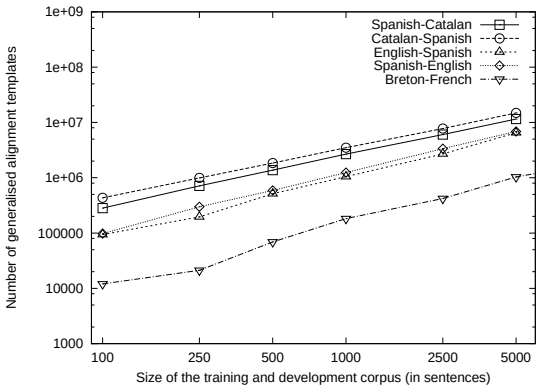


(e) Proportion of words from the test set translated by an alignment template.

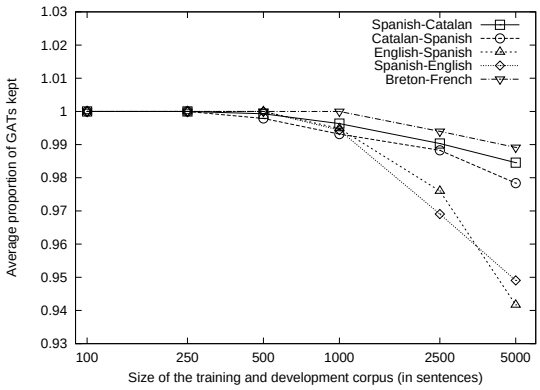


(f) Computing time required to infer alignment templates.

Figure 19: Translation quality, number of alignment templates inferred, coverage (proportion of words in the test set translated by an alignment template) and computing time required to infer alignment templates from the different systems evaluated for the Breton→French language pair. A diagonal cross over a square point indicates that our approach outperforms the baseline approach proposed by Sánchez-Martínez and Forcada (2009) by a statistically significant margin ($p \leq 0.05$). If the cross is over a circle, the baseline outperforms our approach.



(a) Number of GATs initially generated from the set of bilingual phrases extracted from the parallel corpus.



(b) Average proportion of GATs retained after applying the filtering based on the threshold θ.

Figure 20: For each language pair, the number of GATs initially generated from the set of bilingual phrases (left), and average proportion of GATs retained after applying the filtering based on the threshold θ and described at the end of Section 6 (right). The values reported correspond to the filtering performed on the GATs obtained with a value of $\delta = 0$, and a value of θ automatically chosen for each minimisation subproblem to limit the number of input GATs to $1,000$. GATs that do not reproduce at least 2 bilingual phrases have been excluded from the computation of the proportion, since they are always discarded (see Section 6).

be considered. Figure 20 shows the average proportion of GATs retained after applying the filtering based on the threshold θ (described at the end of Section 6) for the different language pairs and training corpus sizes. The proportion of GATs retained after the filtering starts to decrease at a faster pace when the size of the training corpora exceeds $1,000$ sentences. This decrease is less sharp for the Breton→French language pair because the Breton–French dictionaries have a lower *coverage* (see Figure 14)[37] and the amount of bilingual phrases extracted is consequently lower when compared to the other language pairs (see Figure 13). Contrarily, the most pronounced decrease occurs in both directions of the English↔Spanish pair. Even though Spanish↔English is not the language pair for which the highest amount of bilingual phrase pairs are extracted, it is the pair for which a greater amount of GATs are discarded in order to meet the limit of $1,000$ input GATs per minimisation subproblem. This may be explained by the fact that English and Spanish are more distant languages than Spanish and Catalan (which are closely related), for which more bilingual phrase pairs are extracted.

A comparison between the performance of our approach and the alternative approach that does not generalise the morphological inflection attributes (Sánchez-Cartagena et at. - no wildcard, figures 15–19) shows that translation quality grows at a similar rate for both approaches when the corpus size is above $1\,000$ sentences. Recall that when no generalisation of the morphological inflection attributes is performed, no pruning takes place because θ is set to 2 for all the minimisation subproblems. These results suggest that the pruning based on θ has little impact on translation quality since, otherwise, a bigger drop in translation quality would occur.

With respect to the number of GATs eventually included in the rules, also shown in figures 15–19 for the different corpus sizes we have used, for most of the language pairs, the number of GATs inferred is one order of magnitude (and in some cases almost two) lower than the amount of EATs obtained with the baseline approach. The greater expressiveness of our formalism with regard to the baseline approach, and the selection of GATs used to optimise the chunking of the sentences to be translated have led to this reduction in the number of GATs. This reduction is expected to alleviate the effort needed to manually edit the set of inferred rules, if it is necessary to do so. An analysis of the coverage of the test set with the GATs obtained with the different approaches shows another advantage of selecting the GATs to optimise the

[37] Coverage here is defined as the proportion of surface forms (running words) for which there is at least one possible analysis in the dictionaries being used; note that this does not mean that the correct analysis is returned.

chunking and the removal of redundant GATs: our approach achieves better translation quality by applying fewer rules, i.e., only the words which actually need to be processed together are covered by GATs.

As regards the relative impact on the translation quality of the different improvements in comparison to the method proposed by Sánchez-Martínez and Forcada (2009) presented in this paper, it can be observed that the generalisation of morphological inflection attributes with wildcards and reference values brings a clear advantage, but in general, only when the training corpus is really scarce (less than 1 000 sentences). As mentioned previously, the difference between our system and the variant that does not generalise the morphological inflection attributes (Sánchez-Cartagena et al. - no wildcard) disappears or becomes very small for most of the language pairs when the corpus size exceeds 1 000 sentences. We can therefore conclude that the overhead brought by the generalisation of morphological inflection attributes is justified and, when its computational cost starts to be prohibitively high (see the computing time required to infer alignment templates in figures 15–19(f)), the improvement in translation quality that can be expected is really small.

