From free shallow monolingual resources to machine translation systems: easing the task

Helena M. Caseli
NILC – ICMC
University of São Paulo
helename@icmc.usp.br

Maria das Graças V. Nunes
NILC – ICMC
University of São Paulo
gracan@icmc.usp.br

Mikel L. Forcada
DLSI
Universitat d'Alacant
mlf@ua.es

Abstract

The availability of machine-readable bilingual linguistic resources is crucial not only for machine translation but also for other applications such as cross-lingual information retrieval. However, the building of such resources demands extensive manual work. This paper describes a methodology to build automatically bilingual dictionaries and transfer rules by extracting knowledge from word-aligned parallel corpora processed with free shallow monolingual resources (morphological analysers and part-of-speech taggers). Experiments for Brazilian Portuguese–Spanish and Brazilian Portuguese–English parallel texts have shown promising results.

1 Introduction

Two of the main challenges in natural language processing (NLP) are (1) the production, maintenance and extension of computational linguistic resources and (2) the integration of these resources into NLP applications.

In particular, the availability of machine-readable bilingual linguistic resources is crucial not only for rule-based machine translation (RBMT) but also for other applications such as cross-lingual information retrieval. However, the building of such resources (bilingual single-word and multi-word correspondences, translation rules) demands extensive manual work. As a consequence, bilingual resources are usually more difficult to find than shallow monolingual resources such as morphological dictionaries or part-of-speech taggers.

In an attempt to overcome the lack of these bilingual resources, several methods have been proposed to build automatically translation grammars (McTait, 2003; Menezes and Richardson, 2001; Lavie et al., 2004; Carbonell et al., 2002) and bilingual dictionaries (Wu and Xia, 1994; Fung, 1995; Koehn and Knight, 2002; Langlais et al., 2001; Schafer and Yarowsky, 2002).

In line with some of these initiatives, this paper describes a methodology to build automatically both bilingual dictionaries and shallow-transfer rules. These resources are built by extracting knowledge from automatically word-aligned (or *lexically aligned*) parallel corpora which have been processed with shallow monolingual resources (morphological analysers and part-of-speech taggers). The free shallow monolingual resources used in these experiments are available as part of the `Apertium` open-source machine translation (MT) platform.[1]

This methodology is part of the ReTraTos project[2] which aims at inducing linguistic knowledge for Brazilian Portuguese (`pt`), Spanish (`es`) and English (`en`). The MT ex-

[1] `http://www.apertium.org`.

[2] `http://www.nilc.icmc.usp.br/nilc/projects/retratos.htm`

periments carried out for the pt–es and pt–en language pairs produced reasonable results as will be shown here.

There is a distinct advantage in the method proposed in this paper, as compared to other learning approaches to MT (such as statistical machine translation, SMT). It generates dictionaries and rules which may be edited by human experts to improve the performance of the resulting system, or even combined with data written by experts. In particular, there is an ongoing project to convert the data generated by our method to be freely used with Apertium. The induction software will also be distributed as open-source in a near future.

The main contribution of the proposed methods is a way to induce bilingual resources automatically from a parallel corpus using free monolingual resources and tools.

This paper is organized as follows. Section 2 presents related work on automatic induction of bilingual dictionaries and transfer rules. The proposed methods for inducing bilingual dictionaries and transfer rules are described in section 3. The experiments carried out with the pt–es and pt–en language pairs are described in section 4. The paper ends with some conclusions and proposals for future work (section 5).

2 Related work

In this section we present methods to induce automatically bilingual dictionaries (section 2.1) and transfer rules (section 2.2).

2.1 Induction of bilingual dictionaries

A bilingual dictionary —a bilingual list of words and multiword units that are mutual translations— is usually a by-product of a word alignment process (Brown et al., 1993; Och and Ney, 2000; Caseli et al., 2005).[3]

In (Wu and Xia, 1994), an English–Chinese dictionary was automatically induced by means of training a variant of the statistical model described in (Brown et al., 1993). This model was trained on a large corpus (about 3 million words) resulting in a set of about 6,500 English words (on average 2.33 possible Chinese translations for each English word). Evaluation through direct human inspection of a random set of 200 words showed an accuracy lying between 86.0% (completely automatic process) and 95.1% (manual correction).

By contrast, the method proposed by Fung (1995) uses a non-aligned Chinese–English parallel corpus (with about 5,760 English words) to induce bilingual entries for nouns and proper nouns based on co-occurrence (source and target) positions. Three judges evaluated 23.8% of the induced entries and the average accuracy was 73.1%.

