Chapter 4

An Open-Source Toolkit for Integrating Shallow-Transfer Rules into Phrase-Based Statistical Machine Translation

V. M. Sánchez-Cartagena, F. Sánchez-Martínez, J. A. Pérez-Ortiz

Universitat d'Alacant

Abstract

In this paper, we present an open-source toolkit to enrich a phrase-based statistical machine translation system (Moses) with phrase pairs generated from the linguistic resources of a shallow-transfer rule-based machine translation system (Apertium). A system built with this toolkit was not outperformed by any other participant in the shared translation task of the Sixth Workshop on Statistical Machine Translation (WMT 11) for the Spanish–English language pair.

4.1 Introduction

Statistical machine translation (SMT) (Koehn, 2010) 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 and economic resources. When both parallel corpora and linguistic information exist, a hybrid approach may be followed in order to make the most of such resources.

In this paper, we present the free/open-source Rule2Phrase Toolkit, which implements a recently developed hybrid approach (Sánchez-Cartagena et al., 2011, Sánchez-Cartagena et al., 2011a,b) to enrich a phrase-based (Koehn et al., 2003) SMT system with resources from shallow-transfer RBMT; this toolkit is designed to work with the Apertium (Forcada et al., 2011) RBMT platform and the Moses (Koehn et al., 2007) phrase-based SMT system. The Rule2Phrase Toolkit, which is described for the first time in this paper, permits the creation of a set of phrase pairs which encode the knowledge present in the Apertium linguistic resources, and implements different strategies to integrate them in the translation models built with Moses.

Different experiments have been performed previously to validate the hybrid approach using this toolkit. Experiments carried out with small training corpora confirmed its effectiveness (Sánchez-Cartagena et al., 2011a). Experiments performed with bigger training corpora showed that Apertium data is very useful to improve the translation of general domain texts when systems are trained on in-domain corpora (Sánchez-Cartagena et al., 2011b). In addition, a system built under this hybrid philosophy (Sánchez-Cartagena et al., 2011) was not outperformed by a statistically significant margin by any other participant in the shared translation task of the Sixth Workshop on Statistical Machine Translation (WMT 11)(Callison-Burch et al., 2011) for the Spanish-English language pair.

The rest of the paper is organised as follows. Next section overviews the MT systems combined by the Rule2Phrase Toolkit, while section 4.3 presents some similar hybridisation approaches. The hybridisation strategy is described in section 4.4; then, section 4.5 describes the design principles, some implementation details and usage examples of the toolkit. The paper ends with some concluding remarks.

4.2 Translation Approaches

4.2.1 Phrase-Based Statistical Machine Translation

The Moses toolkit, as well as other phrase-based statistical machine translation systems (PBSMT) (Koehn, 2010, ch. 5), translates 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 (possibly with some reordering) the translation of the different source segments according to the phrase table.

4.2.2 Shallow-transfer rule-based machine translation

The RBMT process 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 target-language (TL) representation, and iii) generation of the final translation from the TL intermediate representation.

Shallow-transfer RBMT systems use relatively simple intermediate representations 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 (no full parsing is performed). Apertium, the shallow-transfer RBMT platform our toolkit is designed to work with, splits the transfer step into structural and lexical transfer. The lexical transfer is performed by using a bilingual dictionary which, for each SL lexical form, always provides the same TL lexical form (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 performed by applying a set of rules in a left-to-right, longest-match fashion; rules are applied to sequences of words and prevent word-for-word translation in those cases in which this would result in an incorrect translation. The structural transfer may be split into three levels in order to facilitate the writing of rules, although, for the sake of simplicity, in this paper only a single-level transfer is taken into account.[1]

The SL intermediate representation is obtained by analysing the SL text with a SL monolingual dictionary and a part-of-speech tagger. A *pretransfer* module then splits those lexical forms, such as verbs with enclitic pronouns and contractions, that will be processed as separate units by the transfer module. The final translation is generated from the TL intermediate representation with a TL monolingual dictionary.