It is also worth comparing the results obtained using the alternative approach that does not generalise the morphological inflection attributes (Sánchez-Cartagena et al. - no wildcard) with those obtained using the approach proposed by Sánchez-Martínez and Forcada (2009) with improved chunking (Sánchez-Martínez and Forcada - improved chunking), that are also depicted in figures 15–19. A higher translation quality is generally obtained with the first approach. An analysis of the rules inferred by both systems confirms that GATs with more appropriate lexicalised word classes can be obtained by following our strategy. Moreover, we have detected that the input sentences are not chunked in the most convenient way when the rules are inferred with the approach proposed by Sánchez-Martínez and Forcada (2009), even when it is complemented with the strategy aimed at improving chunking (see Section 4.5); this fact is especially relevant in the Spanish↔Catalan language pairs. These results suggest that the method described in Section 4.5 loses effectiveness when it is not applied to the result of the global minimisation problem.

Given that the positive impact of generalising the morphological inflection attributes is only remarkable for small corpora, disabling it allows us to scale up our approach to bigger corpora. In particular, we have evaluated our approach with two more subsets of the training corpora that contain 10 000 and 25 000 sentences, respectively. The only language pairs used in this evaluation were the English↔Spanish and Breton→French language pairs, since they are those for which the experimental results described previously suggest that translation quality may continue growing at a fast pace with the size of the corpus.

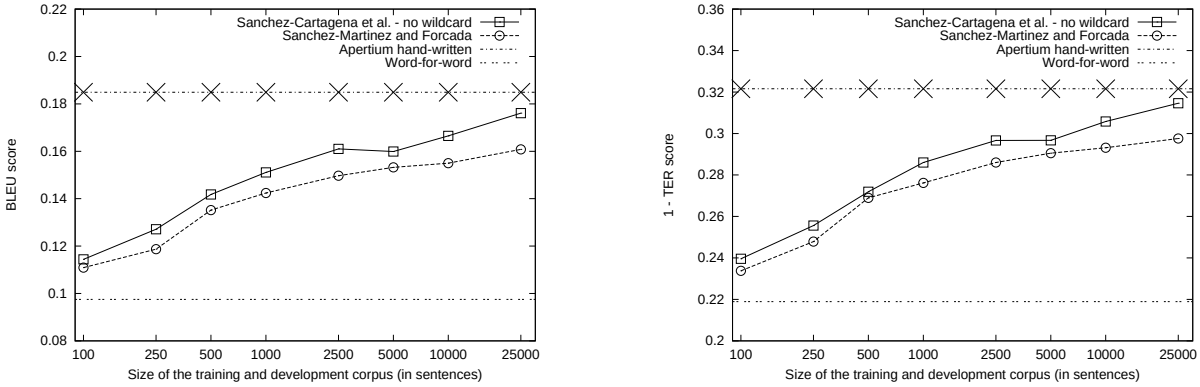
Figures 21–23 show the translation quality achieved by the rules inferred by our approach when no generalisation of the morphological inflection attributes is performed (i.e. without wildcards and reference values) for the aforementioned language pairs and with larger corpora (the results obtained with small corpora are shown for comparison).[38] The performance of the method proposed by Sánchez-Martínez and Forcada (2009) and the hand-written rules is also presented. It can be observed that the translation quality achieved by our approach keeps growing when the size of the corpus is increased, and it still generally outperforms the approach by Sánchez-Martínez and Forcada (2009). Furthermore, as shown in Figure 22, the Spanish→English rules obtained with our method outperform the hand-written rules for the biggest corpus size by a statistically significant margin,[39] according to two of the three evaluation metrics (a diagonal cross is placed on top of the points that represent the results of our approach if they are statistically significantly better than the hand-written rules, and also over the points that represent the hand-written rules if they are statistically significantly better than our approach). Notice that the translation quality for English→Spanish and Breton→French also continues to grow.

Finally, the translation quality (as measured by BLEU (Papineni et al., 2002); the rest of metrics behave in a similar way) achieved by the combination of the hand-written rules in the Apertium project and those inferred by our approach is depicted in Figure 24. Since we wish to assess whether the linguistic information contained in the inferred rules is complementary to that in the hand-written ones or not, they have been combined in such a way that, when the longest matched text segment matches multiple rules, the most

[38] The optimisation of the parameter μ described in Section 4.5 for the sets of rules inferred from 10 000 and 25 000 sentences has been performed by means of a ternary search instead of an exhaustive search in order to speed up the process.

[39] Statistically significance margins have been computed with paired bootstrap resampling (Koehn, 2004a) ($p \leq 0.05$, $1,000$ iterations).

English→Spanish



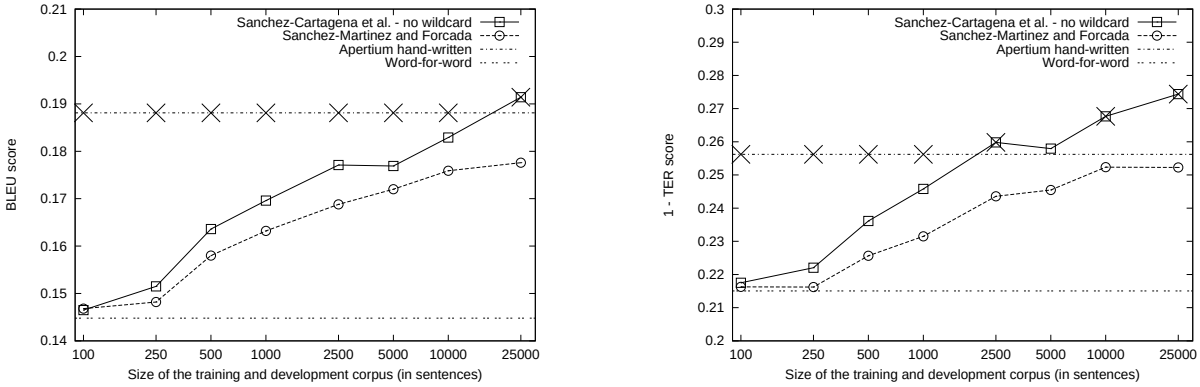
(a) Translation quality measured using BLEU.

