Enriching a Statistical Machine Translation System Trained on Small Parallel Corpora with Rule-Based Bilingual Phrases

Víctor M. Sánchez-Cartagena, Felipe Sánchez-Martínez, Juan Antonio Pérez-Ortiz
Transducens Research Group
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071, Alacant, Spain
{vmsanchez,fsanchez,japerez}@dlsi.ua.es

Abstract

In this paper, we present a new hybridisation approach consisting of enriching the phrase table of a phrase-based statistical machine translation system with bilingual phrase pairs matching structural transfer rules and dictionary entries from a shallow-transfer rule-based machine translation system. We have tested this approach on different small parallel corpora scenarios, where pure statistical machine translation systems suffer from data sparseness. The results obtained show an improvement in translation quality, specially when translating out-of-domain texts that are well covered by the shallow-transfer rule-based machine translation system we have used.

1 Introduction

Statistical machine translation (SMT) (Koehn, 2010) is currently the leading paradigm in machine translation research. SMT systems are very attractive because they may be built with little human effort when enough monolingual and bilingual corpora are available. However, bilingual corpora large enough to build competitive SMT systems are not always easy to harvest, and they may not even exist for some language pairs. On the contrary, rule-based machine translation systems (RBMT) 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.

When both parallel corpora and linguistic information exist, hybrid approaches (Thurmair, 2009) may be followed in order to make the most of such resources. We focus on alleviating the data sparseness problem suffered by phrase-based statistical machine translation (PBSMT) systems (Koehn, 2010, ch. 5) when trained on small parallel corpora. We present a new hybrid approach which enriches a PBSMT system with resources from shallow-transfer RBMT. Shallow-transfer RBMT systems, which are described in detail below, do not perform a complete syntactic analysis of the input sentences, but rather work with much simpler intermediate representations. Hybridisation between shallow-transfer RBMT and SMT has not yet been explored. Existing hybridisation strategies involve more complex RBMT systems (Eisele et al., 2008) which are usually treated as black boxes; in contrast, our approach directly uses the RBMT dictionaries and rules.

We provide an exhaustive evaluation of our hybridisation approach with two different language pairs: Breton–French and Spanish–English. While the first one suffers from actual resource scarceness, many different parallel corpora are available for the second one, which allows us to test our approach on different domains and check if it is able to improve the poor performance of PBSMT systems when translating texts from a domain not covered by the bilingual training data.

The rest of the paper is organised as follows. Next section overviews the two systems we combine in our approach. Then, section 3 outlines related hybrid approaches, whereas our approach is described in section 4. Sections 5 and 6 present the experiments conducted and discuss the results achieved, respectively. The paper ends with our conclusions and future research lines.

2 Translation Approaches

2.1 Phrase-Based Statistical Machine Translation

PBSMT systems (Koehn, 2010, ch. 5) translate sentences by maximising the translation probability as defined by the log-linear combination of a number of feature functions, whose weights are chosen to optimise translation quality (Och, 2003). A core component of every PBSMT system is the phrase table, which contains bilingual phrase pairs extracted from a bilingual corpus after word alignment (Och and Ney, 2003). The set of translations

from which the most probable one is chosen is built by segmenting the source sentence in all possible ways and then combining the translation of the different source segments according to the phrase table. Common feature functions are: source-to-target and target-to-source phrase translation probabilities, source-to-target and target-to-source lexical weightings (calculated by using a probabilistic bilingual dictionary), reordering costs, number of words in the output (word penalty), number of phrase pairs used (phrase penalty), and likelihood of the output as given by a target-language model.

2.2 Shallow-Transfer Rule-Based Machine Translation

The RBMT process (Hutchins and Somers, 1992) can be split into three different steps: analysis of the source language (SL) text to build a SL intermediate representation; transfer from that SL intermediate representation to a target language (TL) intermediate representation; and generation of the final translation from the TL intermediate representation.

Shallow-transfer RBMT systems use relatively simple intermediate representations, which are based on lexical forms consisting of lemma, part of speech and morphological inflection information of the words in the input sentence, and simple shallow-transfer rules that operate on sequences of lexical forms: this kind of systems do not perform a complete syntactic analysis. Apertium (Forcada et al., 2011), the shallow-transfer RBMT platform used to evaluate our approach, splits the transfer stage into structural and lexical transfer. The lexical transfer is done by using a bilingual dictionary which, for each SL lexical form, provides a single TL lexical form; thus, no lexical selection is performed. It is worth noting that multi-word expressions, such as *on the other hand* (which acts as a single adverb), may be analysed to (or generated from) a single lexical form.

