Journal of Artificial Intelligence Research 55 (2016) 17-61 Submitted 03/15; published 01/16

Integrating Rules and Dictionaries from Shallow-Transfer Machine Translation into Phrase-Based Statistical Machine Translation

Víctor M. Sánchez-Cartagena VMSANCHEZ@DLSI.UA.ES
Prompsit Language Engineering
Av. Universitat s/n. Edifici Quorum III
E-03202 Elx, Spain

Juan Antonio Pérez-Ortiz JAPEREZ@DLSI.UA.ES
Felipe Sánchez-Martínez FSANCHEZ@DLSI.UA.ES
Dep. de Llenguatges i Sistemes Informàtics
Universitat d'Alacant
E-03071, Alacant, Spain

Abstract

We describe a hybridisation strategy whose objective is to integrate linguistic resources from shallow-transfer rule-based machine translation (RBMT) into phrase-based statistical machine translation (PBSMT). It basically consists of enriching the phrase table of a PBSMT system with bilingual phrase pairs matching transfer rules and dictionary entries from a shallow-transfer RBMT system. This new strategy takes advantage of how the linguistic resources are used by the RBMT system to segment the source-language sentences to be translated, and overcomes the limitations of existing hybrid approaches that treat the RBMT systems as a black box. Experimental results confirm that our approach delivers translations of higher quality than existing ones, and that it is specially useful when the parallel corpus available for training the SMT system is small or when translating out-of-domain texts that are well covered by the RBMT dictionaries. A combination of this approach with a recently proposed unsupervised shallow-transfer rule inference algorithm results in a significantly greater translation quality than that of a baseline PBSMT; in this case, the only hand-crafted resource used are the dictionaries commonly used in RBMT. Moreover, the translation quality achieved by the hybrid system built with automatically inferred rules is similar to that obtained by those built with hand-crafted rules.

1. Introduction

Statistical machine translation (SMT) (Koehn, 2010) is currently the leading paradigm in machine translation (MT) research. SMT systems are very attractive because they may be built with little human effort when enough monolingual and parallel corpora are available. However, parallel corpora are not always easy to harvest, and they may not even exist for some (under-resourced) language pairs. On the contrary, rule-based machine translation (RBMT) systems (Hutchins & Somers, 1992) may be built without any parallel corpus; however, they need an explicit representation of linguistic information, whose coding by human experts requires a considerable amount of time.

Even when a large parallel corpus is available, SMT systems may still have some limitations as a result of (i) the *data sparseness* problem that makes it difficult to collect enough

phrase pairs covering all the inflected word forms in highly inflected languages, and (ii) the *domain* problem caused when the training parallel corpus belongs to a domain different from that of the texts to be translated. One potential solution is to follow a *hybrid* approach (Thurmair, 2009) and combine an RBMT system with the SMT system in order to mitigate these limitations. This is the approach we follow in this paper, in which linguistic resources from shallow-transfer RBMT are used to enrich the phrase table of a phrase-based SMT (PBSMT) system.

Like any other transfer-based RBMT system, shallow-transfer RBMT systems carry out the translation process in three steps: analysis of the source-language (SL) sentence to produce an SL intermediate representation (IR), transfer from that SL IR to a target-language (TL) IR, and generation of the final translation from the TL IR. Shallow-transfer RBMT systems do not perform a complete syntactic analysis of the input sentences and work with simple IRs consisting of a sequence of *lexical forms*. A lexical form comprises the lemma, lexical category and morphological inflection information of a word.

In shallow-transfer RBMT, as in the Apertium system (Forcada et al., 2011) used in our experiments, after the analysis step, the SL sentence is split into *chunks*. Each chunk is then translated by a shallow-transfer rule and their translations are concatenated in order to build the TL sentence. This process is similar to the process carried out by a PBSMT decoder, which builds translation hypotheses by segmenting the SL sentence into phrases and translating each SL phrase according to the phrase table. As both systems work with flat sub-segments it is easy to integrate chunks from RBMT into the SMT phrase table so that they can be scored by all the feature functions commonly used in PBSMT. Moreover, the use of RBMT dictionaries and shallow-transfer rules allows the PBSMT decoder to choose phrase pairs that go beyond the word-for-word translation of the words in the RBMT dictionaries, as well as translating all the inflected word forms they contain; thus alleviating the data sparseness problem. In addition, the data from a general-purpose RBMT system can help to reduce the bias of an SMT system towards the domain of the training corpus.

Additionally, even if the rules from the RBMT system have not yet been created, they can be automatically inferred from a small fragment of the training parallel corpus by means of the (unsupervised) rule inference approach proposed by Sánchez-Cartagena, Pérez-Ortiz, and Sánchez-Martínez (2015). A better use is therefore made of the training parallel corpus and RBMT dictionaries than in existing approaches (Schwenk, Abdul-Rauf, Barrault, & Senellart, 2009) that simply add the dictionaries to the phrase table. By combining the rule inference algorithm with our hybridisation approach, the translation knowledge contained in the parallel corpus is generalised to sequences of words that have not been observed in the corpus, but share lexical category or morphological inflection information with the words observed.

The enrichment of PBSMT models with RBMT linguistic data has already been explored by other authors (see Section 2.1); however, the approach presented in this paper is the first one specifically designed for use with shallow-transfer RBMT and that takes advantage of the way in which the linguistic resources are used by the RBMT system. To the best of our knowledge, the general approach by Eisele et al. (2008), described in Section 2.1, is the only hybrid approach in literature that can be applied to shallow-transfer RBMT systems.

The experimental results show that our hybrid approach outperforms the strategy developed by Eisele et al. (2008). Moreover, the performance of the hybrid system built using

automatically inferred rules is on a par with the hybrid system built with hand-crafted rules. It is also worth pointing out that a system (Sánchez-Cartagena, Sánchez-Martínez, & Pérez-Ortiz, 2011b) built with our approach and using hand-crafted rules from the Apertium project (Forcada et al., 2011) was one of the winners[1] in the pairwise manual evaluation of the WMT 2011 shared translation task (Callison-Burch, Koehn, Monz, & Zaidan, 2011) for the Spanish→English language pair. The hybridisation approach presented in this paper, together with the aforementioned rule inference algorithm, will contribute to alleviating the data sparseness problem that SMT systems have when highly inflected languages are involved and reducing the corpus size requirements as regards building PBSMT systems.

The remainder of the paper is organised as follows. Section 2 reviews related work on hybrid machine translation, including a description of the limitations of the general hybridisation approach proposed by Eisele et al. (2008). Section 3 describes our hybridisation strategy and a set of different alternatives for scoring the phrase pairs generated from the linguistic resources of the RBMT system. Two different sets of experiments, all of which integrate data from the Apertium RBMT platform (Forcada et al., 2011), are then described in order to evaluate our hybridisation strategy (Section 4) and assess whether the automatically inferred rules can replace hand-crafted ones in the hybrid system (Section 5). The paper ends with a human evaluation and an error analysis (Section 6) and some concluding remarks (Section 7).

2. Related Work

Hybrid approaches related to that presented in this paper can be split into those that integrate RBMT elements into an SMT system (sections 2.1 and 2.2) and those that integrate SMT elements into the RBMT architecture (Section 2.3).[2] Approaches in the first group can in turn be split into two groups: those that use linguistic information from an existing RBMT system (Section 2.1) and those that use linguistic resources inferred from the parallel corpus from which the SMT models are estimated (Section 2.2).

2.1 Integrating Hand-Crafted Linguistic Resources in SMT

Bilingual dictionaries are the most frequently reused resource from RBMT; they have been added to SMT systems since its early days (Brown et al., 1993). One of the simplest strategies, which has already been put into practice with the Apertium bilingual dictionaries (Tyers, 2009), consists of adding the dictionary entries directly to the training parallel corpus. In addition to the obvious increase in lexical coverage, Schwenk et al. (2009) state that the quality of the alignments obtained is also improved when the words in the bilingual dictionary appear in other sentences of the parallel corpus. However, it is not guaranteed that, following this strategy, multi-word expressions from the bilingual dictionary that ap-

1. No other system was found to be statistically significantly better when using the sign test at $p \leq 0.10$.
2. This is not meant to be a strict classification: some of the approaches listed in this section could be included in both groups. Moreover, approaches in which the outputs of different MT systems are just combined without making any modification into the inner workings of the systems involved, such as system combination (Rosti, Matsoukas, & Schwartz, 2007) are not listed in this review because, unlike our approach, they do not involve the creation of new MT architectures that combine elements from SMT and RBMT.

pear in the SL sentences are translated as such because they may be split into smaller units by the phrase-extraction algorithm. Dictionaries have also been added to SMT systems together with other rule-based enhancements, as in the work by Popovic and Ney (2006), who propose combining dictionaries with the use of hand-crafted rules in order to reorder the SL sentences to match the structure of the TL.

Other approaches take advantage of the full RBMT system. Eisele et al. (2008) present a strategy based on the augmentation of the phrase table to include information provided by an RBMT system. Their approach treats the RBMT system as a black box, i.e., the algorithm is not concerned with the inner workings of the RBMT system. The sentences to be translated by the hybrid system are first translated with the RBMT system and a small phrase table is obtained from the resulting parallel corpus (from now on, *synthetic corpus*). This new phrase table is then directly added to the original phrase table obtained from the training parallel corpus. This approach has the following limitations, which are overcome by the hybrid approach described in this paper:

Deficient segment alignment. When phrase pairs are extracted from the synthetic corpus through the usual procedure followed in PBSMT (Koehn, 2010, §5.2.3), unaligned words are included in multiple phrase pairs, since there is no evidence about their correspondence in the other language, and phrase pairs made solely of unaligned words are not extracted. If word alignments are incorrect, phrase pairs that are not mutual translation may be extracted and correct phrase pairs present in the parallel sentences may not be obtained.[3] The less reliable the word alignments are, the more severe this problem becomes.

The word alignment of the synthetic corpus obtained by Eisele et al. (2008) may be unreliable owing to a vocabulary mismatch between the text to be translated and the alignment models, which are inferred from the training corpus.[4] This limitation becomes more evident when the text to be translated does not share the domain with the training corpus, which is actually when the data from the RBMT system is more useful.[5]

Relying on word alignments is a reasonable strategy when extracting phrase pairs from a parallel corpus when we do not know how it was built. However, when we know that an RBMT system has been used to generate the TL side of the corpus, a

3. Consider the following segment of an English–Spanish parallel sentence: *Barcelona City Council – Ayuntamiento de Barcelona*. If the only word alignment between these segments were a link between *Barcelona* in both languages, incorrect phrase pairs such as *Barcelona City Council – Barcelona* would be extracted, whereas the correct phrase pair *City Council – Ayuntamiento* would not be extracted.
4. Alignment models do not contain information about words in the test corpus that are not present in the training corpus, these words are not therefore aligned and it is likely that phrase pairs that are not mutual translation will be extracted from them.
5. This problem could be alleviated by building alignment models from the concatenation of the synthetic corpus and the training corpus, or by incrementally training (Gao, Lewis, Quirk, & Hwang, 2011) the word alignment models. The former would be computationally too expensive, since the process would have to be carried out each time a new text was translated with the resulting hybrid system (e.g. building word alignment models from the English→Spanish parallel corpus with 600 000 sentences described in Section 4.1 took around 6 hours in an AMD Opteron 2 Ghz processor). The latter is likely to cause alignment errors when infrequent words in the synthetic corpus not found in the training corpus are involved.

more precise phrase extraction mechanism that takes advantage of how the RBMT system uses dictionaries and shallow-transfer rules to segment the SL sentences can be used.

Inadequate balance between the different types of phrase pairs obtained. The probabilities derived by Eisele et al. (2008) for the phrase pairs extracted from the synthetic corpus and added to the phrase table are not consistent because they have been independently estimated from two different corpora. On the one hand, if an SL phrase is translated in the same way in the training corpus and by the RBMT system, the probability of the corresponding phrase pair is not increased in comparison with that of other phrase pairs in which that same SL phrase is translated in a different way. On the other hand, when the translations of an SL phrase differ from those produced by the RBMT system, its frequency in the training corpus is not taken into account when scoring the corresponding phrase pairs, and noise may be consequently introduced in the case of SL phrases with a low frequency in the training corpus. For instance, a phrase pair extracted from the training corpus whose SL phrase appears only once is less reliable and should receive a lower score than a phrase pair whose SL phrase appears 10 000 times (see Section 3.2). We overcome this limitation by following a more sophisticated scoring scheme that joins synthetic phrase pairs and phrase pairs obtained from the training corpus in a single list before computing the phrase translation probabilities (see Section 3.2.3).

Another interesting approach is that of Enache, España-Bonet, Ranta, and Màrquez (2012), in which an interlingua RBMT system developed for the limited domain of patent translation is integrated into a PBSMT architecture by generating synthetic phrase pairs from chunks extracted from the SL sentences that can be parsed by the RBMT system.[6] The same philosophy is behind our hybrid approach in which synthetic phrase pairs are generated from the chunks matched by shallow-transfer rules. However, significant differences exist in the method used to score the phrase pairs generated from the RBMT system. Enache et al. use a pre-defined single value for the source-to-target and target-to-source phrase translation probabilities and lexical weightings of the synthetic phrase pairs. As a consequence all the synthetic phrase pairs are equiprobable and their relative weight (compared to the phrase pairs extracted from the training parallel corpus) is not optimised in the tuning step of the SMT training process. In our proposal, however, the relative weight of the synthetic phrase pairs is optimised during the tuning process thanks to the use of a binary feature function, phrases translated in the same way in the parallel corpus and by the RBMT system receive higher scores, and the lexical translation probabilities of the synthetic phrase pairs are computed based on the same principles as in SMT: taking into account the translations of the individual words that make up the phrases.

Finally, Rosa, Mareček, and Dušek (2012) create a set of rules that are applied to the output of an SMT system in order to fix its most common errors. The main difference between their proposal and ours lies in the fact that, although these rules are similar to transfer rules, they operate only on the TL side, and that a syntactic analysis is performed before applying them.

6. A parse tree may not be obtained from the sentences that do not follow the usual structure in the restricted domain. This occurs in the case of 66.7% of the sentences in their test set.

2.2 Adding Morphological Information to SMT

Our hybridisation approach can be combined with the rule inference approach described by Sánchez-Cartagena et al. (2015) in order to integrate a set of structural transfer rules inferred from the SMT training parallel corpus, thereby extending the PBSMT models with new linguistic information. Since shallow-transfer rules operate on lexical forms made of lemma, lexical category and morphological inflection information, the combination of the two approaches can be seen as a novel way of extending PBSMT with morphological features.

In this manner, the resulting approach is related to factored translation models (Koehn & Hoang, 2007), which are an extension of PBSMT in which each word is replaced by a set of factors that represent lemma, lexical category, and morphological inflection information. A phrase-based translation model is inferred for lemmas and an independent one for lexical categories and morphology. A word-based generation model, which can be inferred from additional monolingual data, maps combinations of lemmas, lexical category and morphological inflection information to inflected word forms. The main differences between the factored models and our hybrid approach are as follows:

- In factored models, the translation of lemmas and morphological information is completely independent. As both types of translations are combined in order to generate the final sequence of *surface forms* (running words), a combinatorial explosion is likely to be produced (too many combinations of lemmas and morphological information need to be scored). As all the combinations cannot be explored, correct translation hypotheses may be pruned (Bojar & Hajič, 2008; Graham & van Genabith, 2010). Moreover, idiomatic translations that do not follow the general morphological rules of the TL may be assigned a very low probability by the translation model, even though they would have a high probability in a phrase table built from surface forms. This strategy differs from the one we have followed for the combination of the two approaches, in which translation hypotheses are built from surface-form-based models (like those usually used in PBSMT) that have been enriched with synthetic phrase pairs generated from rules inferred from the training corpus. The complexity of dealing with translations of lemmas and morphological inflection information is moved from decoding to training time, when the rule inference algorithm deals with it.[7]

- Our hybrid approach works with existing bilingual dictionaries, while factored models do not use bilingual dictionaries at all. As a consequence, they translate the morphological inflection information in a different way. In factored models the probability of the TL morphological inflection factors depends solely on the morphological inflection factors of the SL sentence. In contrast, the transfer rules used by our method obtain the morphological inflection attributes of the TL words either from SL words or from their translation according to the bilingual dictionary. This makes the formalism more expressive and eases the treatment of certain linguistic phenomena. Consider, for instance, the case in which there is a morphological inflection attribute that only

7. It is worth noting that Graham and van Genabith (2010) proposed a strategy for partially mitigating the issues caused by the fact that factored models treat lemmas and morphological information as totally independent elements: the extraction from the training parallel corpus of factored templates, which are phrases that will not be decomposed in lemma and morphological information for translation.

exists in the TL (such as gender when translating English into Spanish or French). In our hybrid approach, the structural transfer rule for gender and number agreement between a noun and an adjective would assign the gender of the translation into the TL according to the bilingual dictionary of the SL noun to the TL noun and adjective. This type of rule can be inferred from a very small parallel corpus. In factored models, however, the translation model would presumably assign similar probabilities to TL noun-adjective sequences with both genders, and the success of the agreement would depend solely on the ability of the TL model to differentiate between them.

