Integrating shallow-transfer rules into phrase-based statistical machine translation

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 extensively evaluate a new hybridisation approach consisting of enriching the phrase table of a phrase-based statistical machine translation system with bilingual phrase pairs matching transfer rules and dictionary entries from a shallow-transfer rule-based machine translation system. The experiments conducted show an improvement in translation quality, specially when the parallel corpus available for training is small or when translating out-of-domain texts that are well covered by the shallow-transfer rule-based machine translation system.

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 bilingual corpora are available. However, bilingual corpora 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) (Hutchins and 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.

When both parallel corpora and linguistic information exist, a hybrid approach (Thurmair, 2009) may be taken in order to make the most of such resources. In this paper we present a new hybrid approach which enriches a phrase-based SMT system with resources taken from shallow-transfer RBMT. Shallow-transfer RBMT systems do not perform a complete syntactic analysis of the input sentences, but they rather work with much simpler intermediate representations. Hybridisation between shallow-transfer RBMT and SMT has not yet been explored. Existing hybridisation strategies usually involve more complex RBMT systems and treat them as black boxes, whereas our approach improves SMT by explicitly using the RBMT linguistic resources. We provide an exhaustive evaluation of our hybridisation approach and of the most similar one (Eisele et al., 2008), on the Spanish–English and English–Spanish language pairs by using different training corpus sizes and evaluation corpora.

The rest of the paper is organised as follows. Next section overviews the MT systems we combine in our approach. Section 3 outlines related hybrid approaches, whereas our approach is described in section 4. Sections 5 and 6 present the experiments conducted and the results achieved, respectively. The paper ends with some concluding remarks.

2 Translation approaches

2.1 Phrase-based statistical machine translation

Phrase-based statistical machine translation systems (PBSMT) (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-language (SL) 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 (TL) model.

2.2 Shallow-transfer rule-based machine translation

The RBMT process (Hutchins and Somers, 1992) can be split into three steps: i) analysis of the SL text to build a SL intermediate representation, ii) transfer from that SL intermediate representation to a TL representation, and iii) 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 apply simple shallow-transfer rules that operate on sequences of lexical forms; thus, no syntactic parsing is performed. Apertium (Forcada et al., 2011), the shallow-transfer RBMT platform we have used in our experiments, splits the transfer step into structural and lexical transfer. The lexical transfer is done by using a bilingual dictionary which, for each SL lexical form, always provides the same TL lexical form; thus, no lexical selection is performed. Note 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 from being performed word for word in those cases in which this would result in an incorrect translation. For the translation between non-related languages, such as Spanish and English, the structural transfer may be split into three levels in order to facilitate the writing of rules. The first level performs short-distance operations, such as reorderings and gender and number agreement between nouns and adjectives, and groups sequences of lexical forms into *chunks*; second-level rules perform inter *chunk* operations, such as agreements between distant constituents (i.e. subject and main verb); finally, the third-level transfer unwraps the chunks and generates a sequence of TL lexical forms from each *chunk*.

Suppose that the Spanish sentence *Por otra parte mis amigos americanos han decidido venir* is to be translated into English by Apertium. First, it is analysed as:

```
por otra parte<adv>
mío<det><pos><mf><pl>
amigo<n><m><pl>
americano<adj><m><pl>
haber<vbhaver><pri><p3><pl>
decidir<vblex><pp><m><sg>
venir<vblex><inf>
```

which splits the sentence into seven lexical forms: a multi-word adverb (*por otra parte*), a plural possesive determiner (*mío*), a noun and an adjective in masculine plural (*amigo* and *americano*, respectively), the present third-person plural form of the verb *to be* (*haber*), the masculine singular past participle of the verb *decidir*, and the verb *venir* in infinitive mood. Then, the transfer is executed. It starts by performing the lexical transfer and applying the first-level rules of the structural transfer in parallel. The lexical transfer gives as a result:

```
on the other hand<adv>
my<det><pos><pl>
friend<n><pl>
american<adj>
have<vbhaver><pres>
decide<vblex><pp>
come<vblex><inf>
```