Structural transfer is done by applying a set of rules in a left-to-right, longest-match fashion to prevent the translation to be performed word for word in those cases in which this would result in an incorrect translation. Structural transfer rules process sequences of lexical forms by performing operations such as reorderings and gender and number agreements. For the translation between non-related language pairs, the structural transfer may be split into three levels in order to facilitate the writing of rules by linguists. The first level performs short-distance operations (such as gender and number agreement between nouns and adjectives) and groups word sequences into *chunks*; the second one performs inter *chunk* operations; and the third one generates a sequence of lexical forms from each *chunk*. Note that, although this multi-stage shallow transfer allows performing operations between words which are distant in the source sentence, shallow-transfer RBMT systems are less powerful that the ones which perform full parsing.

3 Related Work

Bilingual dictionaries are the most 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 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 appear in the SL sentences are translated as such by the SMT decoder because they may be split into smaller units by the phrase-extraction algorithm. Our strategy differs from these approaches in that we ensure the proper translation of multi-word expressions, but also add the dictionary entries to the training corpus with the aim of improving word alignment. Other approaches go beyond adding a dictionary to the parallel corpus: dictionary entries may constrain the decoding process (Langlais, 2002), or may be used in conjunction with handcrafted rules to reorder the SL sentences to match the structure of the TL (Popović and Ney, 2006).

Although RBMT transfer rules have also been reused in hybrid systems, they have been mostly used implicitly as part of a complete RBMT engine. For instance, Dugast et al. (2008) show how a PBSMT system can be bootstrapped using only monolingual data and an RBMT engine. Another remarkable study (Eisele et al., 2008) presents a strategy based on the augmentation of the phrase table to include information provided by an RBMT system. In this approach, the sentences to be translated by the hybrid system are first translated with an RBMT system and then a small phrase table is obtained from the resulting parallel corpus. Phrase pairs are extracted following the usual procedure (Koehn, 2010, sec. 5.2.3) which generates the set of all possible phrase pairs that are consistent with the

word alignments. In order to obtain reliable word alignments, they are computed using an alignment model previously built from a large parallel corpus. Finally, the RBMT-generated phrase table is added to the original one. On the contrary, our approach directly generates phrase pairs which match either an entry in the bilingual dictionary or a structural transfer rule; thus preventing them from being split into smaller phrase pairs even if they would be consistent with the word alignments. In addition, our approach does not require a large parallel corpus from which to learn an alignment model. Preliminary experiments show that our hybrid approach outperforms Eisele et al.'s (2008) strategy when translating from Spanish to English.

Other strategies involving neither transfer rules nor bilingual dictionaries may alleviate the data sparseness problem in PBSMT. For example, paraphrases may be derived from a SL monolingual corpus (Marton et al., 2009) and verb forms may be substituted by their lemma when translating into highly-inflected languages (de Gispert et al., 2005).

4 Enhancing Phrase-Based SMT With Shallow-Transfer Linguistic Resources

Our hybridisation strategy modifies two elements of a standard PBSMT system: the word alignments and the phrase translation model.

4.1 Improving Word Alignment with RBMT Bilingual Dictionaries

As improving the quality of the word alignments in a PBSMT system could lead to improvements in translation performance (Lopez and Resnik, 2006), in our approach we add to the original corpus all the entries, after suitably inflecting them, from the Apertium bilingual dictionary, to help the word aligment process. Recall that some multi-word expressions are encoded as single lexical forms in the Apertium dictionaries; therefore, the entries generated from Apertium may contain multi-word parallel segments. Once word alignments have been computed and the probabilistic bilingual dictionary used to compute the lexical weightings of the phrase pairs has been learned, dictionary entries are ignored and no phrase pair are extracted from them. In contrast to Schwenk et al. (2009), we avoid extracting phrase pairs which do not preserve the translation of multi-word expressions as such by including the dictionary entries directly in the phrase table, as discussed next.

4.2 Enriching the Phrase Translation Model

As already mentioned, the Apertium structural transfer detects sequences of lexical forms which need to be translated together to prevent them from being translated word for word, which would result in an incorrect translation. Therefore, adding to the phrase table of a PBSMT system all the bilingual phrase pairs which either match one of these sequences of lexical forms in the structural transfer or an entry in the bilingual dictionary ensures that all the linguistic information of Apertium is encoded with the minimum amount of phrase pairs.