Suppose that the Catalan sentence *La deterioració del senyal* (*the deterioration of the signal* in English) is to be translated into Spanish by Apertium. First, it is analysed as:

```
el<det><def><f><sg> deterioració<n><m><sg>
de<pr>+el<det><def><m><sg>  senyal<n><m><sg>
```

which splits the sentence into four lexical forms: a feminine plural definite determiner (*la*), a noun in feminine singular (*deterioració*), the preposition *de*, joint with a masculine plural definite determiner (*el*), and a noun in masculine singular (*senyal*).
The *pretransfer* module then splits the joint lexical form:

```
el<det><def><f><sg> deterioració<n><f><sg> de<pr>
el<det><def><m><sg> senyal<n><m><sg>
```

After that, the transfer is executed. It performs the lexical transfer and applies the first-level rules of the structural transfer in parallel. On the one hand, the lexical transfer gives as a result:

```
el<det><def><f><sg> deterioro<n><m><sg> de<pr>
el<det><def><m><sg> señal<n><f><sg>
```

On the other hand, a first-level Apertium structural transfer rule is triggered, twice in this case. This rule matches a determiner followed by a noun and makes the determiner to agree in gender and number with the noun. As a result, the final TL lexical forms are obtained:

```
el<det><def><m><sg> deterioro<n><m><sg> de<pr>
el<det><def><f><sg> señal<n><f><sg>
```

Finally, the translation into TL is generated from the TL lexical forms: *El deterioro de la señal.*

[1] Although it facilitates the writing of long rules by linguists, Apertium multi-level transfer has the same expressive power than single-level transfer. However, it is important to remark that the Rule2Phrase Toolkit is also able to work with multi-level transfer rules.

4.3 Related work

Although we are not aware of any other approach which explicitly reuses both structural transfer rules and bilingual dictionaries of a RBMT system in order to improve a SMT one, as does the Rule2Phrase Toolkit, some similar approaches exist.

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, Popovic 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.

All the approaches described above share a feature: the main system is a statistical one and it is enriched (or even built) with resources from RBMT. However, there are other ways of combining RBMT and SMT. For instance, in statistical post-edition (Simard et al., 2007) the output of an RBMT system is coupled to a SMT decoder which tries to correct the errors made by the RBMT engine. A SMT system may also be enriched with other resources, such as phrases from a example-based machine translation system (Dandapat et al., 2010).

4.4 Conceptual Background

As already mentioned, the Apertium structural transfer module detects sequences of lexical forms which need to be processed together to prevent wrong word-for-word translations.

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 incorporate all the linguistic information provided by Apertium. In this section, the generation of these phrase pairs and three different methods to score them are presented; additional implementation details for the Rule2Phrase Toolkit can be found in section 4.5.

4.4.1 Phrase Pair Generation

As described in section 4.5.2, generating bilingual phrase pairs from the bilingual dictionary involves a straightforward combination of the data in the bilingual and monolingual dictionaries.

When generating phrase pairs from the structural transfer rules, the amount of generated pairs is an important issue. 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, producing many meaningless phrases, such as the Spanish *el niño inalámbrico* (*the wireless boy* in English) and making the approach computationally infeasible due to the large number of resulting combinations. In the experiments carried out to evaluate the hybrid approach, this issue was solved by generating only phrase pairs whose source side occurs in the test and tuning sets.

Let the Catalan sentence *El senyal roig*, similar to the example in section 4.2.2, be one of the sentences to be translated. If, in addition to the rule fired by a determiner plus a noun presented in the previous example, there is another rule which matches a determiner followed by a noun and an adjective, the SL sequences *El senyal*, and *El senyal roig* are used to generate bilingual phrase pairs because both match a first-level rule; note that the SL word sequence *El senyal* is used twice because it is covered by two first-level rules.