Four first-level structural transfer rules are triggered: the first one matches a single adverb (the first lexical form of the example); the second one matches a determiner followed by a noun and an adjective (the next three lexical forms); the third one matches a form of the verb *haber* plus the past participle form of another verb (the next two lexical forms); and the last one matches a verb in infinitive mood (last lexical form). Each of these first-level rules group the matched lexical forms in the same *chunk* and perform local operations within the chunk; for instance, the second rule reorders the adjective and the noun:

```
ADV{ on the other hand<adv> }
NOUN_PHRASE{ my<det><pos><pl>
american<adj> friend<n><pl> }
HABER_PP{ have<vbhaver><pres>
decide<vblex><pp> }
INF{ come<vblex><inf> }
```

After that, inter *chunk* operations are performed. The *chunk* sequence *HABER_PP* (verb in present perfect tense), *INF* (verb in infinitive mood) matches a second-level rule which adds a new chunk including the preposition *to*:

```
ADV{ on the other hand<adv> }
NOUN_PHRASE{ my<det><pos><pl>
friend<n><pl> american<adj> }
HABER_PP{ have<vbhaver><pres>
decide<vblex><pp> }
TO{ to<pr> }
INF{ come<vblex><inf> }
```

Third-level structural transfer removes *chunks* to generate a plain sequence of TL lexical forms:

```
on the other hand<adv>
my<det><pos><pl>
american<adj>
friend<n><pl>
have<vbhaver><pres>
decide<vblex><pp>
to<pr> come<vblex><inf>
```

Finally, the translation into TL is generated from the TL lexical forms: *On the other hand my American friends have decided to come*.

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 because they may be split into smaller units by the phrase-extraction algorithm. Other approaches go beyond simply adding a dictionary to the parallel corpus. For instance, Popović and Ney (2006) propose combining that strategy with the use of hand-crafted rules to reorder the SL sentences to match the structure of the TL.

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; RBMT and PBSMT systems can also be combined in a serial fashion (Dugast et al., 2007). 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 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 directly 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.

4 Enhancing phrase-based SMT with shallow-transfer linguistic resources

As already mentioned, the Apertium structural transfer detects sequences of lexical forms which need to be treated together to prevent them from being wrongly translated. 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 suffices to encode all the linguistic information provided by Apertium. Below, the generation of these phrase pairs and three different methods to score them are presented.

4.1 Phrase pair generation

Generating bilingual phrase pairs from the bilingual dictionary is straightforward. First, all the SL surface forms recognised by Apertium and their corresponding lexical forms are listed; then, these SL lexical forms are translated using the bilingual dictionary; finally, their TL surface forms are generated.

Bilingual phrase pairs which match structural transfer rules are generated in a similar way. First, the SL sentences to be translated are analysed to get their SL lexical forms, and then the sequences of lexical forms that either match a first-level or a second-level structural transfer rule are passed through the Apertium pipeline to get their translations. If a sequence of SL lexical forms is covered by more than one structural transfer rule in the same level, they will be used to generate as many bilingual phrase pairs as 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 the Spanish sentence *Por otra parte mis amigos americanos han decidido venir*, from the example in section 3, be one of the sentences to be translated. The SL sequences *por otra parte*, *mis amigos americanos*, *amigos americanos*, *han decidido*, *venir* and *han decidido venir* would be used to generate bilingual phrase pairs because they match a first-level rule, a second-level rule, or both. The SL words *amigos americanos* are used twice because they are covered by two first-level rules: one that matches a determiner followed by a noun and an adjective, and another that matches a noun followed by an adjective. The SL words *han decidido* and *venir* are used because they match first-level rules, whereas *han decidido venir* matches a second-level rule.

Note that the generation of bilingual phrase pairs from the shallow-transfer rules is guided by the text to translate. We decided to do it in this way in order to avoid meaningless phrases and to make our approach computationally feasible. Consider, for instance, the rule which is triggered by a determiner followed by a noun and an adjective. Generating all the possible phrase pairs matching 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*.