This paper proposes a bilingual dictionary induction method based on automatic word alignment as explained in section 3.1.

2.2 Induction of translation rules

In the literature, methods for inducing transfer rules are based on many different approaches. However, all of them get a sentence-aligned parallel corpus (a set of translation examples) as input. The induced rules can, in turn, be used by the MT system to translate source sentences into target sentences.

The method proposed in (McTait, 2003) looks for transfer rules in two steps. In a monolingual step, the method looks for sequences of items that occur at least in two sentences by processing each side (source or target) separately —these sequences are taken as monolingual patterns. In the bilingual step, the method builds bilingual patterns following a co-occurrence criterion.[4] Finally, a bilingual similarity (distance) measure is used to set the alignment between source and target items that form a bilingual pattern. This method achieved 33.9% coverage, considering only full translations, in experiments with a training corpus of 2,500 and a test corpus of 1,000 pairs of en–fr (French) sentences.

The method proposed in (Menezes and Richardson, 2001) aligns the nodes of the

[3] An automatic word aligner is a tool for finding correspondences between words, and sometimes multiword units, in parallel texts.

[4] One source pattern and one target pattern occurring in the same pair of sentences are taken to be mutual translations.

source and target parse trees by looking for word correspondences in a bilingual dictionary. Then, following a best-first strategy (processing first the nodes with the best word correspondences), the method aligns the remaining nodes using a manually created alignment grammar composed of 18 bilingual compositional rules. After finding alignments between nodes of both parse trees, these alignments are expanded using linguistic constructs (such as noun and verb phrases) as context boundaries. Menezes and Richardson (2001) show that their system performed better than BabelFish[5] in 46.5% of test cases in experiments carried out with a training corpus of 161,606 and test corpora of 200-500 pairs of `es`–`en` sentences.

In (Lavie et al., 2004; Carbonell et al., 2002), the method infers hierarchical syntactic transfer rules, initially, on the basis of the constituents of both (manually) word-aligned languages. To do so, sentences from the language with more resources (English, in that case) are parsed and disambiguated. Value and agreement constraints[6] are determined from the syntactic structure, the word alignments and the source and target dictionaries. Lavie et al. (2004) show experiments carried out with RBMT and SMT systems trained with 17,589 lexically aligned sentences and phrases and tested with 258 sentences, for Hindi(`hi`)–`en`. The results show that the RBMT system scored better than the SMT: 11.2 BLEU and 5.32 NIST vs. 10.2 BLEU and 4.70 NIST.

Sánchez-Martínez and Forcada (2007) use an aligned parallel corpus to infer shallow-transfer rules based on the alignment templates approach by Och and Ney (2004). This research makes extensive use of the information in an existing manually-built bilingual dictionary to guide rule extraction. A

[5] `http://babelfish.altavista.com`.

[6] Value and agreement constraints specify which values (value constraints) the morphological features of source and target words should have (for instance, masculine as gender, singular as number and so on) and whether these values should be the same (agreement constraints).

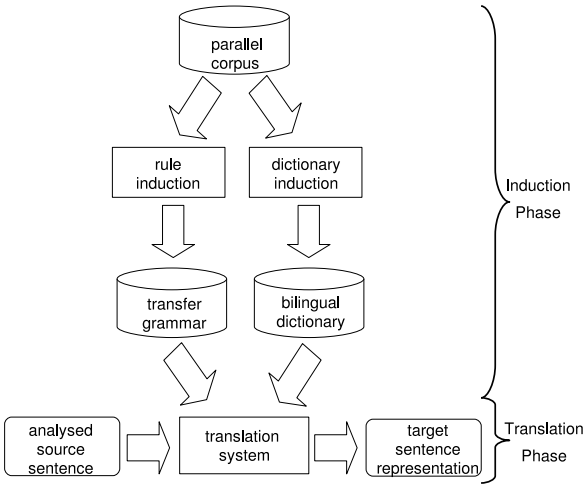

Figure 1: Scheme of proposed induction and translation systems

training corpus composed of 100,834 pairs of `es`–`ca` (Catalan) sentences and a test corpus of about 10,000 words were used to evaluate the induced rules. The evaluation carried out via post-edited translation shows a word error rate (WER) of 8.1–8.5% for automatically induced rules vs. 6.5–6.7% for hand-coded rules.

The method to induce transfer rules presented in this paper brings forth a new approach to induce and filter rules as described in section 3.2.