Other relevant approaches in which morphological attributes are integrated into the translation model of an SMT system can be found in literature. Green and DeNero (2012) define a new feature function that models morpho-syntactic agreements, while the factored language models (Kirchhoff & Yang, 2005) assign probabilities to TL sentences depending on their sequences of word forms and morphological features, among other factors. These approaches differ from the strategy presented in this paper mainly in that they do not perform a generalisation that enriches the translation model with translations of sequences of SL words unseen in the training corpus.

Riezler and Maxwell III (2006) went further and also added syntactic information to SMT. They developed a hybrid RBMT-SMT system which works as follows. The SL sentence is parsed with a lexical functional grammar (Riezler et al., 2002) to obtain an SL intermediate representation (IR). Then the SL IR is transferred into the TL IR by applying a set of probabilistic rules obtained from a parallel corpus. Each rule contains a set of scores inspired by those present in the phrase table of a PBSMT system. Finally, the TL sentence is generated from the TL IR. Since an SL sentence can be parsed in many different ways and many different TL IRs can be generated by applying different rules, a TL model is also used in addition to the aforementioned phrase-table-like features. All these features are finally combined by means of a log-linear model, and their weights are optimised by means of minimum error rate training (Och, 2003) as in SMT. The results show that the grammar used was not able to completely parse half of the sentences of the test set (partial parse trees were obtained instead, but the resulting translation was much worse than the translation of fully parsed sentences), and considering only the sentences that could be fully parsed, there was no statistically significant improvement over a PBSMT system trained using the same data. However, a human evaluation showed an improvement of the grammaticality of the translations. The main differences between this proposal and ours are the following: first, the approach by Riezler and Maxwell III does not use existing bilingual dictionaries; and second, it uses syntactic information that allows the system to perform a deeper linguistic analysis at the expense of not being able to fully parse some input sentences, which results in a drop in translation performance. In contrast, our approach works with lexical categories and morphological inflection information and is more robust to ungrammatical input.

2.3 Integrating Statistical Elements in RBMT

Regarding the enhancement of RBMT systems with statistical elements, it is worth noting that RBMT systems often use statistical methods for part-of-speech tagging (Cutting, Kupiec, Pedersen, & Sibun, 1992) and parsing (Federmann & Hunsicker, 2011). Besides these

components, other elements from SMT have been integrated into RBMT, causing greater changes in the RBMT architecture. For instance, multiple hypotheses can be generated in the transfer step, and the most probable one can then be chosen according to a TL model (Lavie, 2008; Carl, 2007). Another option is to use phrase pairs instead of transfer rules in the transfer step, but keep on using the RBMT analysis and generation modules (Crego, 2014). The approach by Riezler and Maxwell III (2006), discussed previously, also uses a TL model in order to choose among translations generated by applying rules, but it integrates more elements from SMT, such as the feature functions usually encoded in an SMT phrase table.

A different alternative consists of taking advantage of the full syntactic analysis performed by syntactic-transfer RBMT systems to create the structure of the TL sentence, and then insert phrase translations from a PBSMT phrase table in some nodes of the TL parse tree (Labaka, España-Bonet, Màrquez, & Sarasola, 2014). As in SMT, the final translation is that with the maximum probability according to a TL model and to the scores in the phrase table from which the phrases inserted in the tree have been obtained. However, phrase reordering is not allowed, since the structure of the TL sentences is guided by the parse tree. This set-up has also been followed in systems proposed by other authors (Federmann et al., 2010; Zbib et al., 2012).

3. Enhancement of Phrase-Based SMT with Shallow-Transfer Linguistic Resources

If we have access to the inner working of the RBMT system, the correspondence between the SL segments of the input sentence and their translations can be computed without relying on statistical word alignments. In fact, it is not even necessary to translate the whole sentence with the RBMT system. The individual translation according to the bilingual dictionary of each word, and the translation of each segment that matches a shallow-transfer rule constitute the minimum set of bilingual phrases that ensures that all the linguistic information from the RBMT system has been extracted. Another advantage of this method over the approach by Eisele et al. (2008) lies in the fact that rules that match an SL segment but would not be applied by the shallow-transfer RBMT system because of its greedy operating mode are also taken into account.[8] Thus, our hybrid strategy first generates these synthetic phrase pairs from the RBMT linguistic data and the SL text to be translated, and then integrates them into the PBSMT models without further decomposition.

8. Consider, for instance, the English sentence *I visited Bob and Alice's dog was sleeping* to be translated into Spanish with a shallow-transfer RBMT system. Let us suppose that the following segments of the sentence match a shallow-transfer rule: *I visited* matches a rule that removes the personal pronoun (it can be omitted in Spanish), adds the corresponding preposition and generates *visité a*; *Bob and Alice's dog* matches a rule that processes the Saxon genitive, adds the preposition and determiner needed in Spanish and generates *el perro de Bob y Alice*; and *Alice's dog* also matches a rule that processes the Saxon genitive when the noun phrase acting as owner contains a single proper noun, and generates *el perro de Alice*. When the RBMT engine chooses the rules to be applied in a left-to-right, longest match manner, it produces *visité al perro de Bob y Alice estaba durmiendo*, which means *I visited Bob's dog and Alice was sleeping*. The right translation, *visité a Bob y el perro de Alice estaba durmiendo*, can be obtained if the rule that matches *Alice's dog* is applied. If Eisele et al.'s (2008) method is applied to build a hybrid system, the phrase pairs from the correct translation *I visited Bob – visité a Bob* and *Alice's dog was sleeping – el perro de Alice estaba durmiendo* would not be extracted.

3.1 Generation of Synthetic Phrase Pairs

The way in which the synthetic phrase pairs are generated differs depending on which linguistic resources —bilingual dictionaries or shallow-transfer rules— are used. To generate bilingual phrase pairs from the bilingual dictionary, all the SL surface forms recognised by the shallow-transfer RBMT system and their corresponding SL IRs are listed; then, these SL IRs are translated with the bilingual dictionary to obtain their corresponding TL IRs; finally, the corresponding TL word forms are obtained by means of the RBMT generation module.[9] For instance, for the generation of phrase pairs from the English→Spanish bilingual dictionary in the Apertium RBMT platform, mappings between SL surface forms and lexical forms such as *houses* – *house* `N-num:pl` and *however* – *however* `ADV` are generated. They are then translated into the TL by the bilingual dictionary: the resulting phrase pairs are *houses* – *casas* and *however* – *sin embargo*. Since dictionaries may contain multi-word units, the phrase pairs generated may contain more than one word on both (SL and TL) sides. Note that, unlike in the method by Eisele et al. (2008), the sentences to be translated are not used. Thus, the generation of phrase pairs from the bilingual dictionary only needs to be performed once rather than each time a new text is to be translated.

Bilingual phrase pairs matching structural transfer rules are generated in a similar way, but using the SL text to be translated. Thus, this process is repeated each time a new text is to be translated with the hybrid system.[10] First, its SL sentences are analysed in order to obtain their SL IRs, and then the sequences of lexical forms that match a structural transfer rule are passed through the rest of the RBMT pipeline to obtain their translations. If a sequence of SL lexical forms is matched by more than one structural transfer rule, they are used to generate as many bilingual phrase pairs as the different rules it matches. This differs from the way in which Apertium translates, since in these cases only the longest rule would be applied.

Let us suppose the English sentence *My little dogs run fast* to be translated into Spanish. It is analysed by Apertium as the following sequence of lexical forms: *my* `POSP-p:1.num:pl`, *little* `ADJ`, *dog* `N-num:pl`, *run* `VERB-t:inf`, *fast* `ADV`.[11] If the RBMT system only contained two rules, one that performs the swapping and number and gender agreement between an adjective and the noun after it, and another that matches a determiner followed by an adjective and a noun, swaps the adjective and the noun and makes the three words to agree in gender and number, the segments *little* `ADJ` *dog* `N-num:pl` and *my* `POSP-p:1.num:pl` *little* `ADJ` *dog* `N-num:pl` would be used to generate the following bilingual phrase pairs: *little dogs* – *perros pequeños* and *my little dogs* – *mis perros pequeños*.

9. If the TL IR contains missing values for morphological inflection attributes, a different TL phrase for each possible value of the attribute is generated. For instance, from the mapping between the SL (English) word form *beautiful* and the SL lexical form *beautiful* `ADJ-num:sg` two English→Spanish phrase pairs are generated: *beautiful* – *bonito* and *beautiful* – *bonita*; in the first phrase the adjective *beautiful* has been translated as masculine, whereas in the second case it has been translated as feminine.
10. This step can be carried out without dramatically reducing decoding efficiency thanks to the fact that many steps of the Apertium translation pipeline are implemented with partial finite-state transducers (Roche & Schabes, 1997) and are able to process tens of thousands of words per second in an average desktop computer (Forcada et al., 2011, §4.1).
11. The meaning of the abbreviations used to represent lexical categories are: *POSP* = possessive pronoun; *ADJ* = adjective; *N* = common noun; *VERB* = verb; and *ADV* = adverb. Regarding morphological inflection information, `p:1` = *first person*, `num:pl` = *plural number* and `t:inf` = *infinitive mood*.

Note that, unlike the generation of bilingual phrases from the bilingual dictionary, the generation of bilingual phrase pairs from the shallow-transfer rules is guided by the text to be translated.[12] We decided to do this in order to make the approach computationally feasible and avoid meaningless phrases. Consider, for instance, the rule which is triggered by a determiner followed by an adjective and a noun in English. Generating all the possible phrase pairs virtually matching this rule would involve combining all the determiners in the dictionary with all the adjectives and all the nouns, causing the generation of many meaningless phrases, such as *the wireless boy – el niño inalámbrico*.

All the phrase pairs generated are assigned a frequency of 1, since they have not been generated from an actual parallel corpus. These frequencies are used to score the phrase pairs, as described in the next section.

3.2 Scoring the Synthetic Phrase Pairs

PBSMT systems usually attach 4 scores (Koehn, 2010, Sec. 5.3) to every phrase pair in the phrase table (translation model): source-to-target and target-to-source phrase translation probabilities and source-to-target and target-to-source lexical weightings. The source-to-target translation probability $\phi(t|s)$ of a phrase pair (s, t) is usually computed by means of Eq. (1), where count$(\cdot)$ stands for the frequency of a phrase pair in the list of phrase pairs extracted from the training parallel corpus.

$$\phi(t|s) = \frac{\text{count}(s,t)}{\sum_{t_i} \text{count}(s,t_i)} \tag{1}$$

The purpose of lexical weightings is to act as a back-off when scoring phrase pairs with a low frequency (Koehn, 2010, Sec. 5.3.3). The lexical weighting score of a phrase pair is usually computed as the product of the lexical translation probability of each source word and the target word to which it is aligned. Lexical translation probabilities are obtained from a lexical translation model estimated by maximum likelihood from the word alignments of the parallel corpus.

The values of these four scores for the synthetic phrase pairs can be calculated in different ways and this may affect the scores of the phrase pairs extracted from the original training corpus. In this respect, it is desirable that the scoring method applied to both synthetic and corpus-extracted phrase pairs increases the probability of those phrase pairs whose SL phrase are translated in the same way in the training corpus and by the RBMT system. In addition, the scoring method should also consider the frequency in the parallel corpus of the SL phrases when a translation performed by the RBMT system does not agree with that found in the training corpus. Finally, it is also desirable that the addition of the synthetic phrase pairs to the statistical models does not involve a big computational effort, since it is executed for every text to be translated.

12. If bilingual phrase pairs were generated from the segments from the training corpus that match a rule, the method would be less effective when dealing with data sparseness, since synthetic phrases generated from rules would only be available for the sequences of words present in the training corpus.

In this section, we propose a method[13] for integrating the set of synthetic phrase pairs obtained from the RBMT data in the PBSMT system that meets the aforementioned requirements. The remainder of this section contains, in addition to our method, the description of other phrase scoring approaches that can be found in literature and their limitations.[14] All the strategies presented below have been evaluated as will be described in Section 4.

3.2.1 Creating an Additional Phrase Table

One simple strategy for integrating the synthetic phrase pairs in the hybrid SMT system is that of putting them in an different (synthetic) phrase table, as Koehn and Schroeder (2007) propose in the context of domain adaptation. When the decoder builds hypotheses, it looks for phrase pairs in both phrase tables and if the same phrase pair is found in both, one instance from each phrase table is used to build the hypotheses. It is for this reason that some authors refer to this approach as *alternative decoding paths*. Each score in each phrase table receives a different weight during the tuning process, which should help the hybrid system to obtain the appropriate relative weighting of both sources of phrase pairs.

When this scoring strategy is used to integrate the synthetic phrase pairs into the PB-SMT models, the phrase translation probabilities in the synthetic phrase table are computed by means of Eq. (1), as is done with the phrase pairs extracted from the parallel corpus, and using the counts within the set of synthetic phrase pairs. The lexical weighting scores of each phrase pair are computed from a set of word alignments and a lexical translation model as described by Koehn (2010, §5.3.3). The lexical translation model to be used is estimated from a synthetic corpus generated only from the RBMT bilingual dictionary, as described in Section 3.1; the word alignments used are those obtained by tracing back the operations carried out by the RBMT engine.[15]

Since both phrase tables are computed in a totally independent way, the phrase translation probabilities of the phrase pairs which appear in both phrase tables are not increased in comparison with those of the phrase pairs that appear in only one of them. Consider, for instance, that the SL phrase a has two different translations according to the RBMT system: b and c. The source-to-target phrase translation probabilities in the synthetic phrase table

13. This method has already been described by Sánchez-Cartagena, Sánchez-Martínez, and Pérez-Ortiz (2011a); however, this is the first time it has been systematically compared to other scoring methods found in the literature and evaluated with automatically inferred rules.

14. Methods in which the relevance of the phrase tables being combined must be defined in advance (i.e., there is a primary and a secondary phrase table), such as *fill-up* (Bisazza, Ruiz, & Federico, 2011), are not described in this section and have not been evaluated. We leave the responsibility of adapting the relative relevance of both types of phrase pairs to the type of texts to be translated to the tuning step of the SMT training process.

15. The Apertium engine keeps track on each step of its translation pipeline of the input word from which each output word has been obtained. The path starting from each input SL surface form is then followed in order to obtain the TL surface form aligned to it. An exception is made when a step of the pipeline converts an input word into multiple output words or vice-versa. In that case, the words involved are left unaligned; this is done to avoid generating too many word alignments that could be incorrect. Let us suppose that the Spanish sentence *Por otra parte mis amigos americanos han decidido venir* is translated into English as *On the other hand my American friends have decided to come* by Apertium. The Spanish phrase *Por otra parte* is analysed by Apertium as a single lexical form. After being translated into English, it produces the segment *on the other hand* in the generation step. If the exception were not made, the SL word *por* would be aligned with the four TL words *on*, *the*, *other* and *hand* and the SL words *otra* and *parte* would also be aligned with the same set of TL words.

for the resulting phrase pairs would be $\phi_{\text{synth}}(b|a) = 0.5$ and $\phi_{\text{synth}}(c|a) = 0.5$. Let us also suppose that, after extracting phrase pairs from the parallel corpus, the phrase pairs (a, b) and (a, d) have the same frequency, and there are no other phrase pairs with a as a source. The resulting source-to-target phrase translation probabilities would be $\phi_{\text{corpus}}(b|a) = 0.5$ and $\phi_{\text{corpus}}(d|a) = 0.5$. Although there is evidence that suggests that b is a more likely translation than c and d, the three translations have the same probability.