4.4.2 Scoring the New Phrase Pairs

The Moses PBSMT system attaches 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 phrase pairs generated from Apertium may be calculated in three different ways which we describe next (computation of the phrase penalty is trivial). As in previous experiments (Sánchez-Cartagena et al., 2011b) neither of the three strategies clearly outperformed the others, the three approaches are implemented by the toolkit.

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 obtained from 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.

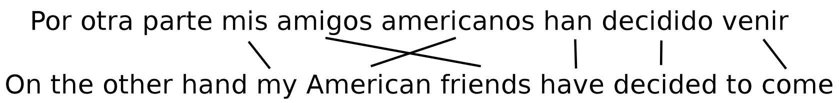
Por otra parte mis amigos americanos han decidido venir
On the other hand my American friends have decided to come

Figure 4.1: Example of word alignment obtained by tracing back the operations performed 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 4.4.2).

Expanding the Phrase Table (*phrase-rules*). Apertium-generated phrase pairs are added 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, an additional boolean score to flag Apertium-generated phrase pairs can be introduced. [2]

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 (see section 4.5.2). Only those words which are neither split nor joined 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 4.1 shows the alignment between the words of a sentence in Spanish and its English translation with Apertium which would be obtained with this strategy. 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 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. Therefore, phrase pairs extracted from Apertium which have been split will be present in the phrase table, but they will have lower scores than those which have not been split. In addition, as in the *corpus-rules* approach, the alignment model is built from a bigger corpus, and so is the reordering model.

[2] Phrase pairs generated from Apertium which are also extracted from the corpus are flagged as Apertium-generated too.

4.5 Description of the Toolkit

4.5.1 Overview

The Rule2Phrase Toolkit implements the hybridisation strategy described above by easily allowing its users to perform these two main steps:

1. Generate a list of phrase pairs and the alignments between their words from the Apertium linguistic resources (see section 4.4.1).
2. Integrate the resulting Apertium-generated phrases in a PBSMT system built with Moses following the three strategies previously presented (see section 4.4.2).

As trying to generate all the SL phrases which match a transfer rule would result in an excessive amount of meaningless phrases (see section 4.4.1), a mechanism to filter them and generate only sentences which are likely to appear in the texts the hybrid system will have to translate is needed. In the experiments performed to validate this hybrid approach (Sánchez-Cartagena et al., 2011, Sánchez-Cartagena et al., 2011a,b) this objective was accomplished by simply generating only phrases present in the tuning and test corpora. However, a different solution is needed when the resulting hybrid system is to be used in a real scenario, where the texts to be translated are not known a priori.

Our toolkit is able to deal with both scenarios by performing a n-gram based filtering. A list of allowed n-grams must be provided when generating the phrase pairs from the transfer rules, so that those SL sequences containing exclusively n-grams from the allowed list are generated. Therefore, if the test corpus is provided, the list of allowed n-grams is extracted from it. If not, one can simply use the most probable n-grams of a source-language model. This second strategy is partially implemented, and still requires a systematic evaluation.

Regarding the integration of the Apertium-generated corpus in the models of PBSMT system, our toolkit provides a wrapper over the Moses training scripts, which facilitates the integration the Apertium-generated corpus in the PBSMT models following any of the three strategies defined in section 4.4.2.

4.5.2 Design Principles

The design of the two modules of our toolkit and their interaction with Apertium and Moses are discussed in this section.

Phrase Generation Module

The generation of the Apertium-generated phrases from the dictionaries is straightforward. All the SL surface forms recognised by Apertium and their corresponding lexical forms are obtained from the SL monolingual dictionary; then, these SL lexical forms are translated using the bilingual dictionary; finally, their TL surface forms are generated using the TL monolingual dictionary.