4.2 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 and lexical weightings, and phrase penalty. The phrase translation probabilities and lexical weightings of the the phrase pairs generated from Apertium may be calculated in three different ways which we describe next.

Augmenting the training corpus (*corpus-rules*). The simplest approach involves appending the Apertium-generated phrase pairs to the training corpus and running the usual PBSMT training algorithm. This improves the alignments of the original training corpus and enriches both the phrase table and the reordering model. However, the phrase extraction algorithm (Koehn, 2010, sec. 5.2.3) may split the resulting bilingual phrase pairs into smaller units which may cause multi-word expressions not to be translated in the same way as they appear in the Apertium bilingual dictionary.

Directly expanding the phrase table (*phrase-rules*). Apertium-generated phrase pairs are added once to the list of corpus-extracted phrase pairs; then, the phrase translation probabilities are calculated by relative frequency as it is usually done (Koehn, 2010, sec. 5.2.5). As they 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 very frequent, single 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 Apertium-generated phrase pairs.

To calculate the lexical weightings (Koehn, 2010, sec. 5.3.3) of the RBMT-generated phrase pairs, a probabilistic bilingual dictionary and the alignments between the words in the source side and those in the target side are needed. These word alignments are obtained by tracing back the operations carried out in the different steps of Apertium. 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, so that multi-word expressions are assigned a lexical weighting of 1.0. Figure 1 shows the alignment between the words in the running example. Regarding the probabilistic bilingual dictionary, it is usually computed from the word alignments extracted from the training corpus. Our approach also takes advantage from the Apertium bilingual dictionary to obtain a richer probabilistic bilingual dictionary.

Combining both approaches (*pc-rules*). The two previous approaches may be combined by appending the RBMT bilingual phrase pairs to both the

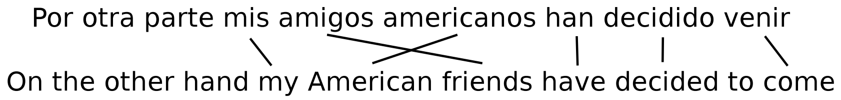
Por otra parte mis amigos americanos han decidido venir
On the other hand my American friends have decided to come

Figure 1: Example of word alignment obtained by tracing back the operations done by Apertium when translating from Spanish to English the sentence *Por otra parte mis amigos americanos han decidido venir*. Note that *por otra parte* is analysed by Apertium as a multi-word expression whose words are left unaligned for convenience (see section 4.2).

training corpus and the phrase table. Following this strategy, the list of phrase pairs from which the phrase table is built will contain each Apertium-generated pair twice, but each sub-phrase identified by the phrase-extraction algorithm only once.

5 Experimental settings

We evaluated our RBMT–PBSMT hybridisation approach on the two translation directions of the Spanish–English language pair and with different corpus sizes to test the translation scenarios in which it best fits. Small corpus sizes allows us to test if our hybrid approach is useful in the translation between less-resourced language pairs where one of the languages is highly inflected (Spanish) while the other one (English) is not. We compare the performance of our approach to that by Eisele et al. (2008) because it is the most similar to ours.

PBSMT systems for both directions were trained from the Europarl v5 parallel corpus (Koehn, 2005). Its whole target side, except for the Q4/2000 portion, was used to train a language model. Regarding the translation model, we learned it from corpora of different sizes; more precisely, we used fragments of the Europarl corpus consisting of 5 000, 10 000, 20 000, 40 000 and 80 000 parallel sentences. In addition we also used the whole Europarl corpus and an empty training set. 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.