3.2.2 Phrase Table Linear Interpolation

Alternatively, once the two phrase tables have been built, they can be linearly interpolated into a single one (Sennrich, 2012, §2.1). The scores attached to each phrase pair in the resulting phrase table are obtained as the linear interpolation of the value of the corresponding score in the corpus-extracted phrase table and in the synthetic phrase table. For instance, the source-to-target phrase translation probability is computed as shown in Eq. (2) below, in which $\text{count}_{\text{synth}}(\cdot)$ is the frequency of a phrase pair in the list of phrase pairs generated from the RBMT system, $\text{count}_{\text{corpus}}(\cdot)$ is the frequency of a phrase pair in the list of phrase pairs extracted from the parallel corpus and λ_{corpus} and λ_{synth} are the weights for both phrase tables; obviously $\lambda_{\text{corpus}} + \lambda_{\text{synth}} = 1$. These weights are optimised by means of perplexity minimisation on a phrase table built from a development set (Sennrich, 2012, §2.4).

$$\phi(t|s) = \lambda_{\text{corpus}} \frac{\text{count}_{\text{corpus}}(s,t)}{\sum_{t_i} \text{count}_{\text{corpus}}(s,t_i)} + \lambda_{\text{synth}} \frac{\text{count}_{\text{synth}}(s,t)}{\sum_{t_i} \text{count}_{\text{synth}}(s,t_i)} \tag{2}$$

This method, unlike that which uses two independent phrase tables and is described in Section 3.2.1, increases the phrase translation probability of the phrase pairs that appear in both phrase tables over those that are only present in one of them. For the phrase pairs (a, b), (a, c) and (a, d) mentioned above, the resulting probabilities would be $\phi(b|a) = 0.5\lambda_{\text{synth}} + 0.5\lambda_{\text{corpus}} = 0.5$; $\phi(c|a) = 0.5\lambda_{\text{synth}}$; and $\phi(d|a) = 0.5\lambda_{\text{corpus}}$. However, this method does not use the frequency of the source phrases in the training corpus when interpolating the phrase tables. If the source phrase x is found only once in the training corpus, and it is aligned with y, but its only possible translation according to the RBMT system is z, the source-to-target phrase translation probabilities of both phrase pairs would be $\phi(y|x) = \lambda_{\text{corpus}}$ and $\phi(z|x) = \lambda_{\text{synth}}$, respectively. If x were found 10 000 times in the training corpus, and always translated as y, the probabilities would be exactly the same because the weights λ_{corpus} and λ_{synth} are the same for all the phrase pairs. However, the phrase pair (x, y) is much more reliable when it is found in the training corpus 10 000 times than when it is found only once. If the probabilities in the resulting phrase table reflected this difference, the decoder would presumably be able to choose better phrase pairs and produce more reliable translations.

3.2.3 Proposed Strategy: Directly Expanding the Phrase Table

One way of taking into account the absolute frequency of the different phrases in the training corpus is to join synthetic phrase pairs and corpus-extracted phrase pairs and calculate the phrase translation probabilities by means of relative frequency as usual. The source-to-target phrase translation probabilities of the resulting phrase table are therefore computed

as follows:

$$\phi(t|s) = \frac{\text{count}_{\text{corpus}}(s,t) + \text{count}_{\text{synth}}(s,t)}{\sum_{t_i}(\text{count}_{\text{corpus}}(s,t_i) + \text{count}_{\text{synth}}(s,t_i))} \quad (3)$$

Since $\text{count}_{\text{synth}}(\cdot) = 1$ for all the synthetic phrase pairs, when a synthetic phrase pair share its SL side with a corpus-extracted phrase pair, the source-to-target phrase translation probability of the synthetic phrase pair may be too small compared to the phrase pair extracted from the training corpus.[16] Depending on the texts to be translated, it may be desirable for a synthetic phrase pair to have a higher phrase translation probability than a corpus-extracted phrase pair with the same SL side. In order to adapt their relative weight to the texts to be translated, an additional binary feature function that flags synthetic phrase pairs is added to the phrase table.[17]

The lexical weighting scores of the phrase table built with this combination method are obtained by using the same lexical translation model for both types of phrase pairs. The model (actually, one model for source-to-target and another model for target-to-source lexical weighting) is obtained from the concatenation of the training parallel corpus and the synthetic phrase pairs generated from the RBMT bilingual dictionary. The lexical weighting scores are then computed using the word alignments obtained by statistical methods for the corpus-extracted phrase pairs, as usual (Koehn, 2010, §5.2.1), and those obtained by tracing back the operations carried out in the different translation steps of Apertium for the synthetic phrase pairs (see Section 3.2.1).

3.2.4 Augmenting the Training Corpus

Finally, the simplest approach involves appending the RBMT-generated phrase pairs to the training corpus and running the usual PBSMT training algorithm. Unlike in the previous approaches, this improves the alignments of the original training corpus and enriches the lexicalised reordering model (Koehn, 2010, §5.4.2), in addition to the phrase table. The phrase extraction algorithm (Koehn, 2010, §5.2.3) may, however, split the resulting bilingual phrase pairs into smaller units which may signify that multi-word expressions are not translated in the same way as they appear in the RBMT bilingual dictionary.

16. The same applies to phrase pairs that share their TL side and the target-to-source phrase translation probability.

17. In order to take into account the absolute frequencies in the parallel corpora from which the two phrase tables to be combined have been obtained, Sennrich (2012, §4.2) proposes the *weighted counts* interpolation method, which is similar to that presented in this paper. There are two main differences between both approaches. Firstly, in order to adapt the weight of both types of phrases to the texts to be translated, the *weighted counts* approach multiplies the frequency of each phrase pair by a factor before building the phrase table; depending on the origin of the phrase, a different factor is used. On the contrary, our method adds a binary feature function to the phrase table. And secondly, the *weighted counts* approach optimises the factors that determine the relative weight of each type of phrase pair by means of perplexity minimisation on a phrase table built from a development set (Sennrich, 2012, §2.4) in isolation, i.e., with no connection to the rest of the elements present in the log-linear model. In contrast, the new method optimises the weight of the binary feature function together with the rest of the elements in the log-linear model during the tuning process. Given the poor results obtained by the phrase table interpolation method —in which the weights are also optimised by means of perplexity minimisation— in the experiments reported in Section 4.2, *weighted counts* has not been included in the experimental setup.

Although this strategy is not feasible in a real-world environment because of the computational cost of word aligning the whole training corpus for each document to be translated,[18] it is worth evaluating it because it is the only strategy that enriches the data from which the lexicalised reordering model is obtained.

4. Evaluation with Hand-Crafted Resources

A set of experiments whose objective was evaluating the feasibility of the hybridisation strategy described in Section 3 when using hand-crafted linguistic resources in the Apertium RBMT platform has been conducted. We compare, for different language pairs, training corpus sizes and text domains, the translation quality achieved by a baseline PBSMT system, by the RBMT system from which the data is extracted, by Eisele et al.'s (2008) approach and by a set of hybrid systems using the phrase scoring alternatives described in Section 3.2.

4.1 Experimental Setup

The language pairs used for evaluation are Breton→French[19] and English↔Spanish.[20] Breton→French has been chosen because it has the problem of resource scarceness: there are only around 60 000 parallel sentences available for this pair (Tyers, 2009; Tiedemann, 2012). English↔Spanish have been chosen because they have a wide range of parallel corpora available and this allows us to perform both in-domain and out-of-domain evaluations. Moreover, as Spanish is a highly inflected language and English is not, the results for both directions of the English↔Spanish language pair allow us to evaluate in detail the impact of the hybrid strategy in the translation of highly inflected languages.

The translation model of the PBSMT systems for English↔Spanish has been trained on the Europarl parallel corpus (Koehn, 2005) version 5; [21] the TL model has been trained on the same corpus. In both cases, the Q4/2000 portion has been set aside for evaluation purposes. Different subsets of the parallel corpus with different number of sentences have been used to build the systems; however, in all cases the language model was trained on the whole TL side of the Europarl corpus. These subsets have been randomly chosen in such a way that larger corpora include the sentences in the smaller ones. The different subcorpora contain 10 000, 40 000, 160 000, 600 000 and 1 272 260 sentences; the latter corresponds to the whole training corpus.

Regarding Breton→French, the translation model has been built using the only freely-available parallel corpus for this language pair (Tyers, 2009; Tiedemann, 2012), which contains short sentences from the tourism and computer localisation domains. Different training corpus sizes have been used too, namely 10 000, 25 000 and 54 196 parallel sentences. The latter corresponds to the whole corpus except for the subsets reserved for tuning and testing. As in the English→Spanish language pair, sentences have been randomly chosen in such a way that larger corpora include the sentences in the smaller ones. The TL model

18. Recall that a different set of synthetic phrase pairs is generated for each SL text to be translated.
19. There is not French→Breton RBMT system in the Apertium platform.
20. The symbol → means that the first language acts as the the SL and the second one as the TL. The symbol ↔ means that the evaluation has been performed in both translation directions.
21. `http://www.statmt.org/europarl/archives.html#v5`

has been learnt from a monolingual corpus built by concatenating the target side of the whole parallel training corpus and the French Europarl corpus provided for the WMT 2011 shared translation task.[22]

Although there are larger monolingual corpora available for the target languages included in the evaluation setup, they have not been used because our experiments are focused on evaluating the impact of the RBMT data on the PBSMT translation model. By learning the TL model from a monolingual corpus that does not exceed the size of the biggest parallel corpus used in the experiments, the risk that a huge language model will overshadow the impact of the RBMT data on the SMT translation model is reduced. Note that, in a real-world environment, the size of the TL model may need to be limited if the hybrid MT system is required to have a reduced memory footprint, for example, because it is going to be executed in a handheld device.[23]

Breton→French systems were tuned using 3 000 parallel sentences randomly chosen from the available parallel corpus and evaluated using another randomly chosen subset of the same size; obviously both subsets were not used for training. Only an in-domain evaluation could be performed for this language pair. Regarding English↔Spanish, both in-domain and out-of-domain evaluations have been carried out. The former was performed by tuning the systems with 2 000 parallel sentences randomly chosen from the Q4/2000 portion of Europarl v5 corpus (Koehn, 2005) and evaluating them with 2 000 random parallel sentences from the same portion of the corpus; special care was taken to avoid the overlapping between the test and tuning sets. The out-of-domain evaluation was performed by using the *newstest2008* set for tuning and the *newstest2010* test for testing; both sets belong to the news domain and are distributed as part of the WMT 2010 shared translation task.[24] Table 1 summarises the data concerning the corpora used in the experiments. Sentences that contain more than 40 tokens were removed from all the parallel corpora, as is customary, in order to avoid problems with the word alignment tool GIZA++ (Och & Ney, 2003).[25]

All the experiments were carried out with release 2.1 of the free/open-source PBSMT system Moses (Koehn et al., 2007) together with the SRILM language modelling toolkit (Stolcke, 2002), which was used to train a 5-gram language model using interpolated Kneser-Ney discounting (Goodman & Chen, 1998). Word alignments were computed by means of GIZA++ (Och & Ney, 2003). The weights of the different feature functions were optimised by means of minimum error rate training (Och, 2003). The parallel corpora were lowercased and tokenised before training, as were the test sets used to evaluate the systems.

The hand-crafted shallow-transfer rules and dictionaries were borrowed from the Apertium platform (Forcada et al., 2011). In particular, the engine and the linguistic resources for English↔Spanish, and Breton→French were downloaded from the Apertium Subversion

22. `http://www.statmt.org/wmt11/translation-task.html`
23. There are also more English↔Spanish parallel corpora available, but they have not been used in the experiments because one of the main objectives of the hybrid approach presented in this paper, as pointed out in the introduction, is to alleviate the data sparseness problem in SMT.
24. `http://www.statmt.org/wmt10/translation-task.html`
25. Preliminary experiments showed that, when sentences contained more than 40 tokens, GIZA++ was not able to align some of them. Sentences with more than 40 tokens were also removed from the tuning and test sets in order to ensure that the approach by Eisele et al. (2008) is able to extract all the phrase pairs needed. Recall that this method needs to align the sentences in the test set with their RBMT translations.

Corpus	#sentences	Source		Target	
		#words	#voc	#words	#voc
Language model (English)	1 650 152	-	-	45 712 294	110 018
Language model (Spanish)	1 650 152	-	-	47 734 244	165 896
training	10 000	209 562	11 561	216 187	15 884
	40 000	836 194	20 883	862 789	30 583
	160 000	3 341 577	36 798	3 452 067	55 584
	600 000	12 546 758	61 654	12 971 035	94 315
	1 272 260	26 595 542	82 585	27 496 270	125 813
Europarl tuning	2 000	42 642	5 157	43 348	6 411
Europarl testing	2 000	42 114	5 080	42 661	6 289
newstest2012 tuning	1 732	34 878	6 209	36 410	7 085
newstest2013 testing	2 215	48 367	7 701	50 745	9 277

(a) English↔Spanish

Corpus	#sentences	Source		Target	
		#words	#voc	#words	#voc
Language model (French)	2 041 625	-	-	60 356 583	155 028
training	10 000	146 255	16 711	146 556	17 588
	25 000	365 856	27 606	369 396	28 333
	54 196	795 045	41 157	801 780	40 279
tuning	3 000	44 586	8 340	45 086	8 907
testing	3 000	43 276	8 119	43 419	8 832

(b) Breton→French

Table 1: Number of sentences, words, and size of the vocabulary of the training, tuning and test sets used in the experiments.

repository.[26] The Apertium linguistic data contains 326 228 entries in the English→Spanish bilingual dictionary, 284 English→Spanish shallow-transfer rules and 138 Spanish→English shallow-transfer rules. Regarding Breton→French, the bilingual dictionary contains 21 593 entries and there are 254 shallow-transfer rules.[27]

For each language pair, domain, and training corpus size, the following systems were built and evaluated:

- *baseline*: a standard PBSMT system.[28]

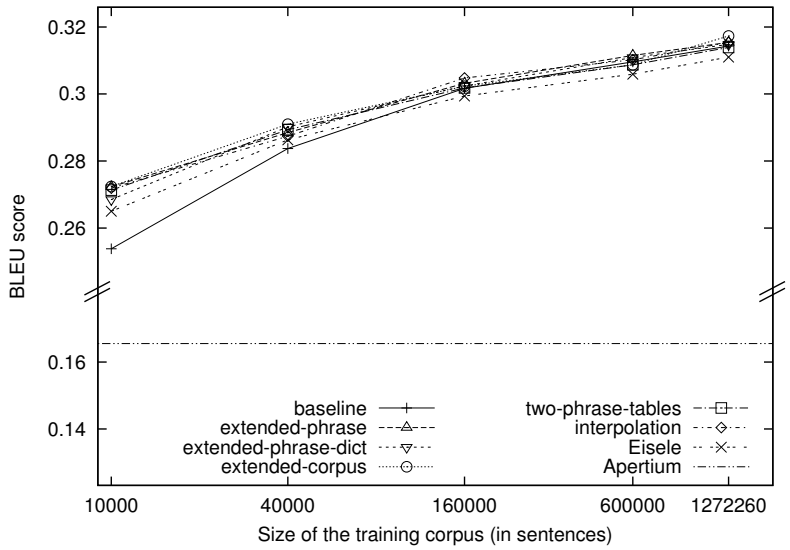
26. Revisions 24177, 22150 and 28674, respectively.
27. The transfer phase is split by Apertium in three steps (Forcada et al., 2011) for the language pairs we have used, and each step works with its own set of rules. Specifically, the Apertium linguistic data contains 216 *chunker* rules, 60 *interchunk* rules, and 7 *postchunk* rules for English→Spanish; 106 chunker rules, 31 interchunk rules, and 7 postchunk rules for Spanish→English; and 169 chunker rules, 79 interchunk rules and 6 postchunk rules for Breton→French.
28. With the same features as the baseline system of the WMT 2011 shared translation task: `http://www.statmt.org/wmt11/baseline.html`.