The generation of phrase pairs from the Apertium shallow-transfer rules, which is summarised in figure 4.2, is performed as follows. Firstly, all the SL lexical form sequences (extracted from the SL monolingual dictionary) which match a first-level transfer rule and

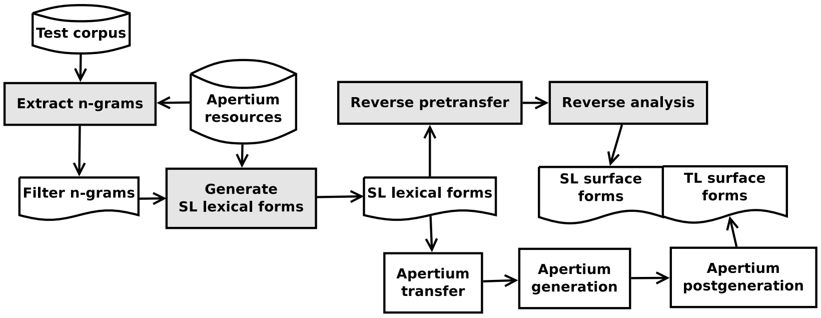

Figure 4.2: Steps carried out by the Rule2Phrase Toolkit to generate a set of phrase pairs from the Apertium transfer rules (grey-shadowed boxes).

whose subsequences are present in the list of allowed n-grams are generated. Note that transfer rules are applied to the output of the *pretransfer* module, which means that, at this step, lexical forms which would have been split by the *pretransfer* module (such as contractions and verbs plus enclitic pronouns) must appear as independent lexical forms. Therefore, when extracting the n-grams from the test corpus (see section 4.5.1), it must be analysed and passed through the *pretransfer* module first.

Then, for each SL lexical form sequence, two processes are carried out in order to obtain, respectively, the TL and the SL side of the resulting bilingual phrase pair.

In the first process, each SL lexical form sequence is passed through the rest of the Apertium pipeline to obtain a TL surface form sequence. The alignments between the input and output sequences of lexical forms of each module are also computed. For instance, consider the following SL lexical form sequence obtained when generating phrase pairs from Apertium for translating from Catalan to Spanish and that a rule matching the preposition *de* plus a determiner followed by a noun is applied:

```
de<pr> el<det><def><m><sg> senyal<n><m><sg>
```

The transfer module produces the following TL lexical forms:

```
de<pr> el<det><def><f><sg> señal<n><f><sg> (1-1 2-2 3-3)
```

Alignments are represented as pairs of numbers $i - j$, where i is the position of the SL word aligned with the TL word at position j. Finally, the generation module produces the following TL surface forms:

```
de la señal (1-1 2-2 3-3)
```

In the second process, the SL surface forms are obtained by firstly passing each SL lexical form sequence through a new module which joins the words split by the pretransfer:

```
de<pr>+el<det><def><m><sg> senyal<n><m><sg> (1-1 1-2 2-3)
```

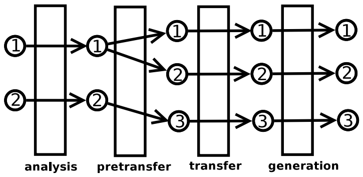

Figure 4.3: Alignments obtained from the different Apertium modules when translating the sentence *Del senyal* from Catalan to Spanish.

And then by using the SL monolingual dictionary to obtain the corresponding surface forms:

```
del senyal (1-1 2-2)
```

Finally, the end-to-end alignments between SL and TL surface forms are obtained by joining the alignments of each module. Alignments are combined in a transitive manner: if module A output is connected to module B input, word i in the input of module A is aligned with word j in its output, and word j in the input of B aligned with word k in its output, we can state that word i is aligned with word k. Figure 4.3 shows how the final alignments of the running example (*1-1 1-2 2-3*) are calculated. At this point the toolkit permits keeping only the alignments which meet the restrictions defined in section 4.4.2.

Integration Module

Implementing the strategy *corpus-rules*, defined in section 4.4.2, only requires concatenating the Apertium-generated phrases with the original training corpus and running the Moses training script as-is. However, the other two strategies involve adding additional steps to the standard Moses training process.