3 Induction and translation in the ReTraTos environment

The general scheme of the proposed induction and translation systems is shown in Figure 1. A PoS-tagged and word-aligned parallel corpus is given as input to our bilingual dictionary and transfer rule induction systems.

The induced sets of transfer rules (transfer grammar) and bilingual entries (bilingual dictionary) are then used by a shallow-transfer MT system to translate source sentences into target sentences.

The induction systems are introduced in the next two sections while the translation system is described in section 3.3.

3.1 Inducing the bilingual dictionary

A brief description of the bilingual dictionary induction process is presented in this section. For a more complete description see Caseli and Nunes (2007).

The bilingual dictionary induction process comprises the following steps: (1) the compilation of two bilingual dictionaries, one for each translation direction (one source–target and another target–source); (2) the merging of these two dictionaries; (3) the generalization of morphological attribute values in the bilingual entries; and (4) the treatment of morphosyntactic differences related to entries in which the value of the target gender/number attribute has to be determined from information that goes beyond the scope of the entry itself.[7]

3.2 Inducing the transfer rules

The transfer rule induction process is briefly described in this section and in detail in Caseli et al. (2008).

In contrast with other rule induction methods, our method follows an *alignment block* based approach. Specifically, it does not learn rules from the whole pairs (source language, target language) of aligned sentences but from sequences of contiguous word-aligned items[8] in these pairs: the alignment blocks.

Figure 2 shows the three types of alignment blocks considered in this approach: omissions (type 0), alignments preserving item order in sentence (type 1) and reorderings (type 2). In this figure, source and target items are accompanied by their positions in the source and target sentences. For example, the source items *a* and *b* are aligned to *a'*, *a"* and *b'* in a way that preserves item order; therefore, they form an alignment block of type 1. Furthermore, they are also part of an alignment block

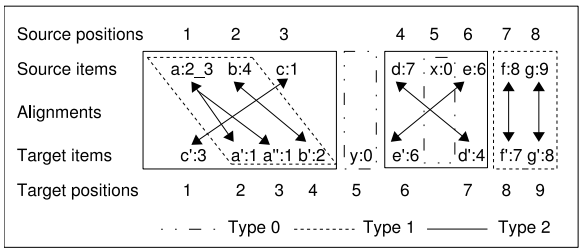

Figure 2: Types of alignment blocks

of type 2, since the source item *c* has a cross-link to *c'*.[9]

After building these alignment blocks, the rules are induced from each type separately, following four phases: (1) pattern identification, (2) rule generation, (3) rule filtering and (4) rule ordering.

First, analogously to (McTait, 2003), the bilingual patterns are extracted in two steps: monolingual and bilingual. In the monolingual step, source patterns are identified by an algorithm based on the Sequential Pattern Mining (SPM) technique and the `PrefixSpan` algorithm (Pei et al., 2004). In the bilingual step, the target items aligned to each source pattern are examined (in the parallel translation example) to form the bilingual pattern.

In pattern identification, the frequency threshold necessary to call a sequence of items a pattern is different for each type of alignment block (0, 1 or 2). This frequency threshold is calculated as a percentage p (an input parameter) of the total amount of blocks of each type.[10] The idea behind alignment-block-guided induction is that if we were using the same absolute frequency threshold for all types of alignments, very few relevant patterns coming from less frequent alignment types would be identified.

Second, the rule generation phase encompasses the building and the generalization of constraints between values on one (monolingual) or both (bilingual) sides of a bilingual

[7] For example, the `es` noun *tesis* (thesis) is valid for both number (singular and plural) and it has two possible `pt` translations: *tese* (singular) and *teses* (plural).

[8] For example, o/o<det><def><m>:5 is an item found in `pt` sentences where *o* (the) is the original word; o<det><def><m> is its lemma, PoS and morphological features; and 5 is the position of the word aligned with it. For details on how these information are obtained see section 4.1.

[9] Only alignment blocks of type 2 can include other alignment blocks (types 0 and 1).

[10] For example, suppose that we have 10,000 alignment blocks of type 1, 100 of type 2 and an input percentage of 15%. So, the frequency threshold for identifying patterns of type 1 is 1,500 while it is only 15 for patterns of type 2.

pattern. Constraints are derived from feature values (morphological information) in translation examples. Two kinds of constraints can be built —value constraints and agreement/value constraints— as in (Carbonell et al., 2002).[11]

Third, the induced rules are filtered to solve ambiguities. For all ambiguous rules —those with the same source side (sequence of source PoS tags) but different target sides— the filtering approach looks for source feature and lexical values which can distinguish the ambiguous rules.