- *Apertium*: the Apertium shallow-transfer RBMT engine, from which the dictionaries and transfer rules were borrowed.
- *extended-phrase*: the hybrid system described in Section 3 following the strategy for scoring the phrase pairs generated from the RBMT data described in Section 3.2.3.
- *extended-phrase-dict*: the same as above, but using only the dictionaries of the RBMT system (without shallow-transfer rules). The comparison between this system and *extended-phrase* permits the evaluation of the impact of the use of shallow-transfer rules.
- *extended-corpus*: the hybrid system described in Section 3 following the strategy used to score the synthetic phrase pairs which simply involves adding the synthetic phrase pairs to the training corpus (see Section 3.2.4).
- *two-phrase-tables*: the hybrid system described in Section 3 following the strategy used to score the synthetic phrase pairs based on two independent phrase tables (Koehn & Schroeder, 2007) (see Section 3.2.1).
- *interpolation*: the hybrid system described in Section 3 following the strategy used to score the synthetic phrase pairs based on the linear interpolation of two phrase tables (Sennrich, 2012, §2.1) (see Section 3.2.2). The interpolation weights were obtained by means of perplexity minimisation on a phrase table built from the tuning set.
- *Eisele*: the approach by Eisele et al. (2008), using the alignment model learnt from the training corpus to obtain the word alignments between the source sentences and the RBMT-translated sentences.

4.2 Results and Discussion

Figures 1–5 show the BLEU (Papineni, Roukos, Ward, & Zhu, 2002) automatic evaluation score for the systems evaluated; TER (Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006) and METEOR (Banerjee & Lavie, 2005) behave in a similar manner. In addition, the statistical significance of the difference between the BLEU, TER and METEOR scores obtained by the hybridisation approach *extended-phrase* (see Section 3.2.3) and those obtained by the other systems has been computed by means of paired bootstrap resampling (Koehn, 2004) ($p \leq 0.05$; 1 000 iterations).[29] The results of this pair-wise comparison are reported in a table, included in each figure, in which each cell represents the reference system to which the approach *extended-phrase* is compared and the training corpus size; the table contains the results for the three evaluation metrics: BLEU (B), TER (T) and METEOR (M). An arrow pointing upwards (⇑) means that *extended-phrase* outperforms the reference system, an arrow pointing downwards (⇓) means that the reference system outperforms *extended-phrase*, and an equal sign (=) means that the difference between both systems is not statistically significant.

29. Only the *extended-phrase* is compared to the other systems because it is expected to achieve the highest translation quality of the different hybrid approaches, as in theory it overcomes most of the limitations of the other approaches (see Section 3.2).



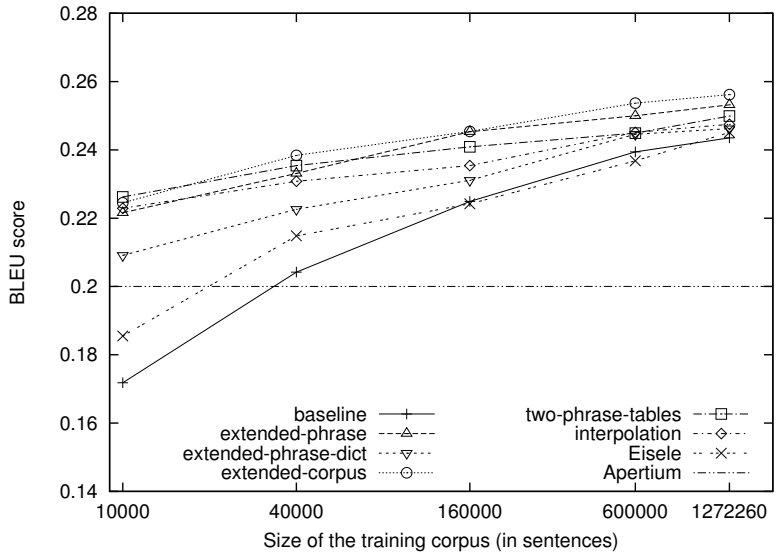
(a) BLEU scores.

system	10 000	40 000	160 000	600 000	1 272 260
metric	B T M	B T M	B T M	B T M	B T M
baseline	⇑ ⇑ ⇑	⇑ ⇑ ⇑	= ⇑ ⇑	= = =	= = ⇑
Apertium	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase-dict	⇑ ⇑ ⇑	= = ⇓	= = =	= = =	= ⇑ =
extended-corpus	= = =	= = =	= ⇑ ⇑	⇑ ⇑ =	= = =
two-phrase-tables	= = =	= = =	= ⇑ ⇑	= ⇑ ⇑	= ⇑ =
interpolation	= = =	= = =	= = =	= = =	= = =
Eisele	⇑ ⇑ ⇑	= = ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑

(b) Paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between *extended-phrase* and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU (B), TER (T) and METEOR (M). ⇑ means that *extended-phrase* outperforms the reference method by a statistically significant margin, ⇓ means the opposite, and = means that there is no statistically significant difference between them.

Figure 1: For the English→Spanish in-domain evaluation, automatic evaluation scores obtained for the baseline PBSMT system, Apertium, the hybrid approaches described in Section 3.2, and the hybrid approach by Eisele et al. (2008). The table shows a pair-wise comparison with the system *extended-phrase* (see Section 3.2.3).

These results show that the hybrid approach described in Section 3 (*extended-phrase*) outperforms both the RBMT and the baseline PBSMT system by a statistically significant margin in different scenarios. Namely, when translating out-of-domain texts (texts whose domain is different from the domain of the parallel corpus used; this occurs for all training corpus sizes and language pairs) and when translating in-domain texts with an SMT system trained on a relatively small parallel corpus. Thus, as was found in literature (see Section 2),



(a) BLEU scores.

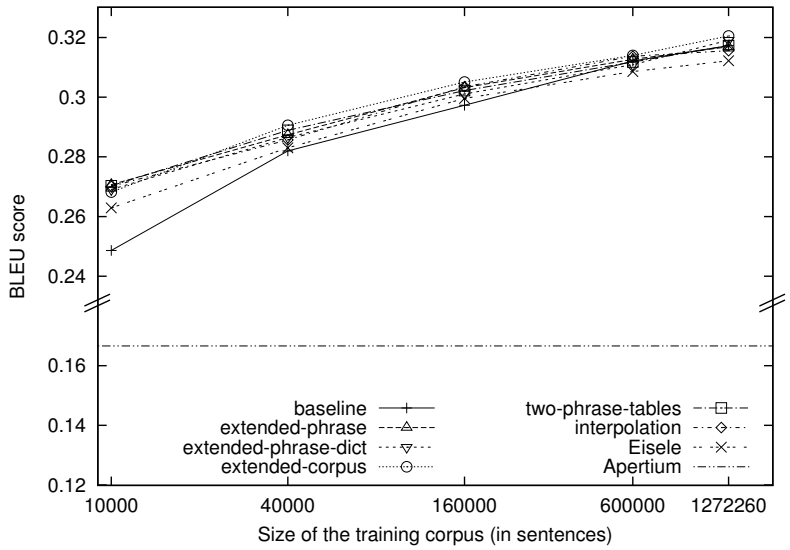
system	10 000	40 000	160 000	600 000	1 272 260
metric	B T M	B T M	B T M	B T M	B T M
baseline	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Apertium	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase-dict	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-corpus	= = ⇓	⇓ ⇓ ⇓	= = ⇓	⇓ = =	⇓ = =
two-phrase-tables	⇓ ⇓ ⇓	= ⇓ =	⇑ ⇑ =	⇑ ⇑ ⇑	⇑ ⇑ ⇑
interpolation	= = =	= = ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Eisele	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑

(b) Paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between *extended-phrase* and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU (B), TER (T) and METEOR (M). ⇑ means that *extended-phrase* outperforms the reference method by a statistically significant margin, ⇓ means the opposite, and = means that there is no statistically significant difference between them.

Figure 2: For the English→Spanish out-of-domain evaluation, automatic evaluation scores obtained for the baseline PBSMT system, Apertium, the hybrid approaches described in Section 3.2, and the hybrid approach by Eisele et al. (2008). The table shows a pair-wise comparison with the system *extended-phrase* (see Section 3.2.3).

it is possible to confirm that shallow-transfer RBMT and PBSMT systems can be combined in a hybrid system that outperforms both of them.

With regard to the differences observed in the results for the in-domain and out-of-domain evaluations, it is important to state that, for English↔Spanish, the out-of-domain tuning and test sets come from a general (news) domain and the RBMT data has been developed bearing in mind the translation of general texts (mainly news). In this case, Apertium-generated (synthetic) phrase pairs, which contain hand-crafted knowledge from



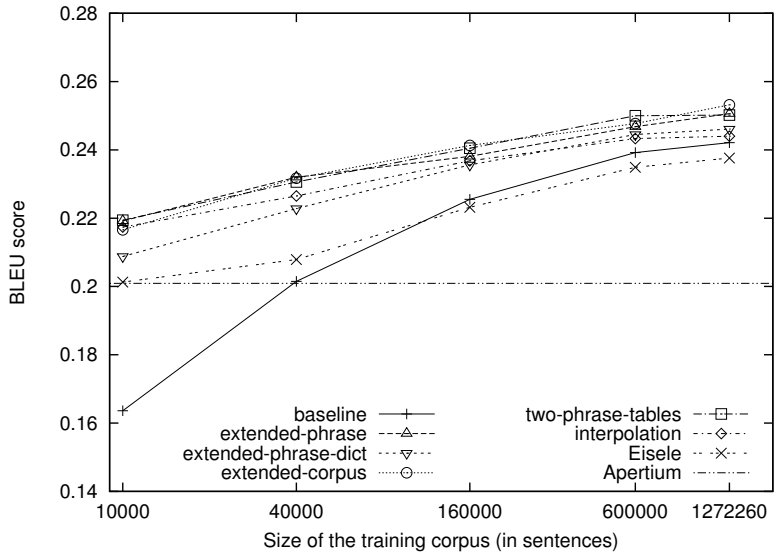
(a) BLEU scores.

system	10 000	40 000	160 000	600 000	1 272 260
metric	B T M	B T M	B T M	B T M	B T M
baseline	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	= = =	= = =
Apertium	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase-dict	= = ⇑	= = =	= = =	= ⇓ =	= ⇓ =
extended-corpus	= = ⇑	⇓ = ⇓	= = =	= = =	⇓ ⇓ ⇓
two-phrase-tables	= = =	= = =	= = ⇑	= = =	= = =
interpolation	= = ⇑	= = =	= = =	= = =	= = =
Eisele	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ = ⇑	⇑ = ⇑	⇑ = ⇑

(b) Paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between *extended-phrase* and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU (B), TER (T) and METEOR (M). ⇑ means that *extended-phrase* outperforms the reference method by a statistically significant margin, ⇓ means the opposite, and = means that there is no statistically significant difference between them.

Figure 3: For the Spanish→English in-domain evaluation, automatic evaluation scores obtained for the baseline PBSMT system, Apertium, the hybrid approaches described in Section 3.2, and the hybrid approach by Eisele et al. (2008). The table shows a pair-wise comparison with the system *extended-phrase* (see Section 3.2.3).

a general domain, cover sequences of words in the input text which are not covered, or are sparsely found, in the original training corpus. Contrarily, the in-domain tests reveal that, as soon as the PBSMT system is able to learn some reliable information from the parallel corpus, the synthetic RBMT phrase pairs become useless because the in-domain test sets come from the specialised domain of parliament speeches. For Breton→French, given the small size of the corpus available, the hybrid approach outperforms both pure RBMT and PBSMT approaches in all the experiments performed.



(a) BLEU scores.

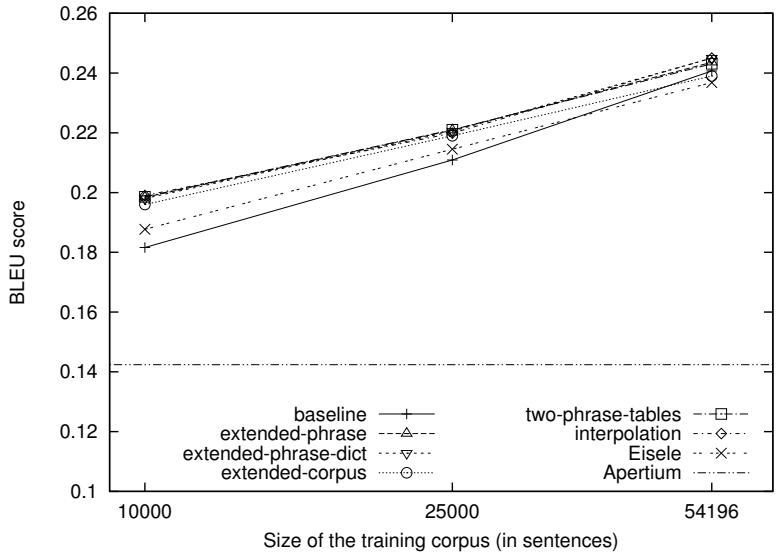
system	10 000	40 000	160 000	600 000	1 272 260
metric	B T M	B T M	B T M	B T M	B T M
baseline	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Apertium	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase-dict	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	= = ⇑	⇑ ⇑ ⇑
extended-corpus	= ⇓ ⇑	= = ⇓	⇓ = ⇓	= = =	= = ⇓
two-phrase-tables	= = =	= = ⇓	= ⇓ ⇓	⇓ ⇓ =	= = =
interpolation	= = =	⇑ ⇑ =	= = ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Eisele	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑

(b) Paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between *extended-phrase* and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU (B), TER (T) and METEOR (M). ⇑ means that *extended-phrase* outperforms the reference method by a statistically significant margin, ⇓ means the opposite, and = means that there is no statistically significant difference between them.

Figure 4: For the Spanish→English out-of-domain evaluation, automatic evaluation scores obtained for the baseline PBSMT system, Apertium, the hybrid approaches described in Section 3.2, and the hybrid approach by Eisele et al. (2008). The table shows a pair-wise comparison with the system *extended-phrase* (see Section 3.2.3).

An analysis of the proportion of synthetic phrase pairs included by the decoder in the final translation[30] for the different evaluation scenarios, depicted in figures 6–8, confirms the reason for the differences between the in-domain and out-of-domain results. For each English↔Spanish training corpus size and hybrid system, the proportion of synthetic phrases is higher in the out-of-domain evaluation.

30. If a synthetic phrase pair has also been obtained from the parallel corpus, it is not considered as synthetic in figures 6–8.