(b) Translation quality measured using TER.

(c) Translation quality measured using METEOR.

Figure 21: Translation quality of the different systems evaluated for the English→Spanish language pair with larger corpora subsets than those used in the primary evaluation. A diagonal cross over a square point indicates that our approach outperforms the hand-written rules by a statistically significant margin ($p \leq 0.05$). A diagonal cross over the top horizontal line means that the hand-written rules outperform our approach by a statistically significant margin ($p \leq 0.05$).

Spanish→English

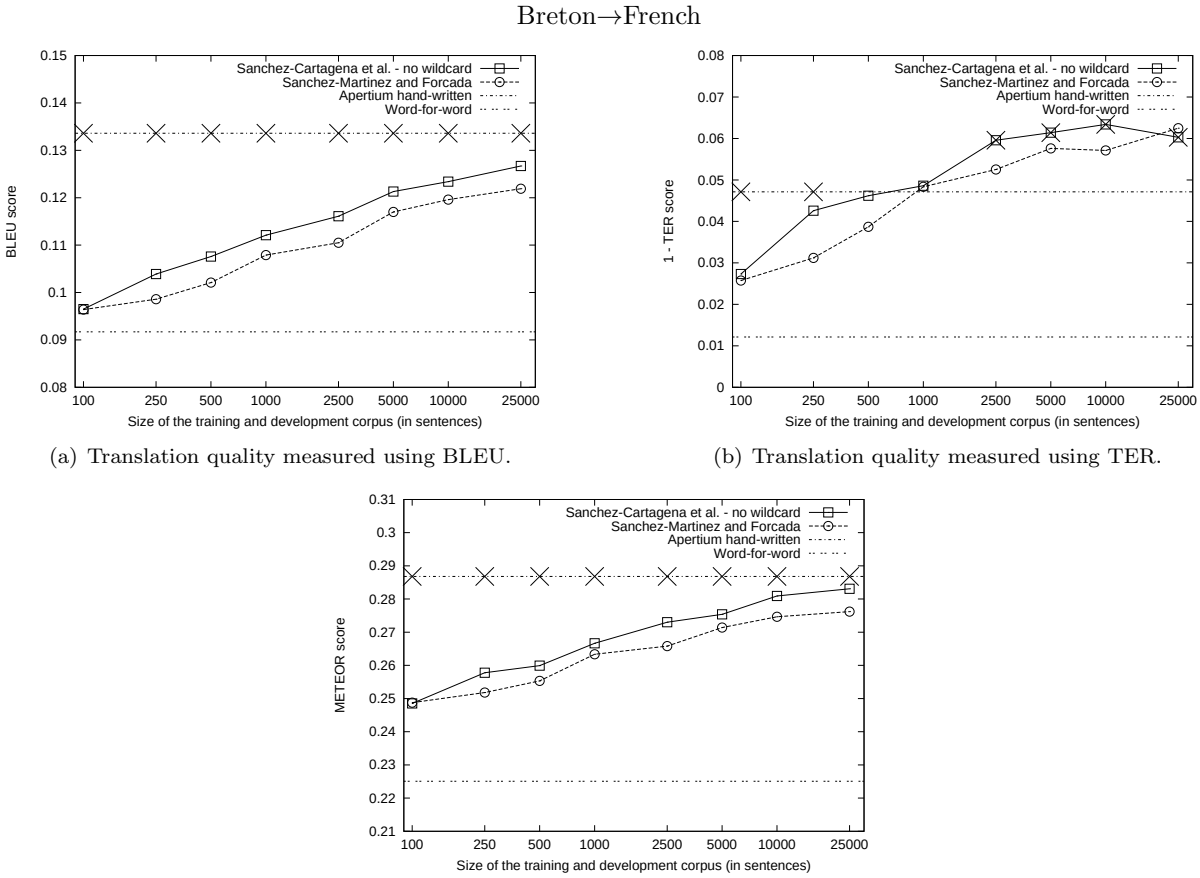


(a) Translation quality measured using BLEU.

(b) Translation quality measured using TER.

(c) Translation quality measured using METEOR.

Figure 22: Translation quality of the different systems evaluated for the Spanish→English language pair with larger corpora subsets than those used in the primary evaluation. A diagonal cross over a square point indicates that our approach outperforms the hand-written rules by a statistically significant margin ($p \leq 0.05$). A diagonal cross over the top horizontal line means that the hand-written rules outperform our approach by a statistically significant margin ($p \leq 0.05$).



(a) Translation quality measured using BLEU.

(b) Translation quality measured using TER.

(c) Translation quality measured using METEOR.

Figure 23: Translation quality of the different systems evaluated for the Breton→French language pair with larger corpora subsets that those used in the primary evaluation. A diagonal cross over a square point indicates that our approach outperforms the hand-written rules by a statistically significant margin ($p \leq 0.05$). A diagonal cross over the top horizontal line means that the hand-written rules outperform our approach by a statistically significant margin ($p \leq 0.05$).

specific hand-written rule is applied. If there is no hand-written rule to apply, the most specific inferred one is used.[40] The results show that the combination does not improve performance. On the contrary, in most cases it causes a degradation of the translation quality originally achieved using the hand-written rules. These results suggest that the linguistic information inferred by our approach has already been encoded by the experts who wrote the rules. It is also worth considering that when an inferred rule matches a segment that is not matched by any hand-written rule and translates it, it may prevent a hand-written rule from being applied afterwards owing to the greedy rule matching mechanism followed by the Apertium engine. Thus, it may be worth considering in the future the application of the method described in Section 4.5 for optimising chunking to the combination of hand-written and inferred rules. Nevertheless, the strategy for rule combination that might be most profitable is the use of our approach to infer a set of rules that are then improved or edited by human experts.