Finally, the rule ordering specifies the order in which transfer rules should be applied by the MT system. It is done implicitly by setting the frequency and weight (the probability of its occurrence) of each rule, each target side and each constraint set.[12]

3.3 Translating sentences

The induced resources are used in the MT task by means of a simple translation system (see figure 1). The input of this system is an already analysed source sentence, that is, a sequence of source lexical forms (each one consisting of lemma, PoS tag and morphological inflection attributes).

The MT system implemented has two modes of translation: word-by-word and transfer. The former uses only the induced bilingual dictionary, while the latter uses both the induced dictionary and transfer rules.

In transfer mode, the system chooses and applies the best suitable transfer rules following a left-to-right longest-match procedure. The "best suitable rule" is the most frequent rule which: (i) matches the source sequence, (ii) matches a set of source constraints (there can be more than one) and (iii) this source constraint set is the most frequent.

[11]The value constraints here are the same as (Carbonell et al., 2002), but the agreement/value constraints are quite different from the agreement constraints used by them since, here, the morphological value is also specified in agreement/value constraint.

[12]The weight of a rule is calculated as its frequency divided by the total frequency of the whole set of rules. The weight of each target side and each constraint set are calculated in a similar way.

Unlike in `Apertium`, a backtracking approach is used in transfer translation: if a source pattern *abcd* matches the input sentence but cannot be applied because it has no compatible constraint, the system will try to apply the sub-pattern *abc*. This backtracking goes on until the sub-pattern has just one item and, in this case, word-by-word translation is applied.

4 Experiments and results

The next sections describe the corpora used to induce the linguistic resources (4.1) and the evaluation settings and results (4.2).

4.1 Preprocessing of bilingual corpora

The experiments described in this paper were carried out using two training parallel corpora. One corpus consists of 18,236 pairs of `pt`–`es` parallel sentences with 503,596 tokens in `pt` and 545,866 tokens in `es`. The other corpus consists of 17,397 pairs of `pt`–`en` parallel sentences with 494,391 tokens in `pt` and 532,121 tokens in `en`. Both corpora contain articles from the online version of a Brazilian scientific magazine, *Pesquisa FAPESP*.[13] It contains parallel texts written in `pt` (original), `en` (version) and `es` (version).

These corpora were PoS-tagged using the morphological analyser and the PoS tagger available in `Apertium` (Armentano-Oller et al., 2006). The morphological analysis provides one or more lexical forms for each surface form (the form as it appears in the text) using a monolingual morphological dictionary. The PoS tagger chooses the best possible lexical form based on a first-order hidden Markov model (HMM).

The original morphological dictionaries available at `Apertium` were enlarged in the ReTraTos project with entries from `Unitex`[14] (`pt` and `en`) and from the `es`–`ca` MT system `InterNOSTRUM`[15] (`es`).[16] So, the original morphological dictionaries for `pt` and `es` available

[13]`http://revistapesquisa.fapesp.br`.

[14]`http://www-igm.univ-mlv.fr/~unitex/`.

[15]`http://www.internostrum.com/`.

[16]The `es` new entries were provided by the Transducens group from the Universitat d'Alacant.

in the Apertium es–pt linguistic data package (version 0.9) were enlarged to cover 1,136,536 and 337,861 surface forms, respectively. The en morphological dictionary available in the Apertium en–ca linguistic data package (version 0.8) was enlarged to cover 61,601 surface forms.[17]

After PoS tagging, the translation examples were word-aligned using two different tools: LIHLA (Caseli et al., 2005) and GIZA++ (Och and Ney, 2000). Experiments have shown that LIHLA had a better alignment error rate (AER) performance than GIZA++ on pt–es parallel texts (5.39% AER vs. 6.35% AER). But GIZA++ had a better performance on pt–en (15.44% AER vs. 8.61% AER) (Caseli et al., 2008). The translation examples were aligned in both directions (source–target and target–source) and the alignments were merged using the union algorithm proposed by Och and Ney (2003).

4.2 Evaluation settings and results

The linguistic resources induced from the two parallel corpora described in section 4.1 were: one bilingual dictionary for each pair of languages and some sets of transfer rules induced using different input parameters.

The induced bilingual dictionaries have 23,450 pt–es entries and 19,191 pt–en entries. The best set of transfer rules was obtained using a percentage $p = 0.07\%$ to calculate the frequency thresholds for pattern identification of each block type. With this parameter, 1,421 pt–es transfer rules, 1,329 es–pt transfer rules, 647 pt–en transfer rules and 722 en–pt transfer rules were induced.

