On the automatic learning of bilingual resources: some relevant factors for machine translation

Helena de M. Caseli and Maria das Graças V. Nunes and Mikel L. Forcada

[1] NILC – ICMC, University of São Paulo
CP 668P – 13.560-970 – São Carlos – SP – Brazil
e-mail: `helename,gracan@icmc.usp.br`
[2] Departament de Llenguatges i Sistemes Informàtics,
Universitat d'Alacant, E-03071 Alacant, Spain
e-mail: `mlf@ua.es`

Abstract. In this paper we present experiments concerned with automatically learning bilingual resources for machine translation: bilingual dictionaries and transfer rules. The experiments were carried out with Brazilian Portuguese (`pt`), English (`en`) and Spanish (`es`) texts in two parallel corpora: `pt`–`en` and `pt`–`es`. They were designed to investigate the relevance of two factors in the induction process, namely: (1) the coverage of linguistic resources used when preprocessing the training corpora and (2) the maximum length threshold (for transfer rules) used in the induction process. From these experiments, it is possible to conclude that both factors have an influence in the automatic learning of bilingual resources.

Keywords: Machine translation, bilingual resources, automatic learning, parallel corpora

1 Introduction

The ability to translate from one language to another has became not only a desirable but also a fundamental skill in the multilingual world every day accessible through the Internet. Due to working or social needs, there is a growing necessity of being able to get at least the gist of a piece of information in a different language. Unfortunately, this ability is not inherent.

The good news is that computers are getting more and more available to ease this task. This is one of the challenges of machine translation (MT) research and also of this paper. In particular, this paper is concerned not only with translating from one language to another but also with the automatic learning of bilingual resources useful for rule-based machine translation (RBMT) systems.

Traditionally, the bilingual resources used in RBMT systems are built by means of hard manual work. However, in the last years several methods have been proposed to automatically learn bilingual resources such as dictionaries [1–4] and transfer rules [5–7] from translation examples (parallel corpora). A bilingual dictionary is a bilingual list of words and multiword units (possibly accompanied by morphological information) that are mutual translations. A transfer (or translation) rule, in turn, is a generalization of structural, syntactic or lexical correspondences found in the parallel sentences (translation examples).

In line with these initiatives, this paper presents experiments carried out to investigate the relevance of certain factors when automatically learning bilingual dictionaries and transfer rules following the methodology of the ReTraTos project.[3]

Following this methodology, the bilingual dictionaries and the transfer rules are induced from automatically word-aligned (or lexically aligned) parallel corpora processed with morphological analysers and part-of-speech (PoS) taggers. The aspects under investigation in this paper are: (1) the coverage of linguistic resources used in preprocessing the training corpora and (2) the maximum length threshold used in the transfer rule induction process, that is, the number of source items that a transfer rule can contain. The evaluation of the translation quality was carried out with Brazilian Portuguese (`pt`), Spanish (`es`) and English (`en`) texts in two parallel corpora —`pt`–`es` and `pt`–`en`— by using the automatic metrics BLEU [8] and NIST [9].

Our interest in the above factors is motivated by previous results [10]. We now intend to investigate if a better coverage of the preprocessing linguistic resources can bring better MT results, and if the previous poor `pt`–`en` MT performance was due to the threshold length (too restrictive) used during the induction of the transfer rules. We think that the small length of source patterns was not sufficient to learn relevant syntactic divergences between those languages. To our knowledge, it is the first time that such a study is carried out for Brazilian Portuguese or other languages.

This paper is organized as follows. Section 2 presents related work on automatic induction of bilingual dictionaries and transfer rules. Section 3 describes briefly the induction methodology under study in this paper. Section 4 shows the experiments and their results, and section 5 ends this paper with some conclusions and proposals for future work.

2 Related work

According to automatic evaluation metrics like BLEU [8] and NIST [9], the phrase-based statistical MT (SMT) systems such as [11] and [12] are the state-of-the-art in MT. Despite this fact, this paper is concerned with RBMT since the symbolic resources (dictionaries and rules) of RBMT suit better than the models of SMT to our research purpose: to know how translation is performed and what affects its performance.