8. Concluding remarks

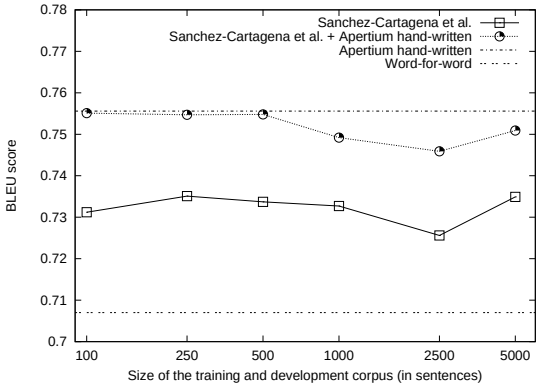
We have described a new alignment-template-based formalism and a language-independent algorithm for the automatic inference of shallow-transfer rules to be used in rule-based MT. This new approach has been evaluated with five different language pairs and with parallel corpora of different sizes. The evaluation performed shows that, in almost all cases and by a statistically significant margin ($p \leq 0.05$), our method outperforms the previous alignment-template-based approach by Sánchez-Martínez and Forcada (2009), which uses a less-expressive formalism and a simpler learning algorithm. In addition, when the languages involved in the translation are closely-related (e.g. Spanish↔Catalan), a few hundred parallel sentence have proved to be sufficient to obtain a set of competitive transfer rules, since the addition of more parallel sentences does not result in great improvements to the translation quality. What is more, this translation quality is close to that obtained with hand-written rules and for some language pairs, our approach is even able to outperform them.

Our approach overcomes many relevant limitations of the previous work, principally those related to the inability to find the appropriate generalisation level for the alignment templates and to select the proper subset of alignment templates which ensures an adequate chunking of the input sentences. Furthermore, the amount of rules inferred by our approach is much smaller than that of the baseline, and this has a positive impact on the possible manual refinement of the resulting rules, since having fewer and more expressive rules eases editing them. In addition, our approach is the first to resolve the conflicts between the inferred rules at a global level by choosing the most appropriate rules according to a global minimisation function rather than by following a pairwise greedy approach. This global minimisation function also allows our method to automatically determine the appropriate level of generalisation of the GATs to be eventually used for rule generation.

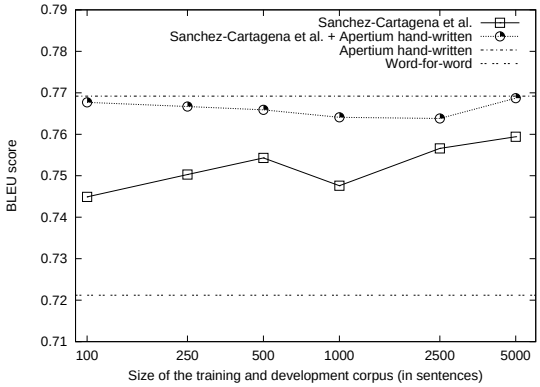
The combinatorial explosion in the generation of GATs with different levels of generalisation and the computational complexity involved in solving the minimisation problem has limited the experiments conducted with our approach to very small parallel corpora, and forced us to introduce some heuristics in order to limit the number of GATs to be considered during the minimisation. It is, however, when the amount of parallel corpora is scarce that our method achieves the greatest improvement when compared to the baseline approach. In addition, disabling the generalisation of morphological inflection attributes with wildcards and reference values has allowed us to scale our approach to bigger corpora and reach, and in some cases surpass, the translation quality of hand-written rules.

In the future, it might be possible to devise a method with which to select the most informative sentences from a monolingual corpus that should be manually translated in order to obtain a parallel corpus for rule inference. A similar scheme has already been proposed for active learning in SMT (Haffari et al., 2009). As regards the optimisation of the thresholds used, the optimum value of δ and θ could be obtained by means of a simplex algorithm (Spendley et al., 1962) rather than trying all the values in the interval $[0, 1]$

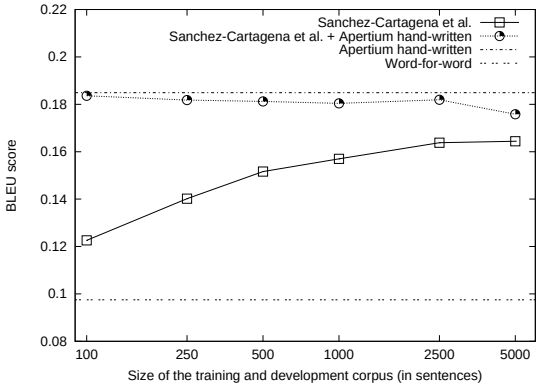
[40] Note that, due to the way in which the hand-written rules are encoded, it is not possible to treat hand-written rules and automatically-inferred ones as a single set and sort them together according to their specificity level.



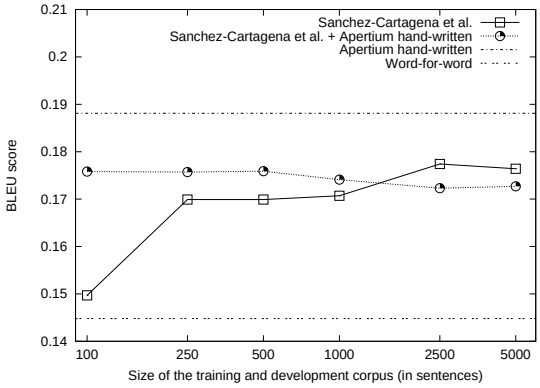
(a) Spanish–Catalan.