4.2.1 Phrase Pair Generation

Generating a phrase pair from every entry in the bilingual dictionary is straightforward: it only involves the inflection of source and target lexical forms. The generation of phrase pairs from the structural transfer rules is performed by finding sequences of SL words in the sentences to be translated that match a structural transfer rule. Each of these sequences constitute the SL side of a bilingual phrase pair; the corresponding TL phrase is obtained by translating the SL side with Apertium.

It is worth noting that the generation of bilingual phrase pairs from the shallow-transfer rules is guided by the test corpus. We decided to do it in this way in order to avoid meaningless phrases and also to make our approach computationally feasible. Consider, for instance, a rule which is triggered every time a determiner followed by a noun and an adjective is detected. Generating phrase pairs from this rule would involve combining all the determiners in the dictionary with all the nouns and all the adjectives, causing the generation of many meaningless phrases, such as *el niño inalámbrico – the wireless boy*. In addition, the number of combinations to deal with would become unmanageable as the length of the rule grows.

4.2.2 Scoring the New Phrase Pairs

State-of-the-art PBSMT systems usually attach 5 scores to every phrase pair in the translation table: source-to-target and target-to-source phrase translation probabilities, source-to-target and target-to-source lexical weightings, and phrase penalty.

To calculate the phrase translation probabilities of the new phrase pairs obtained from the shallow-transfer RBMT resources we simply add them once to the list of corpus-extracted phrase pairs, and then compute the probabilities by relative frequency as it is usually done (Koehn, 2010, sec. 5.2.5). In this regard, it is worth noting that as RBMT-generated phrase pairs are added only once, if one of them

happens to share its source side with many other corpus-extracted phrase pairs, or even with a single, very frequent one, the RBMT-generated phrase pair will receive lower scores, which penalises its use. To alleviate this without adding the same phrase pair an arbitrary amount of times, we introduce an additional boolean score to flag phrase pairs obtained from the RBMT resources.

To calculate the lexical weightings (Koehn, 2010, sec. 5.3.3) of the RBMT-generated phrase pairs the alignments between the words in the source side and those in the target side are needed. They are computed by tracing the operations carried out in the different stages of the shallow-transfer RBMT system. Only those words which are neither split nor joint with other words by the RBMT engine are included in the alignments; thus, multi-word expressions are left unaligned. This is done for convenience since, in this way, the number of lexical probabilities to take into account is reduced, and, as a result, phrase pairs containing multi-word expressions receive higher scores.

5 Experimental Settings

We evaluated our RBMT–SMT hybridisation approach on two different language pairs, namely Breton–French and Spanish–English, and with different small training corpus sizes. While the Breton–French language pair suffers from actual resource scarceness (there are only around 30 000 parallel sentences available), Spanish–English was chosen because it has a wide range of parallel corpora available, which allows us to perform both in-domain and out-of-domain evaluations.

SMT systems for Spanish–English were trained from the Europarl v5 parallel corpus (Koehn, 2005), collected from the proceedings of the European Parliament. Its whole target side, except for the Q4/2000 portion, was used to train the TL model used in the experiments. We learned the translation model from corpora of different sizes; more precisely, we used fragments of the Europarl corpus consisting of 2 000, 5 000, 10 000, 20 000, 40 000 and 80 000 parallel sentences. The sentences in each training set were randomly chosen (avoiding the Q4/2000 portion) in such a way that larger corpora include the sentences in the smaller ones.

Regarding Breton–French, the translation model was built using the only freely-available parallel corpus for such language pair (Tyers, 2009), which contains short sentences from the tourism and computer localisation domains split in different sections for training, tuning and testing. We also used different training corpora sizes, namely 2 000, 5 000, 10 000, 20 000, and 26 835 parallel sentences, the last one corresponding to the whole training section of the corpus. As in the Spanish–English pair, sentences were randomly chosen and larger corpora include the sentences in the smaller ones. The TL model was learnt from a monolingual corpus built by concatenating the target side of the whole bilingual training corpus and the French monolingual data from the Europarl corpus provided for the WMT 2011 shared translation task.[1]

Corpus		Origin	Sentences
Language model		Europarl, Tyers (2009)	1 975 773
Training	$2k$	Tyers (2009)	2 000
	$5k$	Tyers (2009)	5 000
	$10k$	Tyers (2009)	10 000
	$20k$	Tyers (2009)	20 000
	$\approx 27k$	Tyers (2009)	26 835
In-domain tuning		Tyers (2009)	2 000
In-domain test		Tyers (2009)	2 000

Table 1: Description of the Breton–French parallel corpora used in the experiments.

