Transfer rule and bilingual dictionary automatic induction in the ReTraTos project

Helena de M. Caseli[♣♠] and Maria das Graças V. Nunes[♢♠]
[♣]Department of Computer Science, Federal University of São Carlos (Brazil)
[♢]ICMC, University of São Paulo (Brazil)
[♠]NILC – CP 668P – 13.560-970 – São Carlos – SP – Brazil
helenacaseli@dc.ufscar.br,gracan@icmc.usp.br

Abstract. In this paper we present the ReTraTos methodology to automatically induce bilingual resources —transfer rules and bilingual dictionaries— from parallel corpora. These resources are very useful in Machine Translation (MT) and other bilingual Natural Language Processing (NLP) applications. As a result, several automatic approaches have been proposed to avoid the extensive hard work employed to manually build these resources. The automatic approaches described in this paper aim at building bilingual dictionaries and shallow-transfer rules by extracting knowledge from word-aligned and part-of-speech tagged parallel corpora. Experiments carried out with Brazilian Portuguese–Spanish and Brazilian Portuguese–English parallel texts show that the proposed methodologies can speed the development of these valuable computational resources and, thus, help the development of MT systems for new pairs of languages. Furthermore, the rule induction methodology is innovative in the way rules are identified and filtered.
http://www.dc.ufscar.br/~helenacaseli/pdf/2007/TeseDoutorado.pdf[1]

1 Introduction

Machine Translation (MT) is one of the oldest and most important Natural Language Processing (NLP) applications. Since its beginnings several methods have been proposed ranging from the basic level —in which MT is performed by just replacing words in a source language by words in a target language— to more sophisticated ones —which rely on manually created translation rules (Rule-based Machine Translation, RBMT) or automatically generated statistical models (Statistical Machine Translation, SMT).

On the one hand, RBMT is performed based on some machine-readable bilingual linguistic resources such as bilingual single-word and multi-word correspondences, transfer rules and go on. On the other hand, SMT is performed based on language and translation models learned from a huge parallel corpus, that is, a set of translation examples, usually sentences. These bilingual resources

[1] This paper has a strong similarity with [1] since both works describe the ReTraTos project methodology and results.

(dictionaries, rules or corpora) demand extensive manual work to be built. In an attempt to overcome this bottleneck, several automatic methods have been proposed to build bilingual dictionaries [2–5] and translation grammars [6–9].

In this paper we describe the methods proposed in ReTraTos[2] project, which build bilingual dictionaries and shallow-transfer rules from lexically aligned and part-of-speech tagged parallel corpora. The proposed approaches were tested in Brazilian Portuguese (`pt`), Spanish (`es`) and English (`en`) parallel texts. To our knowledge, ReTraTos is the first project to study the automatic induction of bilingual resources for Brazilian Portuguese, and with reasonable results.

This paper is organized as follows. Section 2 presents related work on automatic induction of bilingual dictionaries and transfer rules. The induction methods of ReTraTos are briefly described in Section 3 and Section 4 presents some experiments carried out with them and `pt`–`es` and `pt`–`en` parallel corpora. This paper ends with some conclusions and proposals for future work (Section 5).

2 Related Work

Usually, a bilingual dictionary is obtained as a by-product of an automatic word alignment process [10–12]. In [2], for example, an English–Chinese dictionary was automatically induced by means of training a variant of the statistical model described in [10]. By contrast, the method proposed in [3] 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 [4] and [5]. While [4] builds a bilingual dictionary from unrelated monolingual corpora, [5] 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.

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

The method in [8] 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.

This paper presents an innovative approach to induce transfer rules regarding the way the rules are identified —based on the alignment blocks— and filtered, as explained in Section 3.

3 ReTraTos Environment

Figure 1 shows the general scheme to the induction and translation phases in the ReTraTos environment. The input for both induction methods (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.

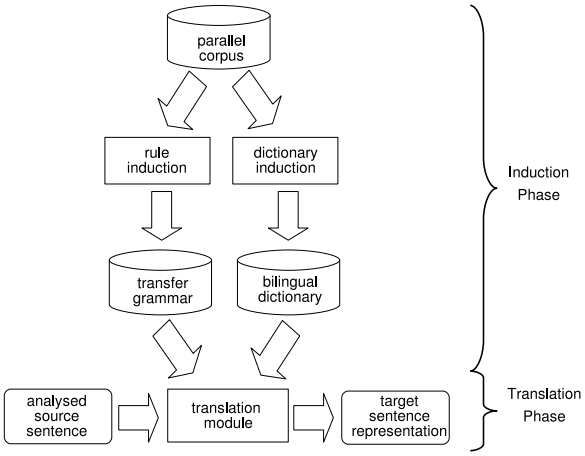

Fig. 1. Scheme of the induction and translation phases in the ReTraTos environment

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.[3]

[3] 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).