This section presents briefly some of the methods proposed in the literature to automatically induce bilingual dictionaries and transfer rules. These methods can follow many different approaches, but usually they perform induction from a sentence-aligned parallel corpus.

Usually, a bilingual dictionary is obtained as a by-product of a word alignment process [13–15]. In [1], for example, an English–Chinese dictionary was automatically induced by means of training a variant of the statistical model

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

described in [13]. By contrast, the method proposed in [2] uses a non-aligned parallel corpus to induce bilingual entries for nouns and proper nouns based on co-occurrence positions. Besides the alignment-based approaches, others have also been proposed in the literature such as [3] and [4]. While [3] builds a bilingual dictionary from unrelated monolingual corpora, [4] combines two existing bilingual dictionaries to build a third one using one language as a bridge.

The transfer rule induction methods also use the alignment information to help the induction process. For example, the method proposed in [6] uses shallow information to induce transfer rules in two steps: monolingual and bilingual. In the 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: one source pattern and one target pattern occurring in the same pair of sentences are taken to be mutual translations. Finally, a bilingual similarity (distance) measure is used to set the alignment between source and target items that form a bilingual pattern.

The method proposed in [7], by its turn, uses more complex information to induce rules. It aligns the nodes of the 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 defined 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.

The method in [16] 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 are set from the syntactic structure, the word alignments and the source/target dictionaries. Value constraints specify which values the morphological features of source and target words should have (for instance, masculine as gender, singular as number and so on). The agreement constraints, in turn, specify whether these values should be the same.

Finally, in [17] the authors used an aligned parallel corpus to infer shallow-transfer rules based on the alignment templates approach [12]. This research makes extensive use of the information in an existing manually-built bilingual dictionary to guide rule extraction.

3 Induction and translation in the ReTraTos environment

The general scheme of the induction and translation in the ReTraTos environment is shown in Figure 1. The input for both induction systems (bilingual dictionary and transfer rule) is a PoS-tagged and word-aligned parallel corpus. After having been induced, the resources —transfer grammar and bilingual dictionary— are used to translate source sentences into target sentences.

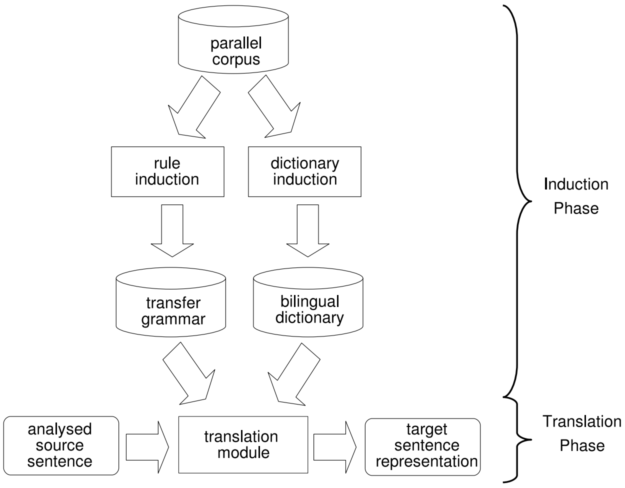

Fig. 1. Scheme of the induction and translation phases in the ReTraTos environment

3.1 Induction in the ReTraTos environment

The induction processes are briefly described in this section. A more detailed description of these processes can be found in [10] and [18].

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 in one specifying the valid translation direction if necessary; (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 bilingual entry itself.[4]

The rule induction method, in turn, induces the transfer rules following four phases: (1) pattern identification, (2) rule generation, (3) rule filtering and (4) rule ordering. Firstly, similarly to [6], 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* technique and the PrefixSpan algorithm [19]. In the bilingual step, the target items aligned to each source pattern are looked for (in the parallel translation example) to form the bilingual pattern.