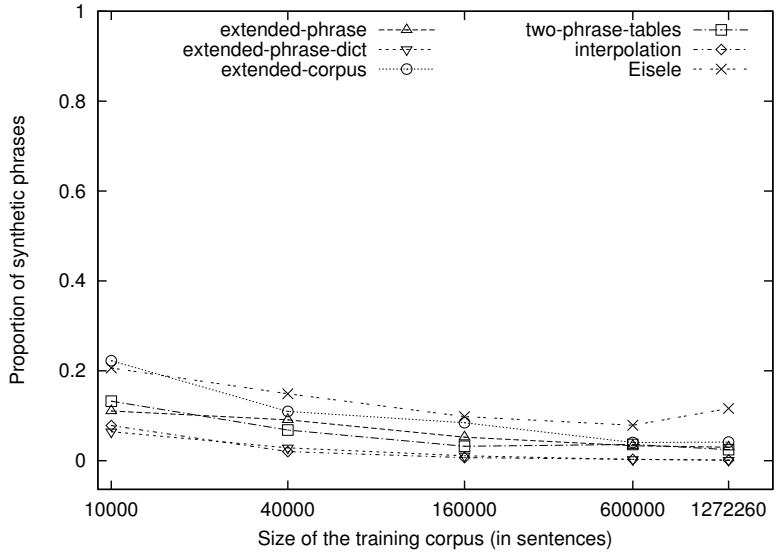
(a) BLEU scores.

system	10 000	25 000	54 196
metric	B T M	B T M	B T M
baseline	⇑ ⇑ ⇑	⇑ ⇑ ⇑	= ⇑ ⇑
Apertium	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase-dict	= = =	= = =	= = =
extended-corpus	= ⇑ =	= ⇑ ⇑	⇑ = ⇑
two-phrase-tables	= ⇓ =	= = =	= = =
interpolation	= ⇓ =	= = ⇑	= = =
Eisele	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑

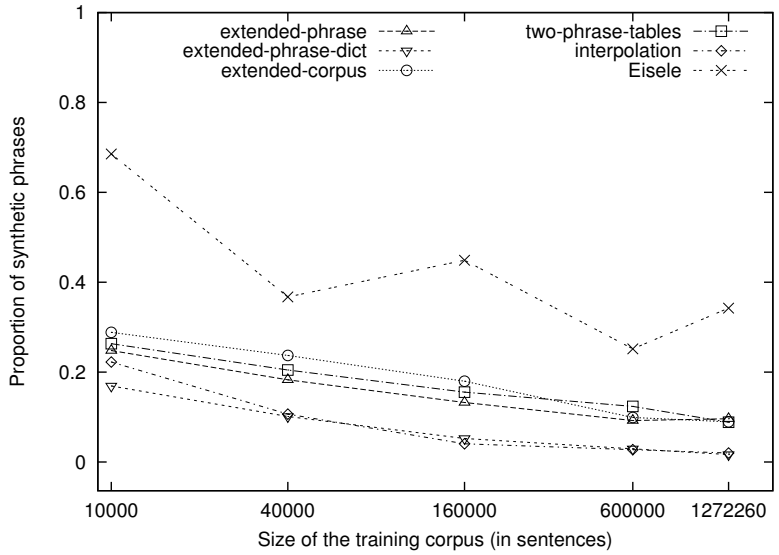
(b) Paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between *extended-phrase* and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU (B), TER (T) and METEOR (M). ⇑ means that *extended-phrase* outperforms the reference method by a statistically significant margin, ⇓ means the opposite, and = means that there is no statistically significant difference between them.

Figure 5: For the Breton→French in-domain evaluation, automatic evaluation scores obtained for the baseline PBSMT system, Apertium, the hybrid approaches described in Section 3.2, and the hybrid approach by Eisele et al. (2008). The table shows a pair-wise comparison with the system *extended-phrase* (see Section 3.2.3).

Regarding the difference between the hybrid systems enriched with all the RBMT resources (*extended-phrase*) and those that only include the dictionary (*extended-phrase-dict*), some patterns can be detected. For English↔Spanish, the impact of the shallow-transfer rules is higher when translating out-of-domain texts and decreases as the training corpus grows. Their impact is therefore higher when the decoder chooses a high proportion of Apertium phrases (see figures 6 and 7). Moreover, the systems including shallow-transfer rules outperform their counterparts which only include the dictionary by a wider margin



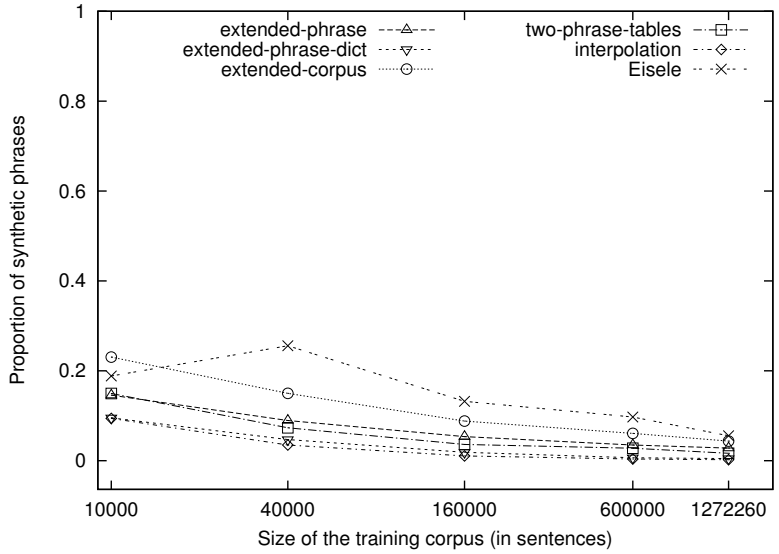
(a) In-domain evaluation.



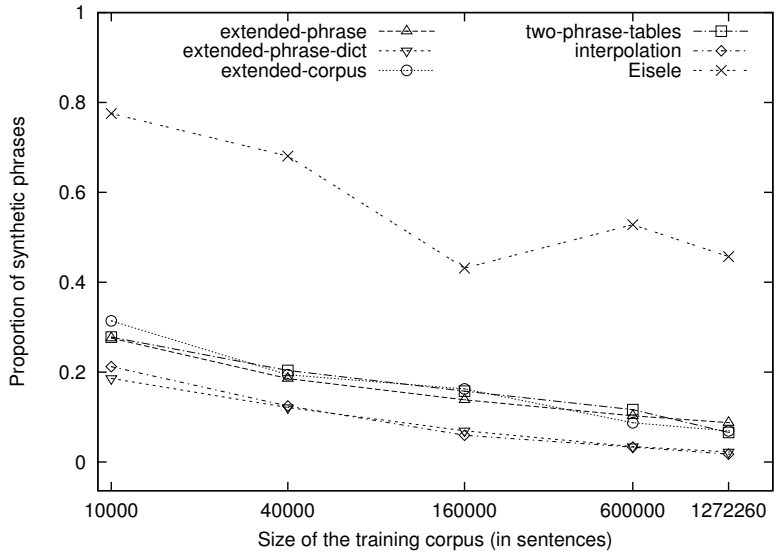
(b) Out-of-domain evaluation.

Figure 6: For English→Spanish, proportion of phrase pairs generated from the RBMT data and chosen by the decoder when translating the test set with the different hybrid approaches described in Section 3.2 and the hybrid approach by Eisele et al. (2008).

when translating out-of-domain texts from English to Spanish than the other way round. As Spanish morphology is richer, transfer rules help to perform more agreement operations when translating into Spanish. On the contrary, when Spanish is the source language, one of the main limitations suffered by the baseline PBSMT system is the high number of out-of-vocabulary (OOV) words, which is already mitigated by integrating the dictionaries into



(a) In-domain evaluation.



(b) Out-of-domain evaluation.

Figure 7: For Spanish→English, proportion of phrase pairs generated from the RBMT data and chosen by the decoder when translating the test set with the different hybrid approaches described in Section 3.2 and the hybrid approach by Eisele et al. (2008).

the phrase table with the *extended-phrase-dict* approach, as shown in figures 9–11.[31] These figures show that the amount of OOV words is much higher for the baseline system when

31. In the approach by Eisele et al. (2008) the number of OOV words is always 0 because the phrase table contains phrase pairs obtained by translating the test set with the RBMT system, and the RBMT system copies verbatim to the output those words that do not appear in its dictionaries.

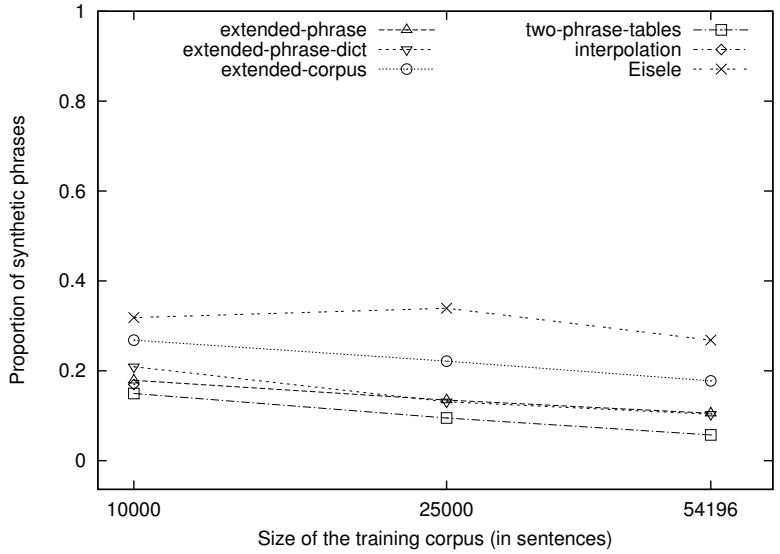
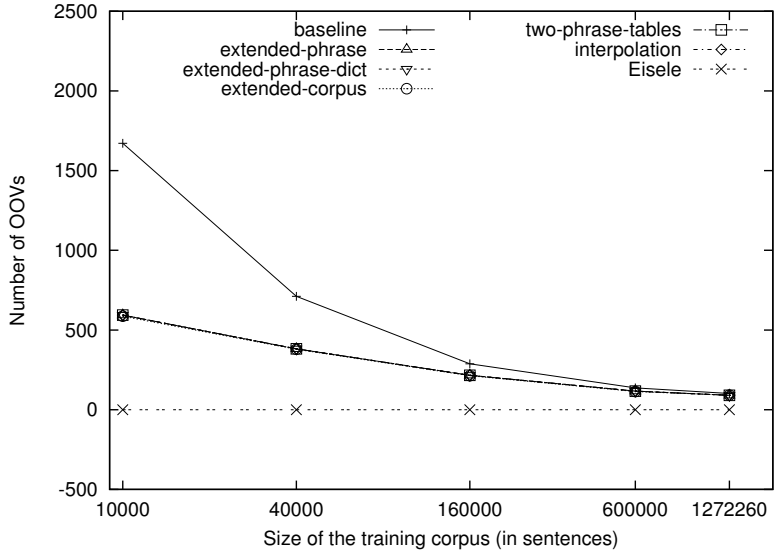


Figure 8: For Breton→French, proportion of phrase pairs generated from the RBMT data and chosen by the decoder when translating the test set with the different hybrid approaches described in Section 3.2 and the hybrid approach by Eisele et al. (2008).

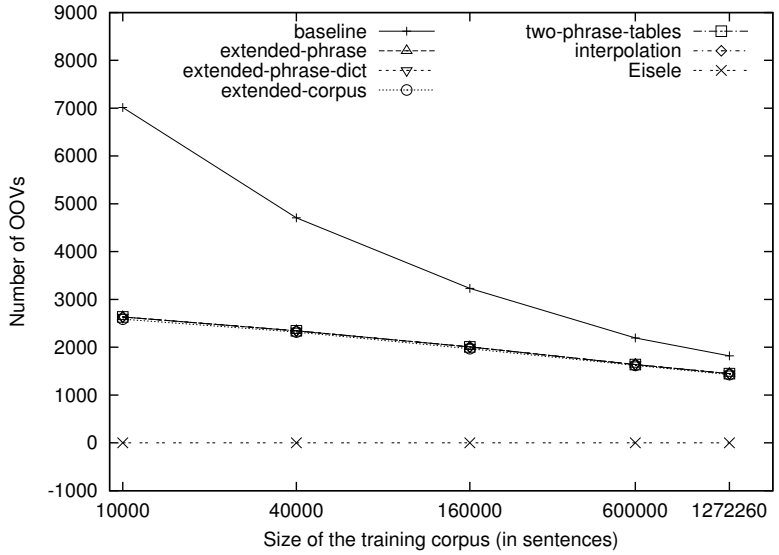
the SL is Spanish than when the SL is English and that the reduction in the amount of OOVs when adding the RBMT dictionaries is consequently also higher in the first case.

In contrast, the positive impact of the rules is very limited in the English↔Spanish in-domain evaluation, where a statistically significant improvement to the hybrid system enriched solely with dictionaries (according to the three evaluation metrics) can only be observed for the smallest English→Spanish training corpus. In fact, for a few training corpus sizes, the inclusion of the shallow-transfer rules in the hybrid system produces a statistically significant drop in translation quality according to one of the three evaluation metrics (METEOR in the case of English→Spanish in-domain evaluation and TER in the case of Spanish→English). When the training parallel corpus belongs to the same domain as the test corpus, corpus-extracted phrase pairs are likely to contain more accurate and fluent translations when compared to the mechanical and regular translations provided by the RBMT shallow-transfer rules. One possible explanation for the fact that the degradation caused by the rules is only measured by TER or METEOR is that we used BLEU for tuning (Och, 2003). Consequently, the weight of the feature function which flags whether a phrase pairs comes from the parallel corpus or from the RBMT system is set so that the inclusion of shallow-transfer rules does not penalise the translation quality as measured by BLEU. The effect of using other evaluation metrics for tuning has yet to be studied.

With regard to Breton→French, the impact of the shallow-transfer rules is also limited: the difference between the hybrid system enriched with shallow-transfer rules and the system enriched only with dictionaries is not statistically significant for any of the training corpus sizes evaluated. The reason is probably that the sentences from the test set do not have a complex grammatical structure: the average sentence length is about 9 words (Tyers, 2009) and it contains many sentences that are simply noun phrases. Another possible reason



(a) In-domain evaluation.

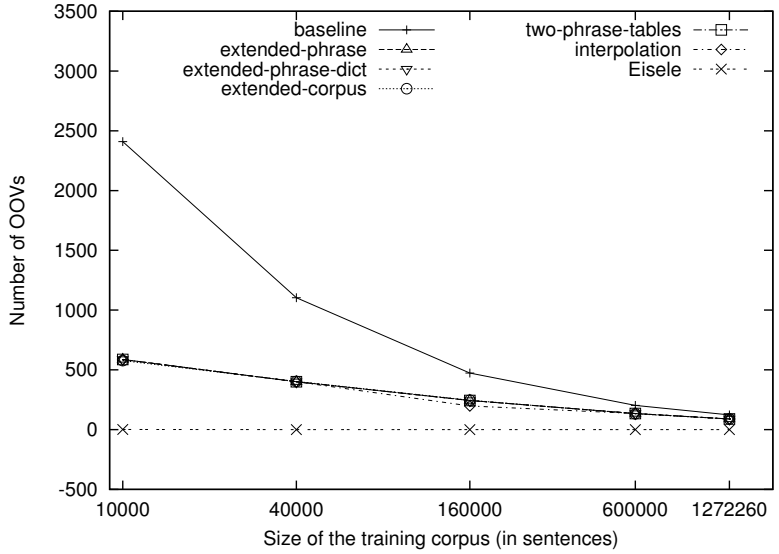


(b) Out-of-domain evaluation.

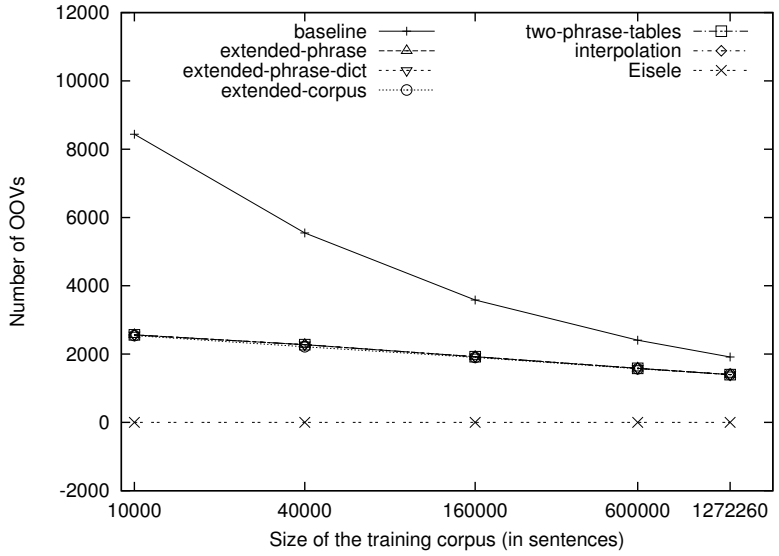
Figure 9: For English→Spanish, number of out-of-vocabulary words in the test set for the different hybrid approaches described in Section 3.2 and the hybrid approach by Eisele et al. (2008).

may be the fact that the quality of the Breton→French shallow-transfer rules may be lower than the quality of the rules used for other language pairs, since the effort spent in their development was smaller.

As regards the different phrase scoring approaches defined in Section 3.2, some differences can be observed. The most remarkable differences show up when the inclusion of synthetic phrase pairs has a great impact, that is, in English↔Spanish out-of-domain eval-



(a) In-domain evaluation.



(b) Out-of-domain evaluation.