The weights of the different feature functions were optimised by means of minimum error rate training (MERT; Och, 2003). Breton–French systems were tuned using the *tuning* section of the parallel corpus by Tyers (2009) and evaluated using the *devtest* section of the same corpus. Note that we can only perform in-domain evaluation for this language pair.

Regarding Spanish–English, we have carried out both in-domain and out-of-domain evaluations. 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 corpus; special care was taken to avoid the overlapping between the test and development 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.[2] Tables 1 and 2 summarise the data about the corpora used in the experiments.

We used the free/open-source PBSMT system Moses[3] (Koehn et al., 2007) together with the

[1] http://www.statmt.org/wmt11/translation-task.html

[2] http://www.statmt.org/wmt10/translation-task.html

[3] Revision 3739, downloaded from https://mosesdecoder.svn.sourceforge.net/svnroot/mosesdecoder/trunk.

Corpus		Origin	Sentences
Language model		Europarl	1 650 152
Training	$2k$	Europarl	2 000
	$5k$	Europarl	5 000
	$10k$	Europarl	10 000
	$20k$	Europarl	20 000
	$40k$	Europarl	40 000
	$80k$	Europarl	80 000
In-domain tuning		Europarl	2 000
In-domain test		Europarl	2 000
Out-of-domain tuning		WMT 2010	2 051
Out-of-domain test		WMT 2010	2 489

Table 2: Description of the Spanish–English parallel corpora used in the experiments.

SRILM language modeling toolkit (Stolcke, 2002), which was used to train a 5-gram language model using interpolated Kneser-Ney discounting (Goodman and Chen, 1998). Word alignments from the training parallel corpus were computed by means of GIZA++ (Och and Ney, 2003). The Apertium (Forcada et al., 2011) engine and the linguistic resources for Spanish–English and Breton–French were downloaded from the Apertium Subversion repository.[4] The Apertium linguistic data contains 326 228 entries in the bilingual dictionary, 106 first-level rules, 31 second-level rules, and 7 third-level rules for Spanish–English; and 21 593, 169, 79 and 6, respectively, for Breton–French (see section 2.2 for a description of the different rule levels).

We have tested the following configurations:

- a state-of-the-art PBSMT system with the feature functions discussed in section 2.1 (*baseline*);
- the Apertium shallow-transfer RBMT engine, from which the dictionaries and transfer rules have been taken (*Apertium*);
- the hybridisation approach described along this paper (*phrase-rules*) and a variation in which only dictionary-matching bilingual phrases are included in the phrase table (*phrase-dict*); and
- a reduced version of our approach in which the entries in the bilingual dictionary are only added to the training corpus for the computation of the word alignments and the probabilistic bilingual dictionary, as explained in section 4.1 (*alignment*).

6 Results and Discussion

Table 3 reports the translation performance as measured by BLEU (Papineni et al., 2002) for the different configurations and language pairs described in section 5. Statistical significance of the differences between systems has been computed by performing 1 000 iterations of paired bootstrap resampling (Zhang et al., 2004) with a p-level of 0.05. In addition, table 4 presents the optimal weight obtained with MERT for the feature function that flags whether a phrase pair has been obtained from the Apertium bilingual resources (dictionaries and rules). Table 5 shows the proportion of RBMT-generated phrases used to perform each translation.