The corpus used to test/evaluate the induced resources consists of 649 parallel sentences from the same domain of the training corpus. The sentences in the test corpus were translated in the four possible directions (pt–es, es–pt, pt–en and en–pt). To evaluate the translations, a reference corpus was created consisting of the corresponding parallel sentences in the test corpus. For example, the reference corpus used to evaluate the translation from pt to es is composed by the es sentences in the test corpus.

The sentences translated automatically were compared automatically with those in the reference corpus by means of the indirect scores BLEU (Papineni et al., 2002) and NIST (Doddington, 2002).

In these experiments, we evaluated the sentences translated by the ReTraTos MT system (see section 3.3) by applying the induced resources in the word-by-word translation (RTT_word-by-word) and in the transfer translation (RTT_transfer). The word-by-word translation was used here with three purposes: (1) to be a baseline for comparison with other systems, (2) to evaluate the quality of the induced vocabulary, and (3) to measure the improvement brought by using transfer rules (RTT_transfer).

We also evaluated translations produced by other MT systems available for the studied languages. For pt–es–pt, we have used two versions of the es–pt data provided in the open-source MT platform Apertium: version 0.9.1, which will be called Apertium and version 0.9.2, using a larger dictionary, which will be called Apertium-P.[18] For pt–en–pt, we have used the MT systems: FreeTranslation,[19] Google[20] and BabelFish.

Table 1 shows the results of pt–es–pt translation. From these values, it is possible to notice that the ReTraTos MT system using only one (RTT_word-by-word) or both (RTT_transfer) the induced linguistic resources obtained scores that were slightly higher than Apertium's versions, with a more significant difference in the es–pt direction.

In the pt–es direction, when compared to Apertium-P, the RTT_transfer had an improvement of around 2 points in BLEU and 0.2 in NIST; while in the es–pt direction, this improvement was twice as large: 4 points in BLEU and 0.4 in NIST.

[17] Initially the pt, es and en morphological dictionaries covered 128,772, 116,804 and 48,759 surface forms, respectively.

[18] Version 0.9.2. was the one that could be tried online in April 2007 at http://xixona.dlsi.ua.es/prototype.

[19] http://www.freetranslation.com.

[20] http://translate.google.com.

Table 1: Evaluation of pt–es–pt MT

Lang.	System	BLEU	NIST
pt–es	RTT_transfer	65.13	10.85
	RTT_word-by-word	64.90	10.82
	Apertium	63.82	10.64
	Apertium-P	63.87	10.64
es–pt	RTT_transfer	66.66	10.98
	RTT_word-by-word	66.49	10.95
	Apertium	60.98	10.31
	Apertium-P	62.88	10.51

Table 2: Evaluation of pt–en–pt MT

Lang.	System	BLEU	NIST
pt–en	RTT_transfer	28.32	7.09
	RTT_word-by-word	26.06	6.77
	FreeTranslation	32.94	7.65
	BabelFish	31.61	7.46
	Google	32.95	7.61
en–pt	RTT_transfer	24.00	6.11
	RTT_word-by-word	23.24	6.02
	FreeTranslation	30.53	6.85
	BabelFish	36.66	7.68
	Google	31.21	6.88

The similar scores of the two versions of ReTraTos on pt–es–pt seem to be due to the greater coverage of the induced bilingual dictionary on the texts of the domain. This fact indicates that, for related languages such as pt and es, a greater coverage of the bilingual dictionary has a stronger impact in translation scores than the transfer rules.

Table 2 shows the results of pt–en–pt translation. In the evaluation for this pair of languages, the translations produced by the ReTraTos versions did not score so high as those for the pt–es pair. This result was already expected, since the transfer rule induction system was not designed to deal with more complex changes in the structure of translation, but simply agreement and position changes between close items.

However, it is worth noticing that the improvement attributed to the use of rules (RTT_transfer) compared to the word-by-word (RTT_word-by-word) translation in the pt–en–pt pair is greater (0.76–2.26 BLEU points and 0.09–0.32 NIST points) than in the pt–es–pt pair (less than 0.3 points in BLEU and 0.03 in NIST). This indicates that, albeit simple (in the sense that they perform only shallow changes), the induced rules may indeed improve word-by-word translation between more distant languages.

5 Conclusions and future work

In this paper we have described a methodology to build bilingual dictionaries and translation rules automatically from parallel corpora. The input corpora were processed using word aligners and shallow monolingual resources such as morphological analysers and PoS taggers.