Figure 10: For Spanish→English, number of out-of-vocabulary words in the test set for the different hybrid approaches described in Section 3.2 and the hybrid approach by Eisele et al. (2008).

uations. Firstly, the *interpolation* strategy is frequently outperformed by other strategies, and the hybrid systems built with it usually choose a relatively small proportion of synthetic phrase pairs. In theory, it should outperform the *two-phrase-tables* strategy because it assigns higher probabilities to synthetic phrase pairs that are also found in the training parallel corpus, but actually the *two-phrase-tables* approach generally achieves a higher translation quality. One possible reason for this result may be the fact that, while in the

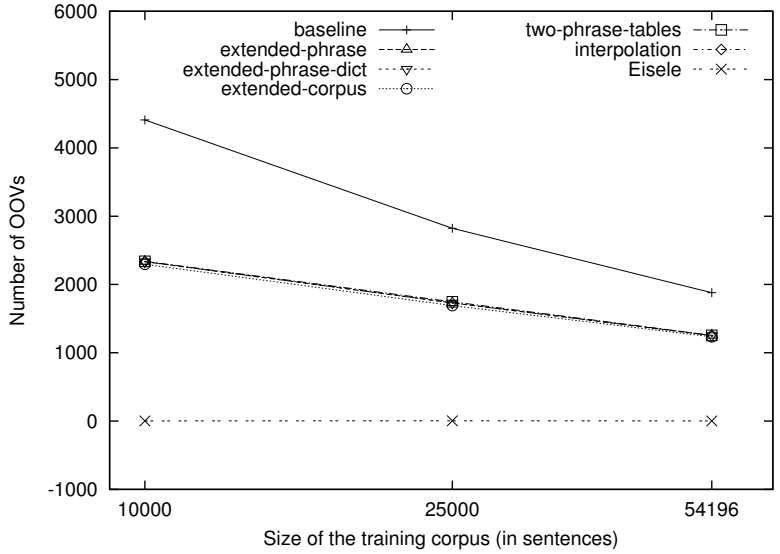


Figure 11: For Breton→French, number of out-of-vocabulary words in the test set for the different hybrid approaches described in Section 3.2 and the hybrid approach by Eisele et al. (2008).

interpolation method the relative weights of the two types of phrase pairs are optimised so as to minimise the perplexity on a set of phrase pairs extracted from a tuning corpus, in the *two-phrase-tables* strategy the relative weights are optimised so as to maximise translation quality by the minimum error rate tuning algorithm. In the latter case, the interaction of phrase pairs with the rest of the elements of the PBSMT system is taken into account during the tuning process. Nevertheless, additional experiments whose objective will be to carry out an in-depth evaluation of the impact of the method used to optimise the relative weight of both types of phrase pairs will need to be carried out. Concerning the *extended-corpus* strategy, it does not consistently outperform the other strategies, probably because the synthetic phrase pairs were too short for their subphrases to clearly improve the reordering model. However, as already stated, this strategy could not be used in a real-world setting because of the high computational cost of aligning the synthetic phrase pairs and the training corpus together for every document to be translated. Finally, the *two-phrase-tables* strategy is outperformed by the *extended-phrase* strategy in the experiments carried out with the English→Spanish language pair (except in the smallest training corpus size, where the effect of increasing the probability of the phrase pairs that appear in both phrase tables, as described in Section 3.2.1, is less relevant). For the reverse language pair, the *two-phrase-tables* strategy is sometimes better, but the three evaluation metrics never agree and the difference between both strategies is small when compared to English→Spanish. These results suggest that, at least in the evaluation scenario where the shallow-transfer rules have the highest impact, the phrase scoring strategy defined in Section 3.2.3 is able to achieve a better balance between the two sources of phrase pairs.

Finally, the hybridisation strategy defined in Section 3, together with the phrase scoring strategy defined in Section 3.2.3, outperforms the approach by Eisele et al. (2008) for all

language pairs, training corpus sizes and domains. The biggest difference between both approaches is observed when small corpora are used for training. As has been anticipated in Section 2.1, under such circumstances, no reliable alignment models can be learnt from the training corpus and therefore no reliable phrase pairs can be obtained from the input text and its RBMT translation. The approach presented in this work, contrarily, is not affected by this issue because it does not rely on word alignments in order to generate phrase pairs from the RBMT system. In addition, there is a significant difference even when the training corpus is relatively big (more than one million parallel sentences). The high proportion of synthetic phrase pairs used when compared to the other hybrid approaches (see figures 6–8) suggests that the approach by Eisele et al. is not able to find an adequate balance between both types of phrase pairs. This may be because synthetic phrase pairs are even extracted from SL segments that do not match a transfer rule and because of the straightforward scoring method used, which simply consists of concatenating the phrase table obtained from the training parallel corpus and that obtained from the RBMT system.

5. Evaluation with Automatically Inferred Rules

As has been empirically proved in the previous section, shallow-transfer rules can improve the performance of PBSMT. However, a considerable human effort and a high level of linguistic knowledge are needed to create them. In order to reduce the degree of human effort required to achieve such improvement, the algorithm proposed by Sánchez-Cartagena et al. (2015) can be used to infer a set of shallow-transfer rules from the training parallel corpus from which the PBSMT models are built, and this set of rules, together with the bilingual dictionary, can be used to enlarge the phrase table as previously described. A significant boost in translation quality could thus be achieved with the sole addition of RBMT dictionaries. In this section, a set of experiments whose objective was to assess the viability of this approach is presented.

The method proposed by Sánchez-Cartagena et al. (2015) uses parallel corpora to infer shallow-transfer rules that are compatible with the formalism used by Apertium (Forcada et al., 2011). Their approach is inspired by the method by Sánchez-Martínez and Forcada (2009), uses a generalisation of the alignment template formalism (Och & Ney, 2004) to encode transfer rules, and overcomes important limitations of the method by Sánchez-Martínez and Forcada (2009). We refer the reader to the paper by Sánchez-Cartagena et al. for a thorough description of these limitations.

The approach by Sánchez-Cartagena et al. (2015) is the first in literature in which the problem of automatically inferring transfer rules is reduced to finding the optimal value of a minimisation problem. They prove that the translation quality achieved with the automatically inferred rules is generally close to that obtained with hand-crafted rules. Moreover, for some language pairs, the automatically inferred rules are even able to outperform the hand-crafted ones.

5.1 Experimental Setup

Two considerations should be borne in mind when inferring a set of shallow-transfer rules to be integrated into the PBSMT system. Firstly, the experiments conducted by Sánchez-Cartagena et al. (2015) concluded that one of the features of the rule inference algorithm, the

generalisation of alignment templates to combinations of values of morphological inflection attributes not observed in the training corpus, is one of the causes of the vast complexity of the aforementioned minimisation problem and brings a significant translation quality boost only when the training corpus is very small (below 1 000 parallel sentences). Given the fact that the parallel corpus sizes for which an SMT system starts to be competitive are much bigger, the generalisation of morphological inflection attributes can be skipped when inferring shallow-transfer rules to be integrated into PBSMT. Moreover, preliminary experiments showed that, even when disabling the generalisation to non-observed combinations of values of morphological inflection attributes, the global minimisation algorithm still needs a huge amount of processing time in order to infer a set of rules from a parallel corpus that contains hundreds of thousands of parallel sentences.

Secondly, the rule inference algorithm by Sánchez-Cartagena et al. (2015) filters the rules to be generated so as to ensure that, when they are applied by a shallow-transfer RBMT system in a greedy, left-to-right, longest-match way, the groups of words which need to be processed together are translated with the same rule. From here on, we shall refer to this process as *optimising the rules for chunking*. Since, in principle, the SMT decoder splits the input sentences in all possible ways, this process might not be needed. Shallow-transfer rules for all the sequences of SL lexical categories present in the corpus would therefore be generated.

We ran some preliminary experiments and the results showed that there are no consistent differences between the systems whose rules have been optimised for chunking and the systems whose rules have not: statistically significant differences can only be found only for some of the evaluation metrics. For Spanish→English, optimising rules for chunking brings a tiny improvement, while for English→Spanish, the effect is the opposite. Since the impact of the rules is higher for the translation of out-of-domain texts, the effect of the optimisation is also more noticeable in this scenario.

The optimisation of rules for chunking affects the resulting hybrid system in two ways. On the one hand, it prevents the inclusion in the phrase table of multiple noisy phrase pairs that were generated from shallow-transfer rules that match sequences of lexical categories that do not need to be processed together when translating between the languages involved. Owing to the fact that the decoder cannot evaluate all the translation hypotheses, these useless phrase pairs may prevent other, more important phrase pairs from being included in the final translation. It may also occur that the language model does not have enough information to properly score the synthetic phrase pairs built from these noisy rules. From this point of view, the optimisation of rules for chunking should have a positive impact on translation quality. Furthermore, since an SMT system does not perform a greedy segmentation of the input sentence, some of the rules discarded during the optimisation for chunking in RBMT may still be useful if they are included in a PBSMT system. Rules that would prevent the application of a more important rule by the RBMT engine do not prevent the application of that rule in the hybrid system because, in principle, all the possible segmentations are taken into account. In the light of our preliminary results, it seems that the former is more relevant for Spanish→English, while the latter has a higher positive impact for English→Spanish. Since Spanish is morphologically more complex, more rules are needed to correctly perform agreements, and more rules discarded during

the optimisation for chunking were probably useful. Nevertheless, these differences have yet to be studied in greater depth.

Bearing these considerations in mind, our experiments have been carried out as follows. For the same language pairs, corpora and RBMT dictionaries used in the previous section, a new system, *extended-phrase-learnt*, has been built; in this system, the rule inference algorithm described by Sánchez-Cartagena et al. (2015) has been applied to the training corpus and the optimisation of rules for chunking has not been performed. The rules inferred, together with the dictionaries, have been used to enrich the PBSMT system following the hybridisation strategy described in Section 3. Because of the time complexity of the minimisation problem to be solved by the rule inference approach, only the first 160 000 sentences of the training corpus have been used for rule inference in those cases in which the corpus was larger than 160 000 sentences. In other words, the systems built from 160 000, 600 000, and the whole set of parallel sentences use exactly the same set of shallow-transfer rules.[32]

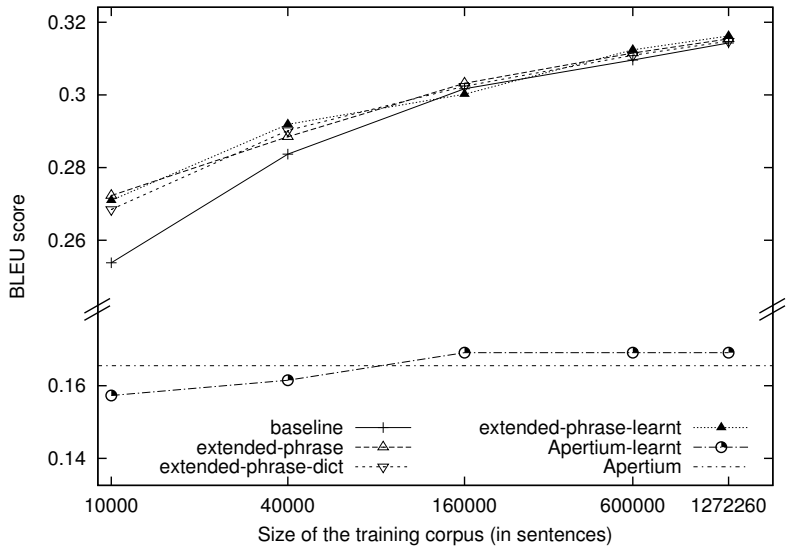
We compare the new system to a pure PBSMT baseline built from the same data, a hybrid system built from the Apertium hand-crafted rules and dictionaries, a hybrid system built with the same strategy but only from the Apertium dictionaries, and two different versions of the RBMT system Apertium: one version using hand-crafted rules and another version with automatically inferred rules. In all the hybrid systems, the scoring method described in Section 3.2.3 has been used, since this is the scoring method that proved to perform best in the experiments described in the previous section.

5.2 Results and Discussion

The comparison of the hybrid approach *extended-phrase-learnt* to the other approaches being considered in this section is presented in figures 12–16. The results show the BLEU (Papineni et al., 2002) automatic evaluation score for the different systems evaluated; TER and METEOR behavie is a similar way. In addition, the statistical significance[33] of the difference between *extended-phrase-learnt* and the other systems is also presented in a table, in the same way as depicted in the previous section.

The comparison to the PBSMT baseline and the pure RBMT system shows that our hybrid approach with automatically inferred rules behaves in the same way as when hand-crafted rules are used: it outperforms both baselines when the training corpus is small or an out-of-domain text is translated. If the comparison is performed with the hybrid system that only uses dictionaries, our hybrid approach also outperforms the dictionary-based approach in almost the same cases as the hybrid approach with hand-crafted rules: out-of-domain evaluation and in-domain evaluation only with the smallest parallel corpus size, although the three evaluation metrics do not agree in the latter case. In other words, with the automatic inference of shallow-transfer rules, a statistically significant improvement to the approach that uses only dictionaries has been achieved without using any additional linguistic resources.

32. In addition, the part of the training corpus used for rule inference has been split into two parts: the first 4/5 of the corpus has been used for actual rule inference, while the last 1/5 has been employed as a development corpus in order to optimise the threshold δ, as in the experiments described by Sánchez-Cartagena et al. (2015). For training corpora bigger than 10 000 sentences, only 2 000 sentences have been used for optimising δ, while the remaining part of the corpus has been used for rule inference.
33. Again obtained through paired bootstrap resampling (Koehn, 2004) ($p \leq 0.05$; 1 000 iterations).



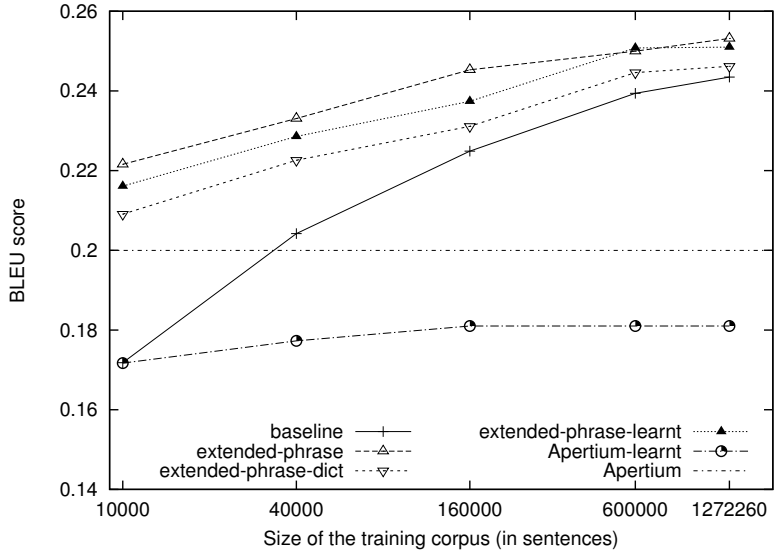
(a) BLEU scores.

system	10 000	40 000	160 000	600 000	1 272 260
metric	B T M	B T M	B T M	B T M	B T M
baseline	⇑ ⇑ ⇑	⇑ ⇑ ⇑	= = =	= = =	= = ⇑
Apertium	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Apertium-learnt	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase-dict	= = ⇑	= = =	= ⇓ ⇓	= = =	= ⇑ =
extended-phrase	= = =	⇑ = =	= ⇓ ⇓	= = =	= = =

(b) Paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between *extended-phrase-learnt* and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU (B), TER (T) and METEOR (M). ⇑ means that *extended-phrase-learnt* outperforms the reference method by a statistically significant margin, ⇓ means the opposite, and = means that there is no statistically significant difference between them.

Figure 12: For the English→Spanish in-domain evaluation, automatic evaluation scores obtained for the baseline PBSMT system, Apertium with hand-crafted rules (*Apertium-learnt*), Apertium with learnt rules (*Apertium-learnt*), and our hybrid approach (described in Section 3.2.3) using hand-crafted shallow-transfer rules (*extended-phrase*), a set of rules inferred from the training corpus (*extended-phrase-learnt*) and no rules at all (*extended-phrase-dict*). The table shows a pair-wise comparison with the system *extended-phrase-learnt*.