The results show that our hybrid approach outperforms both pure RBMT and PBSMT systems in terms of BLEU. However, the difference is statistically significant only under certain circumstances. The in-domain evaluation shows that the statistical significance only holds in the smallest corpus scenarios (i.e., when the training corpus contains at most 40 000 sentences for Spanish–English, and for all the training corpus sizes except 20 000 for Breton–French[5]), and the difference between the baseline PBSMT system and our hybrid approach is reduced as the parallel training corpus grows. Apertium data has been developed bearing in mind the translation of general texts (mainly news) whereas the in-domain test sets come from the specialised domains of parliament speeches (Spanish–English) or tourism and computer localisation (Breton–French). Thus, as soon as the PBSMT system learns reliable information from the parallel corpus, Apertium phrases become useless. On the contrary, the out-of-domain Spanish–English tests, performed on a general (news) domain, show a statistically-significant improvement with all the training corpus sizes tested. In this case, Apertium-generated phrases, which contain handcrafted knowledge from a general domain, cover more sequences of words in the input text which are not covered, or are sparsely found, in the original training corpora. The data reported in tables 4 and 5 support these hypotheses; in the in-domain evaluation, the proportion of phrases generated from Apertium included in the translations drops abruptly as the corpus grows. On the contrary, when evaluating the Spanish–English systems in a different domain, the proportion of Apertium-

[4] Revisions 24177, 22150 and 28674, respectively.

[5] Our results do not agree with those by Tyers (2009), who reported a substantial improvement in BLEU when adding dictionaries to the training corpus. In a personal communication, the author stated that in Tyers (2009) a baseline in which the feature weights were optimised with MERT was compared to a system enriched with the Apertium dictionaries using the default (not optimised) feature weights. Incidentally, not optimising the feature weights provided better results. If the feature weights are optimised in both cases the results obtained are in the line of those reported in this paper.

		In-domain							Out-of-domain						
		$2k$	$5k$	$10k$	$20k$	$\approx 27k$	$40k$	$80k$	$2k$	$5k$	$10k$	$20k$	$\approx 27k$	$40k$	$80k$
es-en	baseline	20.74	24.24	26.46	28.45	-	29.86	30.88	12.59	14.90	16.92	18.63	-	20.32	21.80
	alignment	19.31	23.71	25.89	28.10	-	29.73	30.83	12.06	14.55	16.88	18.66	-	20.34	21.68
	phrase-dict	**24.29**	**26.39**	**27.93**	**29.30**	-	**30.36**	31.14	19.76	20.48	**21.26**	**21.89**	-	**22.67**	**23.20**
	phrase-rules	**24.68**	**26.81**	**28.28**	**29.40**	-	**30.41**	31.02	**20.97**	**21.36**	**22.20**	**22.77**	-	**23.29**	**23.76**
	Apertium	18.00							20.30						
br-fr	baseline	18.86	24.17	28.26	33.17	34.69	-	-	-	-	-	-	-	-	-
	alignment	17.53	23.56	27.82	32.17	34.76	-	-	-	-	-	-	-	-	-
	phrase-dict	**21.57**	**26.39**	**29.66**	33.42	35.50	-	-	-	-	-	-	-	-	-
	phrase-rules	**22.67**	**26.42**	**29.60**	33.14	**35.83**	-	-	-	-	-	-	-	-	-
	Apertium	17.56							-						

Table 3: BLEU score achieved by the different configurations listed in section 5. Hybrid system scores in bold mean that they outperform both Apertium and the PBSMT baseline, and that the improvement is statistically significant. The score of the hybrid system built with the Apertium rules and dictionaries is underlined if it outperforms its dictionary-based counterpart by a statistically significant margin. The $\approx 27k$ corpus size is only tested with the Breton–French language pair because it corresponds to the full Breton–French training corpus size.

generated phrases is higher and falls smoothly, and the value of the feature function is higher than in the in-domain tests.

The inclusion of shallow-transfer rules provides a statistically-significant improvement over the dictionaries for all the training corpus sizes in the Spanish–English out-of-domain evaluation scenario and for the smallest ones in the in-domain tests. That is, shallow-transfer rules are effective when the decoder chooses a high proportion of Apertium-generated phrase pairs.

Finally, the addition of the bilingual dictionaries to the training corpus before the computation of the word alignments and the probabilistic bilingual dictionary results in a small performance drop. It remains to be studied whether the dictionaries improve alignments but not translation performance.

7 Conclusions and Future Work

In this paper we have described a new hybridisation approach consisting of enriching a PBSMT system by adding to its phrase table bilingual phrase pairs matching structural transfer rules and dictionaries from a shallow-transfer RBMT system. The experiments conducted show an improvement of the translation quality when only a small parallel corpus is available. Our approach also helps when training on larger parallel corpora and the texts to translate come from a general (news) domain that is well covered by the RBMT system; in this case, shallow-transfer rules have a greater impact on translation quality than dictionaries.