We 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* set for testing. Both sets are distributed as part of the WMT 2010 shared translation task.[1] Sentences containing more than 40 tokens were removed from all the bilingual corpora to avoid problems with the word alignment tool (Och and Ney, 2003).[2]

Corpus		Origin	Size (sentences)
Language model (es)		Europarl	1 650 152
Language model (en)		Europarl	1 650 152
Training	0	-	0
	2K	Europarl	2 000
	5K	Europarl	5 000
	10K	Europarl	10 000
	20K	Europarl	20 000
	40K	Europarl	40 000
	80K	Europarl	80 000
	all	Europarl	1 272 260
In-domain tuning		Europarl	2 000
In-domain test		Europarl	2 000
Out-of-domain tuning		WMT 2010	1 732
Out-of-domain test		WMT 2010	2 215

Table 1: Data about the Spanish–English parallel corpora used in the experiments.

We used the free/open-source PBSMT system Moses (Koehn et al., 2007) with the 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 were downloaded from the Apertium Subversion repository. The linguistic data contains 326 228 entries in the bilingual dictionary; 106 first-level and 31 second-level structural transfer rules for Spanish–English; and 216

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

[2] We did that also with the tuning and test sets in order to use exactly the same corpora to evaluate our approach and the one by (Eisele et al., 2008). Recall that that approach needs to align the sentences in the test set with their RBMT translations.

		In-domain							Out-of-domain						
		0	5K	10K	20K	40K	80K	all	0	5K	10K	20K	40K	80K	all
es–en	baseline	-	22.69	24.84	26.81	28.23	29.15	31.75	-	14.33	16.54	18.31	19.97	21.63	24.23
	corpus-dict	12.91	**25.62**	**26.54**	**27.91**	28.48	**29.80**	31.78	16.20	**20.99**	**21.67**	**22.51**	**23.11**	**23.43**	**25.15**
	corpus-rules	***17.17***	**25.47**	**26.75**	**27.93**	***29.02***	**29.73**	31.97	16.00	**21.09**	**21.40**	**22.31**	**22.86**	***23.67***	**24.91**
	phrase-dict	15.82	**25.52**	**26.51**	**27.79**	**28.71**	**29.44**	31.57	17.54	20.42	**20.83**	**21.61**	**22.13**	**22.90**	24.50
	phrase-rules	***17.05***	***25.80***	***26.79***	**28.03**	***29.00***	**29.50**	*31.97*	*19.52*	***21.32***	***21.66***	***22.23***	***22.80***	**23.08**	***24.81***
	pc-dict	16.74	**24.93**	**26.66**	**27.52**	**28.56**	**29.60**	31.88	18.10	19.84	**20.73**	**21.53**	**22.06**	**22.87**	**24.58**
	pc-rules	***18.44***	***25.60***	**26.73**	**27.81**	***28.88***	**29.63**	31.84	17.58	***21.23***	***21.73***	***22.31***	***23.02***	***23.27***	**24.81**
	Eisele	-	24.80	26.04	27.28	28.44	28.85	31.89	-	19.84	20.11	20.39	20.93	21.81	23.79
	Apertium	16.63							20.07						
en–es	baseline	-	23.10	25.39	27.14	28.60	29.45	31.51	-	15.08	16.96	18.78	20.26	21.27	24.55
	corpus-dict	12.06	**25.67**	**27.13**	**28.35**	28.86	29.61	31.47	13.65	**20.30**	**20.88**	**21.54**	**22.09**	**22.67**	24.73
	corpus-rules	***17.37***	**26.07**	**27.01**	**28.16**	28.93	**29.75**	31.72	*16.41*	***21.58***	***22.30***	***23.03***	***23.34***	***23.97***	***25.33***
	phrase-dict	14.64	**25.42**	**26.94**	**28.10**	**28.95**	29.55	31.33	16.27	19.96	**20.71**	**21.33**	**22.12**	**22.44**	24.41
	phrase-rules	*16.36*	***25.87***	***27.34***	**28.31**	**28.98**	29.54	31.55	*18.95*	***21.42***	***21.88***	***22.58***	***23.21***	***23.42***	***25.47***
	pc-dict	16.56	**25.48**	**27.00**	**28.02**	28.92	29.50	31.36	17.13	20.06	**20.83**	**21.89**	**22.48**	**22.78**	24.59
	pc-rules	***18.28***	***26.01***	**27.08**	**27.94**	**29.06**	29.51	31.41	17.62	***22.03***	***22.61***	***23.17***	***23.54***	***23.81***	***25.26***
	Eisele	-	25.00	26.31	27.74	28.48	29.43	31.24	-	17.01	18.32	19.90	21.00	21.71	24.35
	Apertium	16.13							19.72						