Secondly, the rule generation phase encompasses: (a) the building of constraints between morphological values on one (monolingual) or both (bilingual) sides of a bilingual pattern and (b) the generalization of these constraints. Two

[4] 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).

kinds of constraints can be built: value constraints and agreement/value constraints. A value constraint specifies which values are expected for the features on each side of a bilingual pattern. An agreement/value constraint, in turn, specifies which items on one or both sides have the same feature values (agreement constraint) and which are these values (value constraint).[5]

Thirdly, the induced rules are filtered to solve ambiguities. An ambiguous rule has the same sequence of PoS tags in the source side, and different PoS tag sequences in the target side. To decide, the filtering module looks for morphological or lexical values, which could distinguish them. For example, it could be possible to distinguish between two ambiguous rules with "`n adj`" (noun adjective) as their sequence of source PoS tags finding out that one rule was induced from examples with feminine nouns and the other, from masculine nouns as stated in their constraints.

Finally, the rule ordering specifies the order in which transfer rules should be applied. It is done implicitly by setting the occurrence frequency of each rule, each target side and each constraint set. The occurrence frequency of a rule is the number of times its source sequence of PoS tags was found in the training corpus. Then, for each rule, the occurrence frequency of a target side (constraint set) is the number of times this target side (constraint set) was found for this specific rule, in the training corpus. The frequencies are used to choose the "best suitable rule" as explained in the next section.

3.2 Translation in the ReTraTos environment

As shown in Figure 1, the induced sets of transfer rules (transfer grammar) and bilingual entries (bilingual dictionary) are used by the MT module to translate an already analysed source sentence into a representation of a target sentence. The analysed source sentence and the representation of a target sentence are both sequences of *lexical forms*, each composed of a lemma, PoS tag, and morphological information.

The ReTraTos MT module, besides looking for translations in the bilingual dictionary, applies the best suitable transfer rules following a left-to-right longest-match procedure. The "best suitable rule" is the most frequent rule which: (a) matches the source pattern (sequence of source PoS tags), (b) matches one of the associated sets of source constraints (there can be more than one) and (c) the matching source constraint set is the most frequent. Therefore, the selected rule might not be the most frequent. Finally, if there is more than one "best suitable rule", the rule defined first is the one chosen.

A backtracking approach is used in rule application: 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 to one word and the process restarts immediately after it.

[5] Our *value* constraints are like in [16], but our *agreement/value* constraints are different from their *agreement* constraints since, here, the values are explicitly defined.

4 Experiments and results

The training and test/reference parallel corpora used in these experiments are described in section 4.1. The training corpora were used to induce the bilingual resources while the test/reference corpora were used to evaluate the performance of the translation based on the induced rules. The effect of the investigated factors in the induction of bilingual resources was measured by means of the BLEU [8] and NIST [9] measures, which give an indication of translation performance. Both take into account, in different ways, the number of n-grams common to the automatically translated sentence and the reference sentence, and estimate the similarity in terms of length, word choice and order.

A baseline configuration was defined as the one resulting from the ReTraTos project and explained in section 4.2. The investigated factors were confronted with the baseline configuration as explained in the following sections: (4.3) the coverage of the preprocessing resources and (4.4) the maximum length threshold.

4.1 Parallel corpora

The experiments described in this paper were carried out using the training and the test/reference `pt–es` and `pt–en` parallel corpora. These corpora contain articles from a Brazilian scientific magazine, *Pesquisa FAPESP*.[6]

The training corpora were preprocessed as explained below. First, they were automatically sentence-aligned usingan implementation of the Translation Corpus Aligner [20]. Then, both corpora were PoS-tagged using the morphological analyser and the PoS tagger available in the `Apertium`[7] open-source machine translation platform. The morphological analysis provides one or more *lexical forms* or analyses (information on lemma, lexical category and morphological inflection) for each surface form using a monolingual morphological dictionary (MD). The PoS tagger chooses the best possible analysis based on a first-order hidden Markov model (HMM) [21].