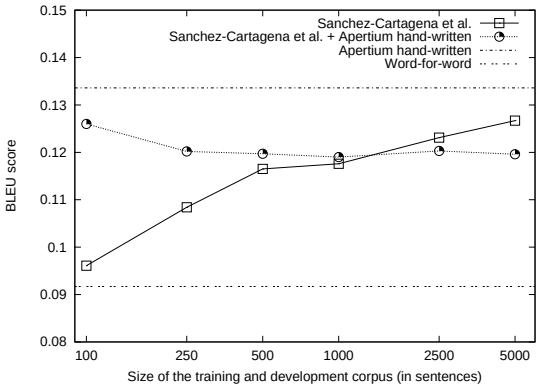
(b) Catalan–Spanish.



(c) English–Spanish.



(d) Spanish–English.



(e) Breton–French.

Figure 24: Translation quality, as measured by BLEU score, of the combination of the rules inferred by our approach with the hand-written rules from the Apertium project. The scores achieved by the hand-written rules alone, the rules obtained by our approach alone, and word-for-word translation are also depicted.

at increments of 0.05 for δ and using θ only to reduce the complexity of the minimisation problem, as we have done. As has already been mentioned, more sophisticated strategies for combining the rules inferred by our approach with hand-written rules could be explored, starting with the improvement of the chunking performed by the combined set of rules.

Alternative approaches could be considered for some of the steps of our rule learning procedure in order to further improve the results obtained. The word alignment quality could be improved by integrating symmetrisation in the training of the alignment models as shown by Liang et al. (2006), who have reported a reduction in the alignment error rate with small parallel corpora. Regarding the optimisation performed to discard rules that cause a deficient chunking of the sentences to be translated, some changes could be made to the evaluation metric used to compute the set of key text segments $\mathcal{I}$; for instance, Nakov et al. (2012) suggest some improvements to the BLEU smoothing, which are well-suited to sentence-level optimisation.

In summary, we have presented a cost-effective approach for the inference of rule-based MT transfer rules that can be applied when monolingual and bilingual dictionaries are available but the amount of available parallel corpora is scarce. Recall that Sánchez-Martínez and Forcada (2009) have already proved that a rule-based MT system with rules inferred from a small parallel corpus outperforms an SMT system trained on the same parallel corpus, even when it is complemented with the entries from the bilingual dictionary of the RBMT system.

We think that the adoption of the method presented in this paper will significantly contribute towards making the development of transfer rules for new language pairs in MT systems like Apertium a much more cost-effective and technically feasible process, thus reducing the total time necessary to deploy working systems.

Acknowledgements

We would like to thank Mikel L. Forcada for his comments on the approach and on an early draft of this paper. Research funded by Universitat d'Alacant through project GRE11-20, by the Spanish Ministry of Economy and Competitiveness through projects TIN2009-14009-C02-01 and TIN2012-32615, by Generalitat Valenciana through grant ACIF/2010/174, and by the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran).

AppendixA. Solving the minimisation problem with integer linear programming

The minimisation problem defined in Section 4.4 can be solved by reformulating it as an integer linear programming problem (Garfinkel and Nemhauser, 1972), which, in turn, allows us to apply existing methods in order to solve it (Xu et al., 2009). An integer linear programming problem involves the optimisation (maximisation or minimisation) of a linear objective function subject to linear inequality constraints, and has the following general form:

- optimise $\sum_{i=1}^{n} c_i x_i$
- subject to m constraints: $\sum_{i=1}^{n} a_{ij} x_i \geq b_j$ with $j = 1, ..., m$
- where $x_i \in \mathbb{Z} \ \ \forall i \in [1, n]$.

Let Z be the set of GATs obtained from the set of bilingual phrase pairs P obtained from the training parallel corpus, and O be the set of GATs we are seeking, i.e., the set with the minimum number of GATs that are needed to correctly reproduce all the bilingual phrases pairs in P. Recall, as stated in Section 4.4, that when there is more than one solution to the minimisation problem, that with the minimum aggregated specificity level is chosen, and that, as a consequence of the filtering of unreliable GATs (see Section 4.3), it may occur that the problem cannot be solved because not all bilingual phrase pairs in P can be reproduced. In this case, only the minimum number of bilingual phrase pairs needed to make the minimisation problem solvable must be removed. We shall refer to the set of bilingual phrase pairs that are reproduced by the set of GAT in O as P_O, obviously $P_O \subseteq P$.

In order to reformulate our minimisation problem using integer linear programming inequations, we define two sets of integer variables: X and Y. The set of integer variables X is associated with the GATs in Z such that $x_i \in X$ equals 1 if the GAT $z_i \in Z$ is part of the solution set O, zero otherwise. The set of integer variables Y is associated with the bilingual phrase pairs $p_j \in P$ so that y_j equals 1 if $p_j \notin P_O$, i.e. if it has been removed to make the minimisation problem solvable.

The function to be minimised (optimised) is defined as:

$$\left(\sum_{i=1}^{|X|} x_i\right) + \left(\sum_{i=1}^{|X|} x_i \cdot \frac{\text{spec\_level}(z_i)}{\max(\{\text{spec\_level}(z_j) : z_j \in Z\})} \cdot \frac{1}{|Z|}\right) + \left(\sum_{j=1}^{|P|} y_j \cdot \text{freq}(p_j) \cdot T\right)$$

where spec_level(z_i) computes the level of specificity of GAT z_i (see the equation on page 19), freq(p_j) is the frequency of the bilingual phrase pair p_j in the parallel corpus, and T is a penalty whose value is set to $|Z| + 2$ (see below).