The rule induction method, in turn, is performed based on *alignment blocks*: sequences of aligned items in the translation examples. Figure 2 shows the three types of alignment blocks: omission (type 0), alignment preserving item order in sentence (type 1) and reordering (type 2).

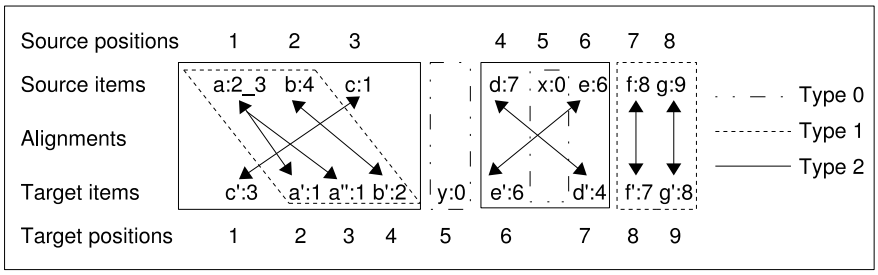

Fig. 2. Types of alignment blocks

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, thus, they form an alignment block of type 1.[4] Furthermore, they are also part of an alignment block of type 2, since the source item *c* has a cross-link to *c'*.[5]

Just as alignment templates [13], these alignment blocks are designed to define the scope for searching patterns. However, alignment blocks are quite different from alignment templates mainly in the way they are built. Whereas alignment blocks are built based on the type of alignment between items, the alignment templates are built on the basis of statistical criteria that do not take into account the type of alignment between items. The assumption behind inducing rules based on alignment blocks is that dealing with each type of alignment separately allows for the identification of relevant patterns even from less frequent alignment types (0 and 2) since patterns are identified based on the number of alignment blocks. After building these alignment blocks, the rules are induced from each type separately, following the 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 [14]. 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 kinds of constraints can be built: value constraints and agreement/value con-

[4] The alignment from *a* to *a'* and *a"* is an example of the opposite of omission, since one source item gives rise to two target items.

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

straints. 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).[6]

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. If it is not possible to find a value to distinguish the several target PoS tag sequences, the most frequent one is chosen.

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.

A more detailed description of the induction processes described in this section can be found in [15] and [16].

4 Experiments and Results

The experiments described in this paper were carried out using the training and the test/reference `pt`–`es` and `pt`–`en` parallel corpora composed of articles from a Brazilian scientific magazine, *Pesquisa FAPESP*.[7]

The training corpora were preprocessed in a three-step process. First, they were automatically sentence-aligned using an implementation of the Translation Corpus Aligner [17]. Then, both corpora were PoS-tagged using the morphological analyser and the PoS tagger available in `Apertium`[8] based on an enlarged version of the Morphological Dictionaries [15, 18]. Finally, the translation examples were word-aligned using `LIHLA` [12] for `pt`–`es` and `GIZA++` [11] for `pt`–`en`. For details of these preprocessed process see [15] and [18]. The resulting `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`.

From these training corpora two bilingual dictionaries and several different configurations of transfer rules were derived considering distinct input param-

[6] Our *value* constraints are like in [8], but our *agreement/value* constraints are different from their *agreement* constraints since, here, the values are explicitly defined.

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

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

eters. One bilingual dictionary was induced for each language pair: one with 23,450 pt–es entries and another with 19,191 pt–en entries. The best configuration for transfer rule induction resulted in 1,421 pt–es, 1,329 es–pt, 647 pt–en and 722 en–pt transfer rules.

To evaluate the performance of the MT based on the induced resources, we use a *test corpus* composed 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. The performance was measured by means of the BLEU [19] and NIST [20] 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.

In these experiments, we evaluated the sentences translated by the ReTraTos MT system by using only the induced dictionary in the word-by-word translation (ReTraTos_word-by-word) or also the transfer rules in the transfer translation (ReTraTos_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.[9] For pt–en–pt, we have used the MT systems: FreeTranslation,[10] BabelFish[11] and Google[12] translators.