In particular, the following steps, summarised in figure 4.4, are automatically executed by the Rule2Phrase Toolkit to integrate the Apertium-generated phrases using the strategy *phrase-rules*:

1. Alignments of the original training corpus are obtained using the Moses toolkit.
2. The probabilistic bilingual dictionary is obtained from the concatenation of the original training corpus and the subset of the Apertium-generated phrases obtained from the dictionaries.
3. Phrase pairs are extracted from the original training corpus.
4. Apertium-generated phrase pairs are appended to the list of corpus-extracted phrase pairs.
5. Phrase pairs are scored to obtain the phrase table.
6. The boolean score which flags Apertium-generated phrase pairs (see section 4.4.2) is added to the phrase table.

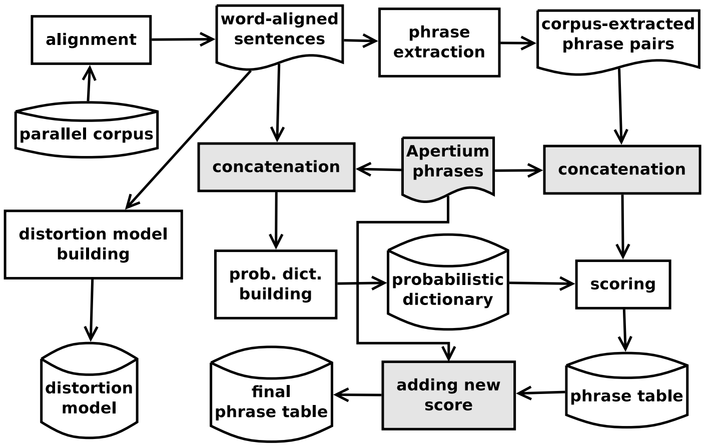

Figure 4.4: Steps carried out by the Rule2Phrase Toolkit to integrate a set of phrase pairs extracted from Apertium directly into the phrase table of a Moses system (grey-shadowed boxes).

7. The remaining standard Moses training pipeline is executed.

The *pc-rules* strategy involves a few less steps:

1. Phrase pairs are extracted from the concatenation of the original training corpus and the Apertium-generated phrases.
2. Apertium-generated phrase pairs are appended to the list of corpus-extracted phrase pairs.
3. Phrase pairs are scored to obtain the phrase table.
4. The boolean score which flags Apertium-generated phrase pairs (see section 4.4.2) is added to the phrase table.
5. The remaining standard Moses training pipeline is executed.

4.5.3 Implementation Details

The different steps of the actions carried out by the toolkit are encoded in a GNU Make Makefile in order to avoid executing some of them when it is not necessary. It is wrapped by a Python script which simplifies the parameter processing, and the main modules are written in Java and Python. Some UNIX utilities such as *sort* and *uniq* are used too.

The strategy to obtain the alignments varies across the different Apertium modules. Obtaining them from the analysis, generation and *pretransfer* modules is relatively straightforward as they keep word order and only split or join words in some cases. Therefore, maintaining a list of multi-word units suffices to obtain the alignments of the words processed by them.

On the contrary, obtaining the word alignments associated with the operations carried out by the transfer module is a more complex task, since transfer may add, delete and reorder words. In order to keep track of these operations, it has been modified to append to its output some extra information, including which transfer rules have been applied and, for each rule, which input SL word corresponds to each output TL word. When generating the phrase pairs from the Apertium linguistic resources, the toolkit executes this modified transfer module and parses its output to obtain the alignments.

4.5.4 Using the Toolkit

The Rule2Phrase Toolkit is licensed under the GNU GPL v3 free software license.[3] It has been tested under GNU/Linux, although it should work under other operating systems, as long as GNU Make and some other UNIX utilities are available for them.

Assuming that Apertium and Moses are already present in the system, installing our toolkit only involves unpacking the binary distribution and patching and recompiling Apertium for obtaining alignment information, a task for which convenient scripts are provided.

