Proceedings of the International Joint Conference IBERAMIA/SBIA/SBRN 2006 - 3rd Workshop on MSc dissertations and PhD thesis in Artificial Intelligence (WTDIA'2006), Ribeirão Preto, Brazil, October 23–28, 2006. CD-ROM. ISBN 85-87837-11-7

Automatic transfer rule induction from parallel corpora

Helena de Medeiros Caseli and Maria das Graças Volpe Nunes

NILC – Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo, CP 668P, 13560-970 São Carlos-SP, Brazil
http://www.nilc.icmc.usp.br

Abstract. Recently, many projects have been proposed aiming at automatically transforming the multilingual information available on parallel texts into linguistic knowledge useful for machine translation. This paper describes an ongoing PhD project in which the main goal is to automatically induce transfer rules and bilingual dictionaries from part-of-speech tagged and lexically aligned parallel corpora. The final goal of this project is to use the induced rules and bilingual entries to translate from (to) Brazilian Portuguese to (from) Spanish and English.
Keywords: transfer rules, automatic induction, machine translation, parallel corpora

1 Introduction

Machine translation (MT) is a hard task, mainly due to the great need of deep linguistic knowledge about two (or more) languages to build resources such as translation grammars and bilingual dictionaries. Even statistical machine translation (SMT) is, sometimes, difficult to be performed once it requires large aligned parallel corpora, impossible to be built for some languages.

The lack of linguistic resources and even the difficulty to build them limit translation systems, for example, to their application domain. So, it is necessary to transform the scarce multilingual information into linguistic knowledge useful for machine translation.

In this context, several methods have been proposed aiming at automatically inducing structural, syntactic or lexical correspondences from parallel texts. These correspondences are, then, generalized to build translation grammars (a set of translation/transfer rules) and other useful resources (such as bilingual dictionaries) for MT systems.

The ReTraTos project, subject of this paper, is also concerned with this goal, more specifically: to induce linguistic knowledge useful for translation – transfer rules and bilingual dictionary – combining different techniques that better fit Brazilian Portuguese (`pt`). We intend to carry out experiments on Spanish (`es`) and English (`en`) – each one belonging to a different language family – forming two language pairs: `pt`–`es` and `pt`–`en`. As far as we know, there is no other similar work on `pt`.

This paper is organized as follows. Section 2 gives an overview of some rule induction methods (mainly those on which ReTraTos is based) and Section 3

summarizes the pre-processing tasks performed, so far, on `pt`–`es` parallel corpus. The transfer rule and bilingual dictionary induction processes are described, respectively, in Sections 4 and 5. This paper ends with some conclusions and future work (Section 6).

2 Related research

Figure 1 shows the general architecture of a system that automatically induces transfer rules and then translates sentences using these rules. In this figure, the dotted line indicates that the use of linguistic or computational resources (such as parsers, bilingual dictionaries and taggers) is optional.

A sentence-aligned parallel corpus (a set of translation examples) is given as input of a rule induction module which produces a set of transfer rules used, in turn, by a rule combination module to translate source sentences into target sentences.

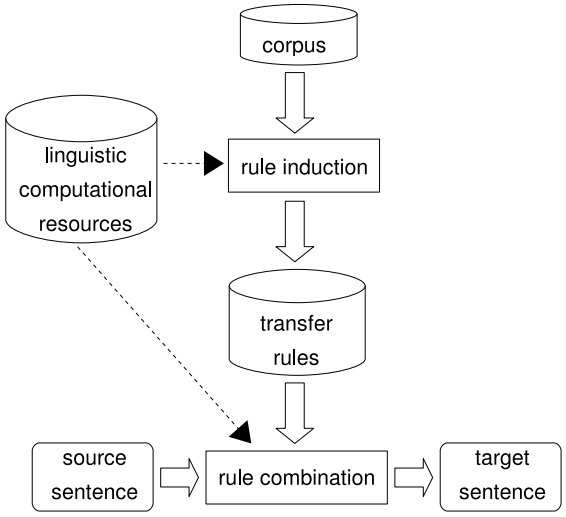

Fig. 1. Architecture of a transfer rule induction system [6]

The method proposed in [6] looks for transfer rules in two steps: monolingual and bilingual. In the monoligual step, the method looks for sequences of items that occur at least in two sentences (patterns), by processing each side (source or target) separately. At the second step, the method builds bilingual patterns following a co-occurrence criterion: one source pattern and one target pattern occurring at the same sentences are taken as translation of each other. Finally, a bilingual similarity (distance) measure is used to help the alignment between source and target items that form a bilingual pattern.