To improve the coverage of the morphological analyser, the original MDs were enlarged with entries from `Unitex`[8] (`pt` and `en`) and from `interNOSTRUM`[9] (`es`).[10] The number of surface forms covered by the original and the extended versions of the morphological dictionaries are: 128,772 vs. 1,136,536 for `pt`, 116,804 vs. 337,861 for `es` and 48,759 vs. 61,601 for `en`. This extension decreased the percentage of unknown words in the corpora used in our experiments from 11% to 3% for `pt`, from 10% to 5% for `es` and from 9% to 5% for `en`. The effect of this

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

[7] The open-source machine translation platform `Apertium`, including linguistic data for several language pairs and documentation, is available at `http://www.apertium.org`.

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

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

[10] The original MDs are available as part of the `apertium-es-pt` (version 0.9) and the `apertium-en-ca` (version 0.8). The new `es` entries derived from `interNOSTRUM` were provided by the Transducens research group from the Universitat d'Alacant.

increase of coverage in the induction of bilingual resources is one of the factors under investigation in this paper (see section 4.3).

Finally, the translation examples were word-aligned using LIHLA [15] for pt–es and GIZA++ [14] for pt–en as explained in [10]. The translation examples were aligned in both directions and the union was obtained as in [22].

The pt–es training corpus consists of 18,236 pairs of parallel sentences with 503,596 tokens in pt and 545,866 in es. The pt–en training corpus, in turn, has 17,397 pairs of parallel sentences and 494,391 tokens in pt and 532,121 in en.

The *test corpus* consists of 649 parallel sentences with 16,801 tokens in pt, 17,731 in es and 18,543 in en (about 3.5% of the size of training corpora). The pt–es and pt–en *reference corpora* were created from the corresponding parallel sentences in the test corpus.

4.2 The baseline

The baseline configuration used in our experiments is the automatic translation produced by ReTraTos (see section 3.2) based on the parallel corpora preprocessed with the *extended MDs* and using a *maximum length threshold* for rule induction set to *5*. Table 1 shows the values of BLEU and NIST for the translation performed in all possible directions and using the bilingual resources induced for both pairs of languages.

Table 1. Values of BLEU and NIST for ReTraTos translation

	pt–es		es–pt		pt–en		en–pt	
	BLEU	**NIST**	**BLEU**	**NIST**	**BLEU**	**NIST**	**BLEU**	**NIST**
baseline	65.13	10.85	66.66	10.98	28.32	7.09	24.00	6.11

4.3 The coverage of preprocessing resources

The first factor investigated in this paper is related to the coverage of the MDs used in the morphological analysis of the training corpora. The effect of using the richer/extended MDs (the baseline) instead of the original ones (see section 4.1) was evaluated in terms of (1) the increase in the number of entries in the induced bilingual dictionary and (2) the effect on overall translation performance.

The number of bilingual entries induced from the training corpora preprocessed with the extended MDs was *twice* as large as using the original ones: 7,862 vs. 4,143 (pt–es) and 10,856 vs. 5,419 (pt–en). The effect on the overall translation was measured by means of BLEU and NIST as shown in Table 2.

From Table 2, it is possible to notice that, for the pt–es language pair, the extended MDs brought about an increase in the measures in relation to the original MDs: 1.96–3.23 points in BLEU and 0.22–0.36 points in NIST. About 23–24% of the sentences in pt–es test corpus were translated differently by the

Table 2. Values of BLEU and NIST according to the original and the extended MDs

MD	pt–es		es–pt		pt–en		en–pt	
	BLEU	**NIST**	**BLEU**	**NIST**	**BLEU**	**NIST**	**BLEU**	**NIST**
Extended	65.13	10.85	66.66	10.98	28.32	7.09	24.00	6.11
Original	63.17	10.63	63.43	10.62	28.12	7.00	23.72	6.08

resources induced from each version of MD. The improvement in pt–es language pair is mainly due to the new transfer rules generated by the extended MDs —293 pt–es and 258 es–pt— since the translations obtained using just the bilingual dictionaries (the word-by-word translation) give only a small improvement of 0.21–1.61 points in BLEU and 0.24–0.28 points in NIST.