Once Apertium has been patched, generating the phrase pairs from it is as easy as typing:

```
$ python rule2Phrase.py --extract-n-grams --test TEST\_CORPUS
                        --output NGRAMS\_DIRECTORY --sl SL --tl TL
```

to extract the n-grams from the test corpus and:

```
$ python rule2Phrase.py --gen-phrases --n-grams NGRAMS\_DIRECTORY
                       --output NEWPHRASES\_DIRECTORY --sl SL --tl TL
```

to get the actual Apertium-generated phrases and their alignments. It is assumed that Apertium is installed under the standard prefix (`/usr/local`), but different installation directories may be chosen.[4]

Regarding the integration of the Apertium-generated phrases in the Moses PBSMT system, the toolkit provides a command for each of the steps described in section 4.5.2 which are not part of the standard Moses training procedure.[5] In addition, the enriched PBSMT system may be built with a single command. For instance, for the *pc-rules* strategy:

```
$ python rule2Phrase.py --buildSMT phrase-rules --synth-phrases
                                 NEWPHRASES\_DIRECTORY --sl SL --tl TL
```

4.6 Concluding Remarks

In this paper, we have presented an open-source toolkit which permits the enrichment of the PBSMT system Moses with phrase pairs generated from the linguistic resources of the shallow-transfer RBMT system Apertium. The hybridisation strategy implemented by the toolkit has already been evaluated with different experiments, which showed that it is very effective when the training corpus is small (Sánchez-Cartagena et al., 2011a)

[3] The toolkit can be downloaded from `http://www.dlsi.ua.es/~vmsanchez/Rule2Phrase.tar.gz`

[4] Run `python rule2Phrase.py -help` for a list of available options.

[5] The integration of Apertium phrase pairs into Moses has been tested with Moses revision 3739.

or the systems are trained on in-domain corpora (Sánchez-Cartagena et al., 2011b) and the texts to be translated are from a general (news) domain. A system built under this hybrid philosophy (Sánchez-Cartagena et al., 2011) was not outperformed by a statistically significant margin by any other participant of the shared translation task from the Sixth Workshop on Statistical Machine Translation (WMT 11)(Callison-Burch et al., 2011) for the Spanish–English language pair. We release the toolkit with the hope that it will be useful to other MT practitioners.

Acknowledgements

This work has been partially funded by Spanish Ministerio de Ciencia e Innovación through project TIN2009-14009-C02-01, by Generalitat Valenciana through grant ACIF/2010/174 from VALi+d programme, and by Universitat d'Alacant through project GRE11-20.

Bibliography

Brown, P. F., 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. ISBN 1-55860-324-7.

Callison-Burch, C., 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.

Dandapat, S., M. L. Forcada, D. Groves, S. Penkale, and A. Way. 2010. *OpenMaTrEx: A Free/Open-Source Marker-Driven Example-Based Machine Translation System*, pages 121–126. Berlin: Heidelberg: Springer. ISBN 978-3-642-14769-2.

Dugast, L., 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.

Dugast, L., 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.

Eisele, A., 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.

Forcada, M.L., 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* 25(2):127–144. Special Issue: Free/Open-Source Machine Translation.

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

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

Koehn, P., 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.

Koehn, P., F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference*, pages 48–54. Edmonton, Canada.

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

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

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

Sánchez-Cartagena, V. M., F. Sánchez-Martínez, and J. A. Pérez-Ortiz. 2011a. Enriching a statistical machine translation system trained on small parallel corpora with rule-based bilingual phrases. In *Proceedings of Recent Advances in Natural Language Processing*, pages 90–96. Hissar, Bulgaria.

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

Sánchez-Cartagena, V. M., F. Sánchez-Martínez, and J. A. Pérez-Ortiz. 2011. The universitat d'alacant hybrid machine translation system for wmt 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 457–463. Edinburgh, Scotland: Association for Computational Linguistics.

Schwenk, H., 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.

Simard, M., N. Ueffing, P. Isabelle, and R. Kuhn. 2007. Rule-based translation with statistical phrase-based post-editing. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 203–206. Prague, Czech Republic.

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 for Machine Translation*, pages 213–217.