The first term in the equation above counts the number of GATs in Z that are part of the solution set O. The second term is introduced to discriminate between different solution sets with the same number of GATs. Here $\frac{1}{|Z|}$ is introduced to ensure that the second term only discriminates between different solutions sets with the same number of GATs and that it does not promote solution sets with a large amount of GATs but a low level of specificity.[41]

The third term counts the number of occurrences in the training corpus of the bilingual phrase pairs that need to be discarded to make the minimisation problem solvable. Here a penalty T is introduced to ensure that only the minimum amount of bilingual phrase pairs needed to make the minimisation problem solvable are removed, i.e. not included in P_O; otherwise we could be removing bilingual phrase pairs not because they cannot be reproduced but because by removing them, the GATs reproducing them could also be removed, thus reducing the size of O. The value of T is set to $|Z| + 2$ because it is the lowest possible value which guarantees that this term is greater than the sum of the first and the second terms when deciding whether or not to remove a bilingual phrase that can be correctly reproduced and by removing it the amount of GATs is also reduced.[42]

The inequations representing the constraints to the minimisation problem are as follows:

- $\mathcal{C}_1$: There may exist at least one GAT in O that reproduces each bilingual phrase pair in P_O:

$$\forall_{p \in P} \sum_{i:p_k \in \mathcal{G}(z_i)} x_i + y_k \geq 1$$

  Note that in order to check this constraint we iterate over all bilingual phrase pairs P, not over P_O, and that if a bilingual phrase pair p_k is not in P_O the constraint is met because $y_k = 1$.

- $\mathcal{C}_2$: For each bilingual phrase pair in P_O matched but not correctly reproduced by a GAT z_i, either z_i is not part of the solution or there is at least one more specific GAT that is part of the solution and correctly reproduces it:

$$\forall_{i \in [1,|Z|]} \forall_{p_k \in \mathcal{B}(z_i)} \sum_{j:j \neq i \wedge p \in \mathcal{G}(z_j)} \Lambda_{ji} x_j + y_k \geq x_i$$

  where Λ_{ji} maps the output of the function more_specific (see Section 4.4 on page 18) to values 0 or 1:

$$\Lambda_{ji} = \begin{cases} 1 & \text{if more\_specific}(z_j, z_i) \\ 0 & \text{otherwise.} \end{cases}$$

  As before, if p_k is not part of P_O the constraint is met because $y_k = 1$.

[41] Note that the value of the second term is always in the range $[0, 1]$.

[42] Note that the sum of the first two terms of the expression is always less than or equal to $|Z| + 1$, and the third term is always greater than or equal to T.

```
<section-def-cats>
  <def-cat n="CAT_VERB">
    <cat-item tags="VERB.*"/>
  </def-cat>
  ...
</section-def-cats>
...
<section-rules>
<rule>
  <pattern>
    <pattern-item n="CAT_VERB"/>
    <pattern-item n="CAT_VERB"/>
  </pattern>
  <action>
    ...
  </action>
</rule>
...
</section-rules>
```

Figure B.25: Header (`pattern` section) of an Apertium shallow-transfer rule containing GAT for the translation of the Catalan verb *anar* in the past tense followed by a verb in infinitive mood into Spanish. The `section-def-cats` section is used to define identifiers for the patterns to be matched by the rules.

AppendixB. Implementing the rules in Apertium

The approach presented throughout this paper has been evaluated with the Apertium shallow-transfer RBMT system (Forcada et al., 2011). This appendix describes how the GATs obtained with our approach are converted to the rule format used by Apertium.

Apertium shallow-transfer rules are encoded in XML format.[43] Each rule consists of a `pattern` section and an `action` section. The `pattern` section is used to specify the lexical category, lemma, and morphological inflection attributes of the lexical forms to be matched; lemma and morphological inflection attributes are optional. The instructions working with the matched lexical forms are placed in the `action` section. Apertium provides instructions that permit access to the SL lexical forms matched by the rule and the translation provided for them in the bilingual dictionary. There are also instructions that permit the TL lexical forms to be built, and these result from the application of the rule by assembling the aforementioned elements. Some flow control structures (mainly, loops and conditionals) are also allowed.

The set of GATs obtained are converted into rules by grouping those GATs that match the same sequence of lexical categories under the same rule. Each rule detects the corresponding sequence of lexical categories in its `pattern` section (regardless of the lemma and morphological inflection attributes). GATs are then included in the `action` section in decreasing order of specificity (see Section 4.6) signifying that the most specific GAT is always applied when more than one GAT can be applied to the sequence of lexical categories matched by the rule. For each GAT, the body of the rule checks whether the lemmas, morphological inflection attributes and restrictions of the sequence of SL lexical categories matched by the rule are compatible with the GAT, and if they are then the GAT is applied and the execution of the rule ends. If after checking all the GATs in a rule, none of them can be applied, the engine attempts to apply a shorter rule to the input text.[44] Recall that the rules to be applied are chosen by Apertium in a greedy, left-to-right, longest-match manner.

The following example illustrates how GATs for the translation of a sequence of two verbs from Catalan to Spanish (like that shown in Figure 12 on page 23) are encoded as an Apertium rule. Figure B.25 shows the pattern section that matches a sequence of two verbs. The action section of the rule consists of several

[43] The Document Type Definition is available at `http://apertium.org/dtd/transfer.dtd`.