In some cases there is no statistically significant difference between the hybrid system with hand-crafted rules and the hybrid system with automatically inferred rules. This occurs, for instance, in the English→Spanish out-of-domain evaluation when the training corpus contains 600 000 sentence pairs. A translation quality similar to that obtained with hand-crafted rules has therefore been attained without the intervention of the human ex-



(a) BLEU scores.

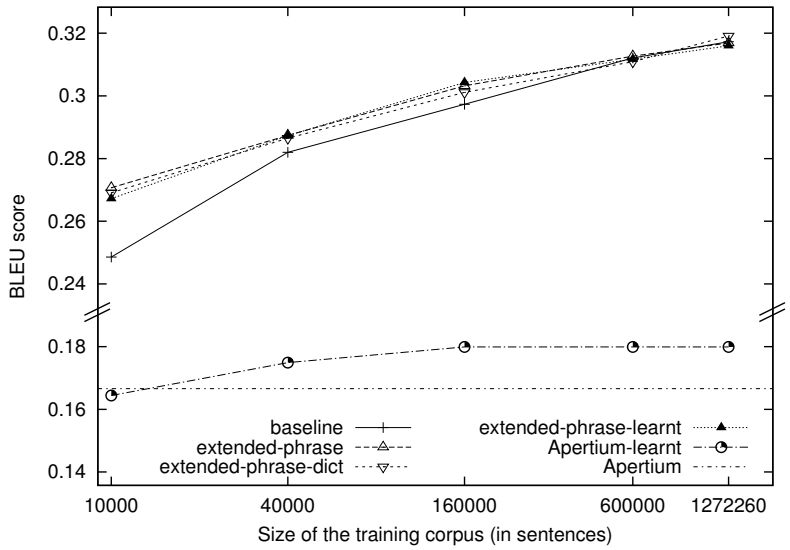
system	10 000	40 000	160 000	600 000	1 272 260
metric	B T M	B T M	B T M	B T M	B T M
baseline	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Apertium	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Apertium-learnt	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase-dict	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase	⇓ ⇓ ⇓	⇓ ⇓ ⇓	⇓ ⇓ ⇓	= = =	= ⇓ ⇓

(b) Paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between *extended-phrase-learnt* and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU (B), TER (T) and METEOR (M). ⇑ means that *extended-phrase-learnt* outperforms the reference method by a statistically significant margin, ⇓ means the opposite, and = means that there is no statistically significant difference between them.

Figure 13: For the English→Spanish out-of-domain evaluation, automatic evaluation scores obtained for the baseline PBSMT system, Apertium with hand-crafted rules (*Apertium-learnt*), Apertium with learnt rules (*Apertium-learnt*), and our hybrid approach (described in Section 3.2.3) using hand-crafted shallow-transfer rules (*extended-phrase*), a set of rules inferred from the training corpus (*extended-phrase-learnt*) and no rules at all (*extended-phrase-dict*). The table shows a pair-wise comparison with the system *extended-phrase-learnt*.

perts who usually create them.[34] In the rest of the cases, where the hybrid system with

34. Although the translation quality of both systems is similar according to automatic evaluation metrics, there are differences in the amount of rules used in each case. While the set of hand-crafted rules in the Apertium platform contains a few hundred rules for each language pair, the number of inferred rules ranges from 2 000 to 75 000, depending on the language pair and size of the training parallel corpus. These figures are not directly comparable, since the rule formalism used for the hand-crafted rules is more expressive than that of the automatically inferred rules (Sánchez-Cartagena et al., 2015,



(a) BLEU scores.

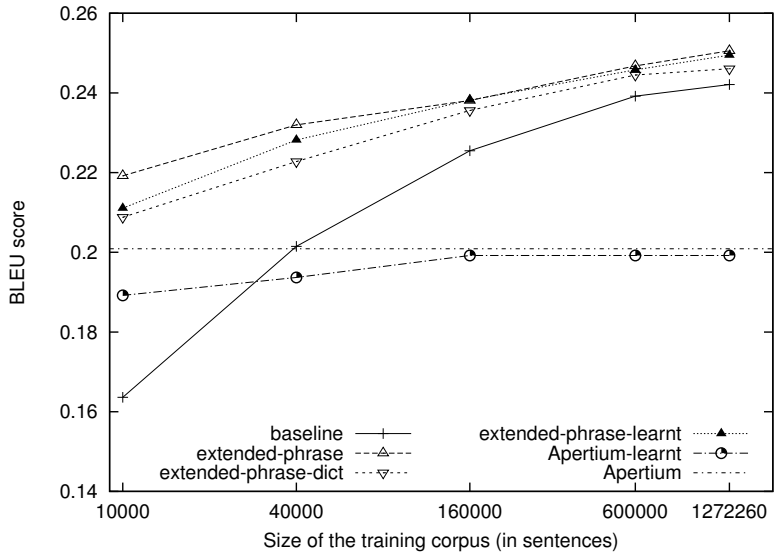
system	10 000	40 000	160 000	600 000	1 272 260
metric	B T M	B T M	B T M	B T M	B T M
baseline	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	= = =	= = =
Apertium	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Apertium-learnt	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase-dict	= = =	= ⇓ =	⇑ = =	= ⇓ =	⇓ ⇓ =
extended-phrase	⇓ = ⇓	= = ⇓	= = =	= = =	= = =

(b) Paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between *extended-phrase-learnt* and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU (B), TER (T) and METEOR (M). ⇑ means that *extended-phrase-learnt* outperforms the reference method by a statistically significant margin, ⇓ means the opposite, and = means that there is no statistically significant difference between them.

Figure 14: For the Spanish→English in-domain evaluation, automatic evaluation scores obtained for the baseline PBSMT system, Apertium with hand-crafted rules (*Apertium-learnt*), Apertium with learnt rules (*Apertium-learnt*), and our hybrid approach (described in Section 3.2.3) using hand-crafted shallow-transfer rules (*extended-phrase*), a set of rules inferred from the training corpus (*extended-phrase-learnt*) and no rules at all (*extended-phrase-dict*). The table shows a pair-wise comparison with the system *extended-phrase-learnt*.

hand-crafted rules outperforms the hybrid system with dictionaries, the translation quality achieved by the hybrid system with automatically inferred rules (*extended-phrase-learnt*) lies in-between.

§3). Nevertheless, the error analysis described in Section 6.2 shows that the automatically inferred rules contain many exceptions applied to particular words that are not included in the hand-crafted ones.



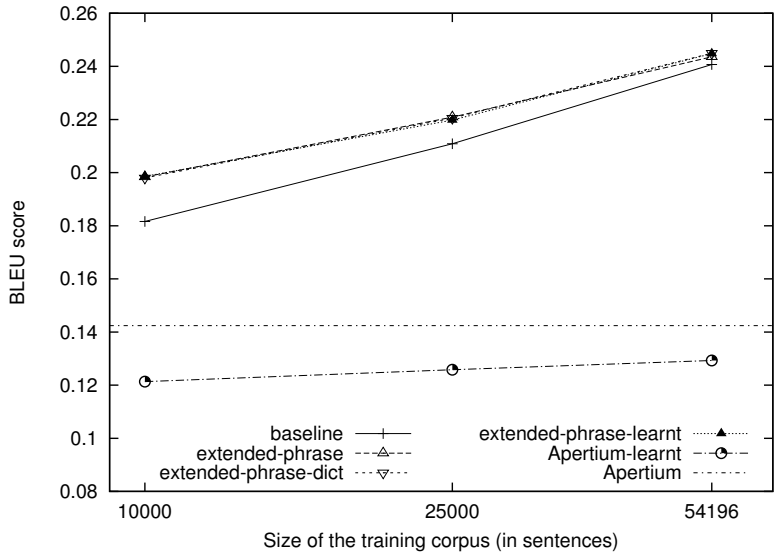
(a) BLEU scores.

system	10 000	40 000	160 000	600 000	1 272 260
metric	B T M	B T M	B T M	B T M	B T M
baseline	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Apertium	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Apertium-learnt	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase-dict	= ⇓ ⇑	⇑ = ⇑	= = ⇑	= = =	⇑ = ⇑
extended-phrase	⇓ ⇓ ⇓	⇓ ⇓ =	= = =	= = ⇓	= ⇓ =

(b) Paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between *extended-phrase-learnt* and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU (B), TER (T) and METEOR (M). ⇑ means that *extended-phrase-learnt* outperforms the reference method by a statistically significant margin, ⇓ means the opposite, and = means that there is no statistically significant difference between them.

Figure 15: For the Spanish→English out-of-domain evaluation, automatic evaluation scores obtained for the baseline PBSMT system, Apertium with hand-crafted rules (*Apertium-learnt*), Apertium with learnt rules (*Apertium-learnt*), and our hybrid approach (described in Section 3.2.3) using hand-crafted shallow-transfer rules (*extended-phrase*), a set of rules inferred from the training corpus (*extended-phrase-learnt*) and no rules at all (*extended-phrase-dict*). The table shows a pair-wise comparison with the system *extended-phrase-learnt*.

In addition, it is worth noting that the translation quality of the approach *extended-phrase-learnt* does not drop when the size of the training corpus exceeds 160 000 sentences and the full training corpus is not used for rule inference. In fact, under these circumstances (600 000 parallel sentences) there are not significant differences between the use of automatically inferred rules and hand-crafted rules in hybrid systems (English→Spanish, out-of-domain evaluation). This observation is probably related to the fact that the trans-



(a) BLEU scores.

system	10 000	25 000	54 196
metric	B T M	B T M	B T M
baseline	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Apertium	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
Apertium-learnt	⇑ ⇑ ⇑	⇑ ⇑ ⇑	⇑ ⇑ ⇑
extended-phrase-dict	= = ⇓	= ⇓ =	= = =
extended-phrase	= = =	= ⇓ =	= = =

(b) Paired bootstrap resampling comparison ($p \leq 0.05$; 1 000 iterations) between *extended-phrase-learnt* and the other methods being evaluated (a method per row). Columns represent training corpus sizes and evaluation metrics: BLEU (B), TER (T) and METEOR (M). ⇑ means that *extended-phrase-learnt* outperforms the reference method by a statistically significant margin, ⇓ means the opposite, and = means that there is no statistically significant difference between them.

Figure 16: For the Breton→French in-domain evaluation, automatic evaluation scores obtained for the baseline PBSMT system, Apertium with hand-crafted rules (*Apertium-learnt*), Apertium with learnt rules (*Apertium-learnt*), and our hybrid approach (described in Section 3.2.3) using hand-crafted shallow-transfer rules (*extended-phrase*), a set of rules inferred from the training corpus (*extended-phrase-learnt*) and no rules at all (*extended-phrase-dict*). The table shows a pair-wise comparison with the system *extended-phrase-learnt*.

lation performance of the automatically inferred rules grows very slowly with the size of the training corpus, and the rules obtained from bigger parallel corpora would probably be similar to those obtained from the fragment of 160 000 sentences. Nevertheless, the exact impact of the proportion of the training corpus used for rule inference for different training corpus sizes, language pairs and domains merits further research.

Finally, it is also worth noting the difference between the hand-crafted rules (*Apertium*) and the automatically inferred rules (*Apertium-learnt*) when they are used in an RBMT system: in some cases (Breton→French and English→Spanish out-of-domain evaluation) the difference in translation performance is considerably higher than the difference between the hybrid systems enriched with hand-crafted rules and with automatically inferred rules (see figures 13 and 16). This occurs because in RBMT the translation is completely led by the shallow-transfer rules, and the possible errors encoded in the automatically inferred rules have a direct impact on the output.

6. Human Evaluation and Error Analysis

This section reports, on the one hand, the results obtained on an out-of-domain human evaluation performed for English→Spanish when the largest training parallel corpus is used, and, on the other, an analysis of the translation errors performed by the different systems evaluated in Section 5.

6.1 Human Evaluation

In order to confirm the results obtained with automatic evaluation metrics, we have performed a human evaluation for English→Spanish and out-of-domain texts. The systems included in this human evaluation were those described in the previous section and trained on the largest parallel corpus used.

We asked 15 users to rank (allowing ties) the translations produced by the baseline PBSMT system (*baseline*), Apertium with hand-crafted rules, our hybrid approach using only dictionaries (*extended-phrase-dict*), our hybrid approach using automatically inferred rules (*extended-phrase-learnt*) and our hybrid approach using hand-crafted rules (*extended-phrase*). Each user ranked the translations of 50 SL sentences from the test set. The users were split in 5 groups, and the users in each group ranked exactly the same set of SL sentences, thus allowing us to compute inter-annotator agreement. In total, the translations of 250 sentences from the test set were ranked. This evaluation method is similar to that followed in the WMT 2012 shared translation task (Callison-Burch et al., 2012).

We computed the ratio of wins for each system (Callison-Burch et al., 2012, Eq. 4) as the proportion of times each system was ranked better than any other system. This score allows us to sort the systems from best to worst, as is shown in the last row of Table 2. The resulting ordering is exactly the same as that obtained with automatic evaluation metrics (see Figure 13).

Table 2 also shows the results of the pairwise comparison between the systems: each cell represents the proportion of sentences for which the system named after the row label outperforms the system named after the column label. A score shown in bold type means that the difference is statistically significant.[35] These results entirely confirm the results obtained with automatic evaluation measures: hybrid systems outperform both RBMT and PBSMT systems, and the automatically inferred rule allow us to build better hybrid sys-

35. According to the Sign Test, for $p \leq 0.10$. We chose a relatively high p-value because of the small amount of human rankings available.

	extended-phrase	extended-phrase-l.	extended-phrase-d.	baseline	Apertium
extended-phrase		**0.55**	**0.60**	**0.60**	**0.68**
extended-phrase-l.	0.45		**0.55**	**0.55**	**0.65**
extended-phrase-d.	0.40	0.46		0.49	**0.65**
baseline	0.40	0.45	0.51		**0.63**
Apertium	0.32	0.35	0.35	0.37	
> other	0.61	0.56	0.51	0.51	0.35

Table 2: Results of the human evaluation; *extended-phrase-l.* is an abbreviation for *extended-phrase-learnt* and *extended-phrase-d.* is an abbreviation for *extended-phrase-dict*. The last row represents the proportion of times each system outperforms any other system, while the remaining cells show the results of a pairwise evaluation: they represent the proportion of sentences for which the system named after the row label outperforms the system named after the column label. A score shown in bold type when the system named after the row label wins more often than the system named after the column label means that the difference is statistically significant.

tems using just dictionaries as an external resource, i.e. *extended-phrase-learnt* outperforms *extended-phrase-dict*.

Finally, the inter-annotator agreement computed as described by Callison-Burch et al. (2012, Sec. 3.2) is $\kappa = 0.503$, which is usually interpreted as a fair agreement.

6.2 Error Analysis

In addition to assessing translation quality by means of automatic evaluation metrics and human ranking, it is also interesting to compare the different types of errors made by the systems evaluated in this section. We compared the translations performed by the different systems used in the human evaluation and found interesting trends that we summarise below. We focused the analysis on English→Spanish because it is the language pair for which the rules have the highest impact (see Section 4.2). Table 3 shows seven examples of translations to which we will refer throughout this section.

A comparison between the pure RBMT system Apertium, the *baseline* PBSMT system and the hybrid system *extended-phrase* shows that the two pure systems are complementary and when they are combined, the number of errors is reduced. When comparing the pure statistical system with the hybrid one, a reduction in the number of OOV words is observed (e.g. the word *patterned* in example #1). There are also words whose translation is too specific to the domain of the parliament speeches when it is performed by the pure PBSMT system, but they are translated in a more appropriate way for the news domain by the hybrid system. An example of this is the word *feel* in example #2. The differences between both systems are not just lexical: the hybrid system produces a better agreement between determiners, nouns and adjectives (see example #3)[36] and correctly translates noun phrases made of adjacent nouns (see example #4),[37] among other grammatical improvements.

36. The grammatically correct translation into Spanish of *specialised category* is *categoría especializada*
37. The correct translation into Spanish of the adjacent nouns *Brno socialists* is *socialistas de Brno*; it literally means *socialists of Brno*.