Table 2: 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. Underlined scores mean that the approach by Eisele et al. (2008) is outperformed and that the improvement is statistically significant. Scores of hybrid systems built with the Apertium rules and dictionaries in italics mean that they outperform their dictionary-based counterpart by a statistically significant margin.

and 60 rules, respectively, for English–Spanish.

We have tested the following configurations:

- a state-of-the-art PBSMT system (*baseline*);
- the Apertium shallow-transfer RBMT engine, from which the dictionaries and transfer rules have been taken (*Apertium*);
- the three hybridisation strategies described in section 4.2 (*corpus-rules*, *phrase-rules*, and *pc-rules*, respectively);
- a reduced version of each of the hybridisation strategies in which only dictionary-matching bilingual phrases are included (*corpus-dict*, *phrase-dict*, and *pc-dict*, respectively); and
- the approach by Eisele et al. (2008), using the alignment model learned from the training corpus to get the word alignments between the source sentences and the RBMT-translated sentences (*Eisele*).

6 Results and discussion

Table 2 reports the translation performance as measured by BLEU (Papineni et al., 2002) for the different systems, training corpora and evaluation corpora previously described. Statistical significance of the difference 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. Table 3 shows the proportion of Apertium phrases chosen by the decoder.

The results show that our hybrid approaches outperform both pure RBMT and PBSMT systems in terms of BLEU. This improvement is statistically significant for all training corpus sizes when translating out-of-domain texts; for in-domain texts the statistical significance only holds when the training corpus is relatively small. Note that The out-of-domain tuning and test sets come from a general (news) domain and Apertium data has been developed bearing in mind the translation of general texts (mainly news). In this case, Apertium-generated phrases which contain hand-crafted knowledge from a general domain cover sequences of words in the input text which are not covered, or are sparsely found, in the original training corpora. Contrarily, the in-domain tests reveal that, as soon as the PBSMT system is able to learn some reliable information from the parallel corpus, Apertium phrases be-

		In-domain						Out-of-domain					
		5K	10K	20K	40K	80K	all	5K	10K	20K	40K	80K	all
es–en	corpus-dict	0.130	0.096	0.064	0.045	0.028	0.004	0.244	0.172	0.139	0.104	0.078	0.017
	corpus-rules	0.210	0.147	0.124	0.101	0.074	0.032	0.267	0.192	0.206	0.128	0.132	0.058
	phrase-dict	0.130	0.086	0.070	0.037	0.028	0.004	0.225	0.194	0.159	0.109	0.088	0.024
	phrase-rules	0.117	0.078	0.061	0.034	0.023	0.004	0.294	0.264	0.236	0.185	0.158	0.070
	pc-dict	0.135	0.093	0.066	0.044	0.030	0.003	0.220	0.185	0.149	0.108	0.090	0.024
	pc-rules	0.204	0.145	0.143	0.098	0.063	0.027	0.268	0.233	0.186	0.196	0.149	0.083
	Eisele	0.218	0.231	0.140	0.064	0.112	0.032	0.884	0.929	0.822	0.574	0.530	0.503
en–es	corpus-dict	0.105	0.063	0.042	0.025	0.018	0.001	0.196	0.148	0.110	0.084	0.073	0.014
	corpus-rules	0.200	0.143	0.095	0.086	0.068	0.034	0.205	0.196	0.196	0.157	0.139	0.062
	phrase-dict	0.100	0.069	0.059	0.030	0.017	0.001	0.212	0.162	0.129	0.100	0.071	0.018
	phrase-rules	0.163	0.106	0.096	0.081	0.053	0.034	0.256	0.248	0.196	0.178	0.135	0.079
	pc-dict	0.104	0.069	0.046	0.024	0.015	0.001	0.209	0.163	0.122	0.091	0.068	0.015
	pc-rules	0.169	0.134	0.102	0.082	0.086	0.038	0.265	0.205	0.175	0.199	0.171	0.089
	Eisele	0.141	0.161	0.107	0.075	0.114	0.053	0.576	0.473	0.378	0.332	0.309	0.390