[44] This behaviour differs from the standard behaviour in Apertium. To implement it, we have modified the Apertium engine in order to add support for the cancellation of the execution of a rule.

```
<action>
  <choose>
    ...
    <when>
      <test><and>
        <equal>
          <clip pos="1" side="sl" part="lemma" />
          <lit v="anar"/>
        </equal>
        <equal>
          <clip pos="1" side="sl" part="tense" />
          <lit-tag v="past"/>
        </equal>
        <equal>
          <clip pos="1" side="tl" part="tense" />
          <lit-tag v="past"/>
        </equal>
        <equal>
          <clip pos="2" side="sl" part="tense" />
          <lit-tag v="inf"/>
        </equal>
        <equal>
          <clip pos="2" side="tl" part="tense" />
          <lit-tag v="inf"/>
        </equal>
      </and></test>
      <out>
         <lu><clip pos="2" side="tl" part="lemma"/><lit-tag v="verb.past"/><clip pos="2"
             side="tl" part="person"/><clip pos="2" side="tl" part="number"/></lu>
      </out>
    </when>
        ...
    <otherwise>
      <reject-current-rule shifting="no" />
    </otherwise>
  </choose>
</action>
```

Figure B.26: Fragment of the `action` section of an Apertium shallow-transfer rule encoding the structural transformation provided by the GAT shown in Figure 12 (see page 23) for the translation of the Catalan verb *anar* in the past tense followed by a verb in infinitive mood into Spanish. The `pattern` section of the rule is shown in Figure B.25.

GATs; Figure B.26 shows the fragment of the `action` section that corresponds to the GAT in Figure 12. The XML tags `choose`, `when`, `test` and `otherwise` work as the `switch` instruction in many programming languages. The first `equal` instruction checks whether the lemma of the first verb is *anar*, the following two instructions ensure that the tense of the first SL verb is *past* and that the result obtained after looking it up in the bilingual dictionary is also in the past tense too (restriction). The two remaining `equal` instructions apply the same verification to the infinitive mood of the second SL verb. If the five tests are passed, one lexical form is generated (defined by the `lu` tag inside the `out` element). Its lemma is obtained by looking up in the bilingual dictionary the second lexical form matched by the rule (first `clip` tag). The lexical categories and the first morphological inflection attribute (verb tense) are explicitly defined with the tag `lit-tag` and the values of the other two morphological inflection attributes (person and number) are obtained using the `clip` tag with the `side` attribute set to "`tl`" (TL references). The `reject-current-rule` instruction discards the rule and attempts to apply other (shorter) rules to the input sequence; it is executed only when none of the GATs in the rule can be applied.

AppendixC. Acronyms and abbreviations

AT	Alignment template
EAT	Extended alignment template
EBMT	Example-based machine translation
GAT	Generalised alignment template
IR	Intermediate representation
MT	Machine translation
RBMT	Rule-based machine translation
SL	Source language
SMT	Statistical machine translation
TL	Target language

References

A. Alcázar. Consumer Corpus: Towards linguistically searchable text. In *Proceedings of BIDE (Bilbao-Deusto) Summer School of Linguistics 2005*, Bilbao, Spain, 2005.

S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005.

O. Bojar, C. Buck, C. Callison-Burch, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia. Findings of the 2013 workshop on statistical machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, 2013.

M. D. Brandt, H. Loftsson, H. Sigurpórsson, and F. M. Tyers. Apertium-IceNLP: a rule-based Icelandic to English machine translation system. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 217–224, 2011.

P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.

R. D. Brown. Adding linguistic knowledge to a lexical example-based translation system. In *Proceedings of the Eighth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pages 22–32, 1999.

M. Carl. Inducing probabilistic invertible translation grammars from aligned texts. In *Proceedings of the 2001 workshop on Computational Natural Language Learning - ConLL '01*, volume 7, pages 1–7, Morristown, NJ, USA, July 2001. Association for Computational Linguistics.

M. Carl and A. Way, editors. *Recent Advances in Example-Based Machine Translation*, volume 21. Springer, 2003.

H. M. Caseli, M. G. V. Nunes, and M. L. Forcada. Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. *Machine Translation*, 20(4):227–245, 2006. Published in 2008.

D. Chiang. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228, Jun 2007. ISSN 0891-2017. doi: 10.1162/coli.2007.33.2.201.

I. Cicekli and H. A. Güvenir. Learning translation templates from bilingual translation examples. *Applied Intelligence*, 15(1): 57–76, 2001.

M. R. Costa-Jussà and M. Farrús. Statistical machine translation enhancements through linguistic levels: A survey. *ACM Comput. Surv.*, 46(3), 2014. doi: 10.1145/2518130.

A. Font-Llitjós. *Automatic Improvement of Machine Translation Systems*. PhD thesis, Carnegie Mellon University, 2007.

M. L. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, G. Ramírez-Sánchez F. Sánchez-Martínez, and F. M. Tyers. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144, 2011. Special Issue: Free/Open-Source Machine Translation.

R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Series of Books in the Mathematical Sciences. W. H. Freeman, 1979.

R. S. Garfinkel and G. L. Nemhauser. *Integer programming*, volume 4. Wiley New York, 1972.

G. Haffari, M. Roy, and A. Sarkar. Active learning for statistical phrase-based machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 415–423, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-41-1.

W. J. Hutchins and H. L. Somers. *An introduction to machine translation*, volume 362. Academic Press New York, 1992.

A. B. Kahn. Topological sorting of large networks. *Communications of the ACM*, 5(11):558–562, 1962.

H. Kaji, Y. Kida, and Y. Morimoto. Learning translation templates from bilingual text. In *Proceedings of the 14th Conference on Computational Linguistics*, pages 672–678. Association for Computational Linguistics, 1992.

P. Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, volume 4, pages 388–395, 2004a.