Our future plans include evaluating the presented hybridisation strategy with more language pairs and bigger training corpora, focusing on test corpora from the news domain, which seems to be the scenario in which our approach better fits. We also plan to further investigate the negative impact that adding the entries in the Apertium bilingual dictionary to the corpus has on translation performance.

Acknowledgments

Work funded by the Spanish Ministry of Science and Innovation through project TIN2009-14009-C02-01 and by Generalitat Valenciana through grant ACIF/2010/174 (VALi+d programme).

References

P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, M. J. Goldsmith, J. Hajic, R. L. Mercer, and S. Mohanty. 1993. But dictionaries are data too. In *Proceedings of the workshop on Human Language Technology*, pages 202–205.

A. de Gispert, J.B. Mariño, and J.M. Crego. 2005. Improving statistical machine translation by classifying and generalizing inflected verb forms. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, pages 3185–3188.

L. Dugast, J. Senellart, and P. Koehn. 2008. Can we Relearn an RBMT System? In *Proceedings of the 3rd Workshop on Statistical Machine Translation*, pages 175–178.

A. Eisele, C. Federmann, H. Saint-Amand, M. Jellinghaus, T. Herrmann, and Y. Chen. 2008. Using Moses to integrate multiple rule-based machine translation engines into a hybrid system. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*, pages 179–182.

M. L. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martnez, G. Ramírez-Sánchez, and F. M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*. doi: 10.1007/s10590-011-9090-0.

		$2k$	$5k$	$10k$	$20k$	$\approx 27k$	$40k$	$80k$
es-en	phrase-dict	0.0025	0.0010	−0.0003	−0.0003	-	0.0007	0.0067
in-domain	phrase-rules	0.0029	−0.0061	0	−0.0023	-	−0.0054	−0.0094
es-en	phrase-dict	0.0202	0.0138	0.0073	0.0162	-	0.0211	0.0259
out-of-domain	phrase-rules	0.0227	0.0219	0.0288	0.0156	-	0.0092	0.0230
br-fr	phrase-dict	0.0106	0.0052	0.0098	0.0079	0.0115	-	-
in-domain	phrase-rules	0.0024	0.0103	0.0020	0.0044	0.0029	-	-

Table 4: Relative weights assigned to the binary feature function that flags whether a phrase pair has been obtained from the Apertium bilingual resources (dictionaries and rules) or from the training parallel corpus in the different evaluation set-ups. Weigths have been normalised by dividing them by the highest weight assigned to a feature of its corresponding set-up.

		$2k$	$5k$	$10k$	$20k$	$\approx 27k$	$40k$	$80k$
es-en	phrase-dict	0.194	0.120	0.100	0.062	-	0.046	0.026
in-domain	phrase-rules	0.172	0.105	0.083	0.047	-	0.033	0.019
es-en	phrase-dict	0.282	0.229	0.188	0.144	-	0.121	0.088
out-of-domain	phrase-rules	0.374	0.319	0.277	0.237	-	0.194	0.167
br-fr	phrase-dict	0.276	0.184	0.133	0.096	0.086	-	-
in-domain	phrase-rules	0.319	0.224	0.181	0.138	0.134	-	-

Table 5: Proportion of phrase pairs used in the translation of the test set that have been generated from the Apertium bilingual resources (dictionaries and rules).

J. Goodman and S. F. Chen. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University, August.

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

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, C. Shen, W.and Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, demonstration session*, pages 177–180.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. *MT summit*, 5:12–16.

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

P. Langlais. 2002. Improving a general-purpose Statistical Translation Engine by terminological lexicons. In *Second International Workshop on Computational Terminology*, pages 1–7.

A. Lopez and P. Resnik. 2006. Word-based alignment, phrase-based translation: Whats the link. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 90–99.

Y. Marton, C. Callison-Burch, and P. Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 381–390.

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

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

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

M. Popović and H. Ney. 2006. Statistical machine translation with a small amount of bilingual training data. In *LREC workshop on Minority Languages*, pages 25–29.

H. Schwenk, S. Abdul-Rauf, L. Barrault, and J. Senellart. 2009. SMT and SPE machine translation systems for WMT'09. In *Proceedings of the 4th Workshop on Statistical Machine Translation*, pages 130–134.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *7th International Conference on Spoken Language Processing*, pages 901–904.

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

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

Y. Zhang, S. Vogel, and A. Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 2051–2054.