One advantage of the method proposed here is that both the inferred dictionaries and the induced rules are written in formats that can be easily edited by humans or combined with manually written rules.

In particular, the rules can be easily converted to the formats used by the open-source MT platform Apertium, and the bilingual dictionary entries are already induced in the formalism used by Apertium. Thus, the induction systems presented in this paper can be used along with the tools and linguistic data distributed with Apertium to ease the task of building new MT systems.

As future work, we intend to finish an ongoing project to adapt the induced resources to Apertium and to implement a *open-source toolchain* for MT. This toolchain will join the already existing free resources from Apertium and from ReTraTos and make them freely available to produce new MT systems.

Other future work includes an evaluation by means of the WER using post-edited output as a reference. We also aim at testing different configurations of ReTraTos to determine to what extent changes in optional modules (rule filtering and rule ordering) affect translation quality. Experiments to compare the performance of the system presented here (using automatically induced transfer rules) and that of a SMT system trained and tested on the same corpora are already been carried out.

Acknowledgements

We thank the financial support of the Brazilian agencies FAPESP (02/13207-8 and 04/06707-0), CAPES (053804-3 and 407507-

2) and CNPq (550388/2005-2), and of the Spanish Ministry of Education and Science; in particular through MEC-CAPES project PHB2005-0052 (Spain) and 116/06 (Brazil).

References

C. Armentano-Oller, R. C. Carrasco, A. M. Corbí-Bellot, M. L. Forcada, M. Ginestí-Rosell, S. Ortiz-Rojas, J. A. Pérez-Ortiz, G. Ramírez-Sánchez, F. Sánchez-Martínez, and M. A. Scalco. 2006. Open-source Portuguese-Spanish machine translation. In *Proceedings of the VII PROPOR*, pages 50–59, Itatiaia-RJ, Brazil.

P. Brown, V. Della-Pietra, S. Della-Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–312.

J. Carbonell, K. Probst, E. Peterson, C. Monson, A. Lavie, R. Brown, and L. Levin. 2002. Automatic rule learning for resource-limited MT. In *Proceedings of AMTA-02*, volume 2499 of *LNCS*, pages 1–10, London, UK.

H. M. Caseli and M. G. V. Nunes. 2007. Automatic induction of bilingual lexicons for machine translation. *International Journal of Translation*, 19:29–43.

H. M. Caseli, M. G. V. Nunes, and M. L. Forcada. 2005. Evaluating the LIHLA lexical aligner on Spanish, Brazilian Portuguese and Basque parallel texts. *Procesamiento del Lenguaje Natural*, 35:237–244.

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

G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of ARPA WHLT*, pages 128–132, San Diego.

P. Fung. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of ACL-95*, pages 236–243.

P. Koehn and K. Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the Workshop of the ACL SIGLEX*, pages 9–16, Philadelphia.

P. Langlais, G. Foster, and G. Lapalme. 2001. Integrating bilingual lexicons in a probabilistic translation assistant. In *Proceedings of the 8th MT Summit*, pages 197–202, Santiago de Compostela, Spain.

A. Lavie, K. Probst, E. Peterson, S. Vogel, L. Levin, A. Font-Llitjós, and J. Carbonell. 2004. A trainable transfer-based machine translation approach for languages with limited resources. In *Proceedings of EAMT-04*, pages 1–8, Valletta, Malta.

K. McTait. 2003. Translation patterns, linguistic knowledge and complexity in an approach to EBMT. In M. Carl and A. Way, editors, *Recent Advances in Example-Based Machine Translation*, pages 1–28.

A. Menezes and S. D. Richardson. 2001. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the Workshop on Data-driven Machine Translation at ACL-01*, pages 39–46, Toulouse, France.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. In *Proceedings of the ACL-00*, pages 440–447, Hong Kong, China, October.

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

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

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the ACL-02*, pages 311–318, Philadelphia, PA.

J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M. Hsu. 2004. Mining sequential patterns by pattern-growth: the PrefixSpan approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(10):1–17.

F. Sánchez-Martínez and M. L. Forcada. 2007. Automatic induction of shallow-transfer rules for open-source machine translation. In *Proceedings of the TMI 2007*, pages 181–190.

C. Schafer and D. Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures an bridge languages. In *Proceedings of CoNLL-02*, pages 1–7.

D. Wu and X. Xia. 1994. Learning an English-Chinese lexicon from parallel corpus. In *Proceedings of the AMTA-94*, pages 206–213, Columbia, MD.