P. Koehn. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In R. E. Frederking and K. B. Taylor, editors, *Machine Translation: From Real Users to Research*, volume 3265 of *Lecture Notes in Computer Science*, pages 115–124. Springer Berlin Heidelberg, 2004b. ISBN 978-3-540-23300-8. doi: 10.1007/978-3-540-30194-3_13.

P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, pages 12–16, Phuket, Thailand, September 2005.

P. Koehn. *Statistical Machine Translation*. Cambridge University Press, 2010.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, 2007.

B. Korte and J. Vygen. *Combinatorial Optimization: Theory and Algorithms*. Springer, 5th edition, 2012.

P. Liang, B. Taskar, and D. Klein. Alignment by agreement. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 104–111, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. doi: 10.3115/1220835.1220849.

Y. Liu and C. Zong. The technical analysis on translation templates. In *Proceedings of the IEEE International Conference on Systems, Man & Cybernetics (SMC)*, pages 4799–4803. IEEE, 2004. ISBN 0-7803-8566-7.

A. Menezes and S. D. Richardson. *Recent Advances in Example-Based Machine Translation*, chapter A best-rirst alignment algorithm for automatic extraction of transfer mappings from bilingua corpora, pages 421–442. Springer, 2003.

P. Nakov, F. Guzmán, and S. Vogel. Optimizing for Sentence-Level BLEU+1 Yields Short Translations. In Martin Kay and Christian Boitet, editors, *COLING*, pages 1979–1994. Indian Institute of Technology Bombay, 2012.

F. J. Och. *Statistical machine translation: from single-word models to alignment templates*. PhD thesis, RWTH Aachen University, 2002.

F. J. Och. Statistical machine translation: Foundations and recent advances. Tutorial at MT Summit X, 2005. (`http://www.mt-archive.info/MTS-2005-Och.pdf`).

F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1): 19–51, 2003.

F. J. Och and H. Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4): 417–449, 2004.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318, 2002. doi: http://dx.doi.org/10.3115/1073083.1073135.

K. Probst. *Automatically Induced Syntactic Transfer Rules for Machine Translation under a Very Limited Data Scenario*. PhD thesis, Carnegie Mellon University, 2005.

K. Probst, L. Levin, E. Peterson, A. Lavie, and J. Carbonell. MT for minority languages using elicitation-based learning of syntactic transfer rules. *Machine Translation*, 17(4):245–270, 2002. ISSN 0922-6567.

F. Sánchez-Martínez and M. L. Forcada. Inferring shallow-transfer machine translation rules from small parallel corpora. *Journal of Artificial Intelligence Research*, 34(1):605–635, 2009. ISSN 1076-9757.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231, 2006.

W. Spendley, G. R. Hext, and F. R. Himsworth. Sequential application of simplex designs in optimisation and evolutionary operation. *Technometrics*, 4(4):441–461, 1962.

G. Thurmair. Comparing different architectures of hybrid Machine Translation systems. In *Proceedings MT Summit XII*, 2009.

A. Toral, M. Ginestí-Rosell, and F. M. Tyers. An Italian to Catalan RBMT system reusing data from existing language pairs. In *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 77–81, 2011.

F. M. Tyers. Rule-based augmentation of training data in Breton–French statistical machine translation. In *Proceedings of the 13th Annual Conference of the European Association of Machine Translation*, pages 213–217, 2009.

F. M. Tyers. Rule-based Breton to French machine translation. In *Proceedings of the 14th Annual Conference of the European Association of Machine Translation*, pages 174–181, 2010.

F. M. Tyers. *Feasible lexical selection for rule-based machine translation*. PhD thesis, Universitat d'Alacant, Alacant, Spain, July 2013.

F. M. Tyers, F. Sánchez-Martínez, and M. L. Forcada. Flexible finite-state lexical selection for rule-based machine translation. In *Proceedings of the 17th Annual Conference of the European Association of Machine Translation*, pages 213–220, 2012.

S. Vogel, H. Ney, and C. Tillmann. HMM-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics - Volume 2*, COLING '96, pages 836–841, Copenhagen, Denmark, 1996. Association for Computational Linguistics. doi: 10.3115/993268.993313.

Y. Xu, T. K. Ralphs, L. Ladányi, and M. J. Saltzman. Computational experience with a software framework for parallel integer programming. *INFORMS Journal on Computing*, 21(3):383–397, 2009.

A. Zollmann and S. Vogel. A word-class approach to labeling pscfg rules for machine translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1–11, Portland, Oregon, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9.

Vitae

Víctor M. Sánchez-Cartagena is a Ph.D. candidate in the Departament de Llenguatges i Sistemes Informàtics at Universitat d'Alacant in Spain. At the same time, he works as a research engineer for Prompsit Language Engineering, a company specialised in the creation of customised machine translation solutions. He received his Master's Degree in Computer Engineering Technologies in 2010. His main research interests are the bootstrapping of machine translation systems from scarce resources and the hybridisation between the different MT approaches.

Juan Antonio Pérez-Ortiz is an associate professor at Universitat d'Alacant in Spain. He received his Ph.D. in computer science in 2002. He has worked on neural networks applied to sequence prediction and, for the last ten years, on machine translation and computer-assisted translation, especially as a member of the team involved in the development of the open-source machine translation platform Apertium. He currently teaches undergraduate and postgraduate courses on machine translation, programming and web development.

Felipe Sánchez-Martínez is lecturer at Universitat d'Alacant (Spain) and member of the European Association for Machine Translation; he received his Ph.D. in computer science in 2008. His main field of research is machine translation and the application of unsupervised corpus-based methods to build some of the modules and linguistic resources needed by rule-based machine translation systems. He also works on the integration of machine translation and computer-aided translation tools based on translation memories. He is part of the team that is responsible for the design, development and maintenance of the Apertium shallow-transfer machine translation platform. Most of his undergraduate and graduate teaching involves translation and language technologies.