On the other hand, the values of BLEU and NIST are roughly the same for pt–en increasing just 0.20–0.28 points in BLEU and 0.03–0.09 points in NIST. Although 17–21% of the sentences in pt–en test corpus were translated differently by the resources induced from each version of MD the values of BLEU and NIST do not change significantly. For pt–en, the new 167 pt–en and 159 en–pt transfer rules did not lead to better values of these measures.

From these values it is possible to conclude that more morphological information has a larger effect on related language pairs, but it does not seem to have influence on the translation for pairs of more distant languages. This fact shows that for more distant languages more than just an improvement in preprocessing linguistic resources is needed to achieve better values for BLEU and NIST.

4.4 The maximum length threshold

An experiment was also carried out to investigate the effect of the maximum length threshold (maximum number of source items that a transfer rule can contain) used in the transfer rule induction process. Thus, the translations produced by ReTraTos based on rules induced using different maximum length thresholds (and the dictionaries generated by the baseline configuration) were analysed.

Table 3 shows the values of BLEU and NIST when the maximum length threshold was set to 3, 4, 5 (the baseline), 6, 7, 8 and 10. It is possible to see that the best maximum length threshold seems to be between 4 and 5 for both language pairs. The similar values of these metrics reflect the small percentage of sentences translated differently when the rules induced using threshold lengths of 4 and 5 were applied: 2–3% for pt–es, es–pt and en–pt and 12% for pt–en.

It is also important to say that although several new transfer rules have been generated when the length threshold increased from 4 to 5 —334 (pt–es), 337 (es–pt), 102 (pt–en) and 113 (en–pt)— less than a half of them were applied to translate the test corpora —45% (pt–es), 28% (es–pt), 30% (pt–en), and 13% (en–pt). Thus, these new rules seem not to be useful to improve translation performance as measured by BLEU and NIST.

Table 3. Values of BLEU and NIST according to the maximum length threshold

Max. length	pt–es		es–pt		pt–en		en–pt	
	BLEU	**NIST**	**BLEU**	**NIST**	**BLEU**	**NIST**	**BLEU**	**NIST**
3	63.40	10.86	65.08	10.91	27.23	7.07	23.48	6.11
4	65.15	10.85	66.68	10.98	28.56	7.12	24.01	6.11
5	65.13	10.85	66.66	10.97	28.32	7.08	24.00	6.11
6–8	63.42	10.86	65.05	10.91	27.19	7.07	23.64	6.06
10	65.17	10.85	66.68	10.97	28.35	7.10	24.00	6.11

5 Conclusions and future work

In this paper we described experiments carried out to investigate the effect of certain factors in the automatic induction of bilingual resources useful for RBMT considering the ReTraTos methodology. The factors under investigation were: (1) the coverage of preprocessing resources and (2) the maximum length threshold used in transfer rule induction process. From the experiments carried out on `pt–es` and `pt–en` language pairs it is possible to draw some conclusions.

First, a great coverage of the morphological dictionaries used in preprocessing the training corpora improved the translation generated based on the induced resources mainly for the `pt–es` pair: 1.96–3.23 points in BLEU and 0.22–0.36 points in NIST. For `pt–en` it seems that more than an improvement in lexical coverage is needed to improve the MT performance since the values increased just 0.20–0.28 points in BLEU and 0.03–0.09 points in NIST. Second, the maximum length threshold used in transfer rule induction process also proved to have a small influence in the translation performance and the best thresholds were 4 and 5. However, thresholds bigger than 5 did not bring measurable improvements.

Future work includes the design of new experiments to investigate the effect of other factors in the transfer rule induction process, such as the application (or not) of rule filtering and rule ordering. We also want to evaluate the translations by means of the word error rate (WER) using post-edited output as a reference. Finally, we are already carrying out experiments to compare the performance of the system presented here (using automatically induced resources) and that of a SMT system trained and tested on the same corpora.