When compared to the Apertium RBMT system, the hybrid system produces more fluent translations in the TL, probably thanks to the use of a TL model. For instance, the hybrid system deals better with sentences that do not have a regular grammatical structure (see the translation of *It should* in example #4).[38] Preposition choices are also generally better in the hybrid system (for instance, the preposition *to* is correctly removed by the hybrid system in example #4), as is the translation of phrasal verbs (see how *closing down* is translated by the different systems in example #5).

The results of the evaluation show that the translation performance of the hybrid system built with automatically inferred rules (*extended-phrase-learnt*) is close to that of the hybrid system built with hand-crafted rules (*extended-phrase*; see Figure 13). A manual inspection of the translations produced reveals that hand-crafted rules and automatically inferred rules do not produce similar translations. On the one hand, automatically inferred rules encode many exceptions to general translation rules, which makes them outperform the hand-crafted ones in the case of some sentences. One common example of this phenomenon is the swapping of the adjective–noun sequence. Some adjectives (prepositive adjectives) must not be swapped when translating them into Spanish and the automatically inferred rules are able to learn this (for instance, the adjective *best* in example #6). On the other hand, hand-crafted rules encode long-range grammatical operations, such as the subject-predicate agreement in example #7 —where *invaded* is translated as *invadieron*, which agrees in person and number with the translation of *Some 150 drivers*—, which could not be automatically inferred because the rule inference algorithm only considers segments of at most 5 tokens.

7. Concluding Remarks

In this paper, a hybridisation approach with which to enrich PBSMT models with the data from shallow-transfer RBMT systems has been presented. It has been confirmed that data from shallow-transfer RBMT can improve PBSMT systems and also that the resulting hybrid system outperforms both pure PBSMT and RBMT systems built from the same data.

Our hybridisation approach overcomes the limitations of the general-purpose strategy that attempts to improve PBSMT models with data from other MT systems (Eisele et al., 2008) thanks to the fact that it takes advantage of the way in which the shallow-transfer RBMT system uses its linguistic resources to segment the SL sentences. The experiments carried out have shown that our hybrid approach outperforms the strategy by Eisele et al. by a statistically significant margin in a wide range of situations. In fact, a system (Sánchez-Cartagena et al., 2011b) built with the hybridisation approach described in this work was one of the winners in the pair-wise manual evaluation of the WMT 2011 shared translation task (Callison-Burch et al., 2011) for Spanish→English.[39] The effectiveness of our hybrid

38. The reference sentence, which is a relatively free translation of the SL sentence, does not contain the word *Debería*, although it is the most appropriate translation of *It should* in that context.

39. This evaluation was performed by asking users to rank the translations produced by the different systems. Users iteratively ranked (from best to worst) the translations of the same SL sentence produced by 5 different systems. We refer the reader to the description of the task by Callison-Burch et al. (2011) for more details about the evaluation. The human evaluation described in Section 6 was carried out in a similar way.

#	system	sentence
1	source	If the inauguration was about thick lace, in Oslo it was **patterned** velvet.
	Apertium	Si la inauguración era sobre encaje grueso, en Oslo sea terciopelo **estampado**.
	baseline	Si la inauguración de centro encaje en Oslo, fue **patterned**.
	extended-phrase	Si la inauguración era sobre encaje grueso, en Oslo, fue terciopelo **estampado**
	reference	Si para la toma de posesión de trataba de un encaje gordo, en Oslo es un terciopelo **con estampado**.
2	source	He **feels** a kick.
	Apertium	**Siente** un chut.
	baseline	Él **considera** una patada.
	extended-phrase	Él **siente** una patada.
	reference	Él **siente** una patada
3	source	The only **specialised category** which nobody won this year was physics.
	Apertium	La **categoría** sólo **especializada** qué nadie ganó este año era física.
	baseline	La única **categoría especializados** que nadie ganó este año ha sido física
	extended-phrase	La única **categoría especializada** que nadie ganó este año ha sido física.
	reference	La única **categoría especializada**, sin una mujer premiada, fue física.
4	source	**It should**, but ecology and aesthetics don't seem **to** interest the **Brno socialists** much.
	Apertium	**Tenga que**, pero la ecología y la estética no parecen **para** interesar los **socialistas de Brno** mucho.
	baseline	**Debería**, pero la ecología y la estética no parecen interés el **Brno socialistas** mucho.
	extended-phrase	**Debería**, pero la ecología y la estética no parecen interés los **socialistas de Brno** mucho.
	reference	Lo que pasa es que la ecología y la estética no parecen interesar demasiado a los **socialistas de Brno**.
5	source	We are opposed on principle to the **closing down** of parties.
	Apertium	Somos opposed encima principio al **encierro abajo** de partidos.
	baseline	Nos oponemos por principio a la **clausura** de partidos.
	extended-phrase	Nos oponemos por principio a la **clausura** de los partidos.
	reference	Por principio nos oponemos a la **clausura** de partidos.
6	source	There's some of the **best skiers** and snow borders in the county here - some real talent," he added.
	extended-phrase-l.	No hay algunos de los **mejores esquiadores** y fronteras de nieve en el condado aquí - un verdadero talento," añadió.
	extended-phrase	No hay algunos de los **esquiadores mejores** y fronteras de nieve en el condado aquí - algunos verdadero talento," añadió.
	reference	Aquí se encuentran algunos de los **mejores esquiadores** y snowboarders del condado, talento verdadero", añadió.
7	source	Some 150 drivers **invaded** a works council meeting [...]
	extended-phrase-l.	Unos 150 conductores **invadido** una reunión del consejo de empresa [...]
	extended-phrase	Unos 150 conductores **invadieron** una reunión de consejo de los trabajos [...]
	reference	Unos 150 conductores **invadieron** un comité [...]

Table 3: Translations into Spanish of different English sentences extracted from the out-of-domain evaluation corpus and produced by the systems evaluated in Section 6. The most remarkable differences are highlighted. *extended-phrase-l.* is the abbreviation of *extended-phrase-learnt*, the hybrid system with automatically inferred rules.

approach is thereby confirmed by both automatic and human evaluation (results in WMT 2011 human evaluation are compatible with those of the human evaluation described in Section 6: in both experiments, a hybrid system built with our method outperforms a pure PBSMT system).

Moreover, it has been proved that the rule inference algorithm presented by Sánchez-Cartagena et al. (2015) can be successfully combined with the hybrid approach, thus allowing a hybrid system to be built using dictionaries as the only hand-crafted linguistic resource. An improvement to translation quality is also achieved in the same way as if hand-crafted shallow-transfer rules had been used. The hybrid system with automatically inferred rules is able to attain the translation quality achieved by a hybrid system with hand-crafted rules and, even when it does not, it often obtains better results than a hybrid system that only uses dictionaries to enrich the PBSMT models. Additionally, the need for a human expert to write the rules is avoided.

According to the results obtained, our hybrid approach is especially recommended when the training parallel corpus (for the translation model) and monolingual corpus (for the language model) have a moderate size and when the domain of the training corpus is different from the domain of the texts to be translated.[40] The use of moderate-sized training corpora may be necessary in order to limit the size of the phrase table and the TL model when the hybrid system must be executed in a mobile device with limited memory. Moreover, the hybrid approach presented in this work can also be safely applied in other scenarios, since drops in translation quality in comparison with a PBSMT baseline have not been detected. If good enough hand-crafted rules are available, it is worth using them instead of inferring rules from the parallel training corpus, but if they are not, applying the rule inference algorithm will not significantly degrade translation quality.[41]

The hybridisation method described in this paper is implemented in a software tool called `rule2Phrase` (Sánchez-Cartagena, Sánchez-Martínez, & Pérez-Ortiz, 2012) that has been released under the GNU GPL v3 free software license. Its source code can be freely downloaded from `http://www.dlsi.ua.es/~vmsanchez/Rule2Phrase.tar.gz`. The tool includes the phrase scoring strategies that have been described in sections 3.2.3 and 3.2.4 of this paper.

40. The English↔Spanish out-of-domain evaluations described in Section 5 were repeated using a TL model estimated from much bigger monolingual corpora. In particular, a portion of the *News Crawl* monolingual corpus provided for the WMT 2011 shared translation task (`http://www.statmt.org/wmt11/translation-task.html`) was concatenated to the Europarl corpus. As a result, English and Spanish monolingual corpora with around 6 200 000 sentences each were obtained. The results of the evaluation showed that when a parallel corpus that contains around 26 000 000 words is used together with these monolingual corpora, the difference between the hybrid system built with automatically inferred rules and the baseline SMT system is not statistically significant for some of the evaluation metrics.

41. Manually creating transfer rules involves a huge human effort. Rule writers must first identify the grammatical divergences between the languages involved that need to be treated by rules and sort them by frequency in the texts that will be translated by the RBMT system. This operation is called *contrastive* analysis. They then write the rules to deal with these divergencies, starting with the most frequent ones and choosing them in case of possible conflicts between rules. Rules written by humans may not be good enough if there are grammatical divergencies not identified during the contrastive analysis, their frequency has not been correctly estimated or enough time has not been invested in writing rules for dealing with the most important grammatical divergencies.

Acknowledgments

Research funded 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).

References

Banerjee, S., & Lavie, A. (2005). 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*, pp. 65–72, Ann Arbor, Michigan, USA.

Bisazza, A., Ruiz, N., & Federico, M. (2011). Fill-up versus interpolation methods for phrase-based smt adaptation. In *Proceedings of the 8th International Workshop on Spoken Language Translation*, San Francisco, California, USA.

Bojar, O., & Hajič, J. (2008). Phrase-based and deep syntactic English-to-Czech statistical machine translation. In *Proceedings of the third Workshop on Statistical Machine translation*, pp. 143–146, Columbus, Ohio, USA.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Goldsmith, M. J., Hajic, J., Mercer, R. L., & Mohanty, S. (1993). But dictionaries are data too. In *Proceedings of the Workshop on Human Language Technology*, pp. 202–205, Princeton, New Jersey.

Callison-Burch, C., Koehn, P., Monz, C., & Zaidan, O. (2011). Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 22–64, Edinburgh, Scotland.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., & Specia, L. (2012). Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pp. 10–51, Montreal, Canada.

Carl, M. (2007). METIS-II: The German to English MT system. In *Proceedings of the XI Machine Translation Summit*, pp. 65–73, Copenhagen, Denmark.

Crego, J. (2014). SYSTRAN RBMT engine: hybridization experiments. In *3rd Workshop on Hybrid Approaches to Machine Translation (HyTra)*, Gothenburg, Sweden.

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. (1992). A practical part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pp. 133–140, Trento, Italy.

Eisele, A., Federmann, C., Saint-Amand, H., Jellinghaus, M., Herrmann, T., & Chen, Y. (2008). Using Moses to integrate multiple rule-based machine translation engines into a hybrid system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pp. 179–182, Columbus, Ohio, USA.

Enache, R., España-Bonet, C., Ranta, A., & Màrquez, L. (2012). A hybrid system for patent translation. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pp. 269–276, Trento, Italy.

Federmann, C., Eisele, A., Uszkoreit, H., Chen, Y., Hunsicker, S., & Xu, J. (2010). Further experiments with shallow hybrid mt systems. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics*, pp. 77–81, Uppsala, Sweden.

Federmann, C., & Hunsicker, S. (2011). Stochastic parse tree selection for an existing rbmt system. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 351–357, Edinburgh, Scotland.

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

Gao, Q., Lewis, W., Quirk, C., & Hwang, M.-Y. (2011). Incremental Training and Intentional Over-fitting of Word Alignment. In *Proceedings of the XIII Machine Translation Summit*, pp. 106–113, Xiamen, China.

Goodman, J., & Chen, S. F. (1998). An empirical study of smoothing techniques for language modeling. Tech. rep. TR-10-98, Harvard University.

Graham, Y., & van Genabith, J. (2010). Factor templates for factored machine translation models. In *Proceedings of the 7th International Workshop on Spoken Language Translation*, pp. 275–283, Paris, France.

Green, S., & DeNero, J. (2012). A class-based agreement model for generating accurately inflected translations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, pp. 146–155, Jeju Island, Korea.

Hutchins, W. J., & Somers, H. L. (1992). *An introduction to machine translation*, Vol. 362. Academic Press New York.

Kirchhoff, K., & Yang, M. (2005). Improved language modeling for statistical machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pp. 125–128, Ann Arbor, Michigan, USA.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Vol. 4, pp. 388–395, Barcelona, Spain.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the X Machine Translation Summit*, pp. 12–16, Phuket, Thailand.

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

Koehn, P., & Hoang, H. (2007). Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 868–876, Prague.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pp. 177–180, Prague, Czech Republic.

Koehn, P., & Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 224–227, Prague, Czech Republic.

Labaka, G., España-Bonet, C., Màrquez, L., & Sarasola, K. (2014). A hybrid machine translation architecture guided by syntax. *Machine Translation*, *28*(2), 91–125.

Lavie, A. (2008). Stat-XFER: A General Search-Based Syntax-Driven Framework for Machine Translation. In Gelbukh, A. (Ed.), *Computational Linguistics and Intelligent Text Processing*, Vol. 4919 of *Lecture Notes in Computer Science*, pp. 362–375.

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

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

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 160–167, Sapporo, Japan.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA.

Popovic, M., & Ney, H. (2006). Statistical machine translation with a small amount of bilingual training data. In *5th LREC SALTMIL Workshop on Minority Languages*, p. 25–29, Genoa, Italy.

Riezler, S., King, T. H., Kaplan, R. M., Crouch, R., Maxwell, III, J. T., & Johnson, M. (2002). Parsing the wall street journal using a lexical-functional grammar and discriminative estimation techniques. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 271–278, Stroudsburg, PA, USA. Association for Computational Linguistics.

Riezler, S., & Maxwell III, J. T. (2006). Grammatical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pp. 248–255, New York City, New York, USA.

Roche, E., & Schabes, Y. (1997). Introduction. In Roche, E., & Schabes, Y. (Eds.), *Finite-state language processing*, pp. 1–65. MIT, Cambridge, Massachusetts, USA.

Rosa, R., Mareček, D., & Dušek, O. (2012). Depfix: A system for automatic correction of czech mt outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pp. 362–368, Montreal, Canada.

Rosti, A.-V., Matsoukas, S., & Schwartz, R. (2007). Improved word-level system combination for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 312–319, Prague, Czech Republic.

Sánchez-Cartagena, V. M., Pérez-Ortiz, J. A., & Sánchez-Martínez, F. (2015). A generalised alignment template formalism and its application to the inference of shallow-transfer

machine translation rules from scarce bilingual corpora. *Computer Speech & Language*, *32*(1), 46–90.

Sánchez-Cartagena, V. M., Sánchez-Martínez, F., & Pérez-Ortiz, J. A. (2011a). Integrating shallow-transfer rules into phrase-based statistical machine translation. In *Proceedings of the XIII Machine Translation Summit*, pp. 562–569, Xiamen, China.

Sánchez-Cartagena, V. M., Sánchez-Martínez, F., & Pérez-Ortiz, J. A. (2011b). The Universitat d'Alacant hybrid machine translation system for WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 457–463, Edinburgh, Scotland.

Sánchez-Cartagena, V. M., Sánchez-Martínez, F., & Pérez-Ortiz, J. A. (2012). An open-source toolkit for integrating shallow-transfer rules into phrase-based satistical machine translation. In *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation*, pp. 41–54, Gothenburg, Sweden.

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

Schwenk, H., Abdul-Rauf, S., Barrault, L., & Senellart, J. (2009). SMT and SPE machine translation systems for WMT'09. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pp. 130–134, Athens, Greece.

Sennrich, R. (2012). Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 539–549, Avignon, France.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th biennial conference of the Association for Machine Translation in the Americas*, pp. 223–231, Cambridge, Massachusetts, USA.

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing*, pp. 901–904, Denver, Colorado, USA.

Thurmair, G. (2009). Comparing different architectures of hybrid Machine Translation systems. In *Proceedings of the XII Machine Translation Summit*, Ottawa, Canada.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pp. 2214–2218, Istanbul, Turkey.

Tyers, F. M. (2009). 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*, pp. 213–217, Barcelona, Spain.

Zbib, R., Kayser, M., Matsoukas, S., Makhoul, J., Nader, H., Soliman, H., & Safadi, R. (2012). Methods for integrating rule-based and statistical systems for Arabic to English machine translation. *Machine Translation*, *26*(1-2), 67–83.