Table 1. Evaluation of pt–es–pt and pt–en–pt MT

	pt–es		es–pt		pt–en		en–pt	
System	**BLEU**	**NIST**	**BLEU**	**NIST**	**BLEU**	**NIST**	**BLEU**	**NIST**
ReTraTos_transfer	65.13	10.85	66.66	10.97	28.32	7.09	24.00	6.11
ReTraTos_word-by-word	64.90	10.82	66.49	10.95	26.06	6.77	23.24	6.02
Apertium	63.82	10.64	60.98	10.30	–	–	–	–
Apertium-P	63.87	10.64	62.88	10.51	–	–	–	–
FreeTranslation	–	–	–	–	32.94	7.65	30.53	6.84
BabelFish	–	–	–	–	31.61	7.46	36.66	7.68
Google	–	–	–	–	32.95	7.61	31.21	6.88

Table 1 shows the results of MT evaluation. From these values, for pt–es–pt, it is possible to notice that the ReTraTos MT system using only one (ReTraTos_word-by-word) or both (ReTraTos_transfer) the induced linguistic resources performed a little better than Apertium's versions. In the pt–es direction, when compared to Apertium-P, the ReTraTos_transfer had an im-

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

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

[11] http://babelifish.altavista.com.

[12] http://www.google.com.br/language-tools.

provement of around 2% in BLEU and NIST; while in the es–pt direction, this improvement was of 6% in BLEU and 4% in NIST.

The similar performances of the two versions of ReTraTos (transfer and word-by-word) seem to be due to the greater coverage of the induced bilingual dictionary on the texts of the domain. From this fact we can conclude that, for related languages such as pt and es, a greater coverage of the bilingual dictionary has a stronger impact in translation than the transfer rules.

In the evaluation of pt–en–pt pair of languages, the translation produced by the ReTraTos versions were not so good 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, very frequent when translating from more distant languages such as pt and en.

However, it is worth noticing that the improvement attributed to the use of rules compared to the word-by-word translation in the pt–en–pt pair is greater (3-8% in BLEU and 1-4% in NIST) than in the pt–es–pt pair (less than 1% in both measures). This means that, albeit simple (in the sense that they perform only shallow changes), the induced rules can significantly improve word-by-word translation between more distant languages. To deal with this far distance between languages, we intend to induce new rules using also syntatic information and measure the gain in MT performance, if any.

5 Conclusions and Future Work

Following the ReTraTos methodology we found that the bilingual resources inferred for Brazilian Portuguese–Spanish and Brazilian Portuguese–English language pairs together with the monolingual resources used to infer them can be combined to build a promising first version of a shallow-transfer MT system.Thus, the proposed induction methods are a promise to a fast development of bilingual resources and MT systems, since only a parallel corpora —preprocessed as explained in Section 4— is required to produce bilingual dictionaries and transfer rules for any pair of languages.[13]

As future work, we intend to develop a new version of our transfer rule induction method which will be based on syntactic information together with part-of-speech tags. In this new version we aim at coping with the problems found in the current version when translating from/to pt to/from en.

Acknowledgements

We thank the financial support of FAPESP, CAPES and CNPq.

References

1. Caseli, H.M., Nunes, M.G.V.: Automatic induction of bilingual resources for machine translation: the ReTraTos project. In: Proc. of the VI Concurso de Teses e Dissertações em Inteligência Artificial (CTDIA). (2008) 1–10

[13] The ReTraTos induction methods are freely available as open-source at: http://sourceforge.net/projects/retratos.

2. Wu, D., Xia, X.: Learning an English-Chinese lexicon from parallel corpus. In: Proc. of AMTA-94, Columbia, MD (October 1994) 206–213
3. 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
4. Koehn, P., Knight, K.: Learning a translation lexicon from monolingual corpora. In: Proc. of SIGLEX-02, Philadelphia (July 2002) 9–16
5. Schafer, C., Yarowsky, D.: Inducing translation lexicons via diverse similarity measures an bridge languages. In: Proc. of CoNLL-02. (2002) 1–7
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 (ACL-01), Toulouse, France (2001) 39–46
8. 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
9. 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
10. 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
11. Och, F.J., Ney, H.: Improved statistical alignment models. In: Proc. of ACL-00, Hong Kong, China (October 2000) 440–447
12. 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
13. Och, F.J., Ney, H.: The alignment template approach to statistical machine translation. Computational Linguistics **30**(4) (2004) 417–449
14. 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
15. 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
16. Caseli, H.M., Nunes, M.G.V.: Automatic induction of bilingual lexicons for machine translation. International Journal of Translation **19** (2007) 29–43
17. 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
18. Caseli, H.M., Nunes, M.G.V., Forcada, M.L.: On the automatic learning of bilingual resources: Some relevant factors for machine translation. In: Proc. of the 19th Brazilian Symposium on Artificial Intelligence (SBIA). Volume 5249., Springer Berlin / Heidelberg (2008) 258–267
19. 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
20. 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