Acknowledgements

We thank the financial support of the Brazilian agencies FAPESP, CAPES and CNPq, and of the Spanish Ministry of Education and Science.

References

1. Wu, D., Xia, X.: Learning an English-Chinese lexicon from parallel corpus. In: Proc. of AMTA-94, Columbia, MD (October 1994) 206–213
2. Fung, P.: A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In: Proc. of ACL-95. (1995) 236–243

3. Koehn, P., Knight, K.: Learning a translation lexicon from monolingual corpora. In: Proc. of SIGLEX-02, Philadelphia (July 2002) 9–16
4. Schafer, C., Yarowsky, D.: Inducing translation lexicons via diverse similarity measures an bridge languages. In: Proc. of CoNLL-02. (2002) 1–7
5. Kaji, H., Kida, Y., Morimoto, Y.: Learning translation templates from bilingual text. In: Proc. of COLING-92. (1992) 672–678
6. McTait, K.: Translation patterns, linguistic knowledge and complexity in an approach to EBMT. In Carl, M., Way, A., eds.: Recent Advances in EBMT. Kluwer Academic Publishers, Printed in Netherlands (2003) 1–28
7. Menezes, A., Richardson, S.D.: A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In: Proc. of the Workshop on Data-driven Machine Translation at ACL-01, Toulouse, France (2001) 39–46
8. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proc. of ACL-02. (2002) 311–318
9. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proc. of ARPA Workshop on Human Language Technology, San Diego (2002) 128–132
10. Caseli, H.M., Nunes, M.G.V., Forcada, M.L.: Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. Machine Translation **20**(4) (2006) 227–245
11. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proc. of HLT/NAACL. (2003) 127–133
12. Och, F.J., Ney, H.: The alignment template approach to statistical machine translation. Computational Linguistics **30**(4) (2004) 417–449
13. Brown, P., Della-Pietra, V., Della-Pietra, S., Mercer, R.: The mathematics of statistical machine translation: parameter estimation. Computational Linguistics **19**(2) (1993) 263–312
14. Och, F.J., Ney, H.: Improved statistical alignment models. In: Proc. of ACL-00, Hong Kong, China (October 2000) 440–447
15. Caseli, H.M., Nunes, M.G.V., Forcada, M.L.: Evaluating the LIHLA lexical aligner on Spanish, Brazilian Portuguese and Basque parallel texts. Procesamiento del Lenguaje Natural **35** (2005) 237–244
16. Carbonell, J., Probst, K., Peterson, E., Monson, C., Lavie, A., Brown, R., Levin, L.: Automatic rule learning for resource-limited MT. In: Proc. of AMTA-02. Volume 2499 of LNCS., London, UK, Springer-Verlag (2002) 1–10
17. Sánchez-Martínez, F., Forcada, M.L.: Automatic induction of shallow-transfer rules for open-source machine translation. In: Proc. of TMI-07. (2007) 181–190
18. Caseli, H.M., Nunes, M.G.V.: Automatic induction of bilingual lexicons for machine translation. International Journal of Translation **19** (2007) 29–43
19. Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., Hsu, M.: Mining sequential patterns by pattern-growth: the PrefixSpan approach. IEEE Transactions on Knowledge and Data Engineering **16**(10) (October 2004) 1–17
20. Hofland, K.: A program for aligning English and Norwegian sentences. In Hockey, S., Ide, N., Perissinotto, G., eds.: Research in Humanities Computing, Oxford, Oxford University Press (1996) 165–178
21. Armentano-Oller, C., Carrasco, R.C., Corbí-Bellot, A.M., Forcada, M.L., Ginestí-Rosell, M., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Scalco, M.A.: Open-source Portuguese-Spanish machine translation. In: Proc. of PROPOR-06, Itatiaia-RJ, Brazil (2006) 50–59
22. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29**(1) (2003) 19–51