Table 3: Proportion of Apertium-generated phrase pairs chosen by the decoder.

come useless because the in-domain test sets comes from the specialised domain of parliament speeches. Data from table 3 support this hypothesis: for each training corpus size and hybrid system, the proportion of Apertium phrases is higher in the out-of-domain evaluation, being the difference between the two evaluation domains specially remarkable for the biggest training corpus size. In addition, the hybrid systems built without training corpus outperform Apertium when translating in-domain texts, probably because the language model helps to choose the best combination of transfer rules, avoiding the inflexibility of the left-to-right, longest match policy.

Regarding the difference between the hybrid systems enriched with all the Apertium resources (in boldface in tables 2 and 3) and those only including the dictionary, some patterns can be detected. The impact of the shallow-transfer rules is higher when translating out-of-domain texts and decreases as the training corpus grows. That is, their impact is higher when the decoder chooses a high proportion of Apertium phrases, according to table 3. Moreover, the systems including shallow-transfer rules outperform their counterparts which only include the dictionary by a wider margin 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 to determine the gender, number and person when translating into Spanish.

As regards the three hybrid strategies we defined, no consistent differences exist between them, probably because the Apertium-generated bilingual phrase pairs were too short for their subphrases to clearly improve the reordering model and because bigger improvements in aligment quality may be needed to improve BLEU (Lopez and Resnik, 2006).

Finally, our hybridisation strategy provides better results than the approach by Eisele et al. (2008), specially when small corpora are used for training. Under such circumstance, no reliable alignment models can be learned for the Eisele's set-up from the training corpus and then no reliable phrases pairs can be obtained from the input text and its rule-based translation. Our approach is not affected by this problem because it does not need any special alignment model. In addition, the approach by Eisele et al. (2008) involves concatenating two phrase tables independently learned, which causes their probabilities not to be consistent.

7 Conclusions and future work

We have described a hybridisation approach consisting of enriching a PBSMT system with bilingual phrase pairs matching transfer rules and dictionaries from a shallow-transfer RBMT system. Automatic evaluation shows a clear improvement of translation quality when the PBSMT system is trained on a small parallel corpus, and when it is trained on larger parallel corpora and the texts to translate come from a general (news) domain that is well covered by the shallow-transfer RBMT system. In the context of the WMT 2011 shared translation task (Callison-Burch et al., 2011), our approach was also manually evaluated in the latter scenario: the human evaluation confirmed the improvements already measured automatically since our system was not statistically significantly outperformed by any other system.

Shallow-transfer rules have shown a strong impact on translation quality when translating from English, a language with a simple morphology, to Spanish, whose morphology is richer, thus making worth to evaluate our hybridisation strategy with other morphologically rich target languages. In addition, we also plan to compare it with other hybrid approaches which combine a PBSMT system with explicit linguistic information, such as the one by Dugast et al. (2007).

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.

C. Callison-Burch, P. Koehn, C. Monz, and O. Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64.

L. Dugast, J. Senellart, and P. Koehn. 2007. Statistical post-editing on SYSTRAN's rule-based translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 220–223.

L. Dugast, J. Senellart, and P. Koehn. 2008. Can we Relearn an RBMT System? In *Proceedings of the Third 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 Third 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-Martínez, G. Ramírez-Sánchez, and F.M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation. Special Issue on Free/Open-Source Machine Translation*, In press.

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 ACL on Interactive Poster and Demonstration Sessions*, 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.

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

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 of the ACL*, 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 of the ACL*, pages 311–318.

M. Popovic 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 Fourth 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 Fourth International Conference on Language Resources and Evaluation*, pages 2051–2054.