The method proposed in [7] aligns the nodes of source and target parse trees by looking for lexical correspondences in a bilingual dictionary. Then, following a best-first strategy (processing, first, the nodes with the best lexical correspondences), the method aligns the remaining nodes using a manually created alignment grammar composed of 18 bilingual compositional rules. The purpose of this grammar is to ensure that only linguistically significant alignments will be generated. After finding alignments between nodes of both parse trees, these alignments are expanded using linguistic constructors (such as noun and verb phrases) as context boundaries.

In [4], the method infers hierarchical syntactic transfer rules based, initialy, on the constituents of both (manually) lexically aligned languages. To do so, sentences from the language with more resources (English, in that case) are parsed and desambiguated. Value and agreement constraints[1] are built based on syntactic structure, lexical alignments and source/target dictionaries. The composition of these initial rules is performed by looking for small rules found in bigger ones.

The method proposed by ReTraTos (see Section 4) is based on several ideas of these and other rule induction methods, but it is important to strengthen some features that can not be found on them. First, ReTraTos is meant to be modular and incremental, that is, it is possible to induce rules only for one specific type of alignment or for some pre-defined PoS (Part-of-Speech).

Furthermore, lexical alignments are taken as the start point on ReTraTos' rule induction process (not performed at the end like in [6]) and used to split the translation examples into three types of alignment blocks according to the alignment between their words. By doing this, ReTraTos can deal with each type of translation event (alignment preserving word order, reordering and omissions) separately (see Section 4).

Finally, all corpus pre-processing tasks (see Section 3) are performed in a completely automatic way and not manually (as the lexical alignment in [4]) or semi-automatically (as in [7]). By doing this, we decrease the need of human specialists on induction process and, although this approach is not error-proof, we are interested in studying if it is possible to induce useful transfer rules and bilingual entries for a bilingual dictionary in a completely automatic way.

3 Corpus pre-processing

For now, only `pt`–`es` parallel corpus was pre-processed to be used on ReTraTos, the `pt`–`en` parallel corpus will be pre-processed in a near future. The `pt`–`es` parallel corpus (`CorpusFAPESP`) is composed of articles from the online version of the Brazilian scientific magazine *Pesquisa FAPESP*.[2] Three pre-processing

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

[2] The *Pesquisa FAPESP* magazine is available at `http://revistapesquisa.fapesp.br` with parallel texts written in Brazilian Portuguese (original), English (version)

tasks were performed with this corpus: sentence alignment, PoS tagging and word alignment.

The 645 parallel articles of `CorpusFAPESP` were automatically sentence-aligned by a version of Translation Corpus Aligner (`TCAalign`) [3]. The automatic process was followed by a manual analysis of alignments different from 1 : 1 (the most common type). The resulting set of 18,209 pairs of parallel sentences (translation examples) was considered as reference to evaluate the automatic process resulting in 94% of precision and 93% of recall.[3]

The set of translation examples was also PoS tagged using tools available in the open-source shallow-transfer machine translation engine `Apertium`[4] and morphological information provided by Spanish and Brazilian groups where this project has being developed. The same tagset was used for tagging `pt` and `es` parallel texts but it is not a requisite of the induction method.

Finally, the translation examples were word-aligned using the `LIHLA` lexical aligner [2] with 90% of precision and 93% of recall evaluated on a small set of 20 parallel texts randomly chosen from the whole set of 645 parallel texts.[5] In contrast to sentence alignment, the output of lexical aligner was not manually verified once it would not be a feasible task.

After corpus pre-processing, it was obtained a set of 18,209 PoS tagged and lexically aligned `pt`–`es` translation examples with 960,690 words (470,489 in `pt` and 490,201 in `es`). Although this set is not large enough yet, it can give good insights about the rule induction procedure and, in the future, be enlarged for better results.

A pre-processed `pt`–`es` translation example from `CorpusFAPESP` is shown in Table 1 in which the `pt` sentence is at the first line and the `es` sentence, at the second one. PoS tags are presented as a sequence <`X`> where `X` can be the grammatical category (`pr` stands for preposition, `det` for determiner and so on) or inflectional information (`def` stands for definite, `f` for feminine, `pl` for plural and go on) of each word. Alignment information is at the end of each item indicated by one or more positions of the correspondent items on the parallel sentence.

In this table, each surface form (the item as it appears in the text, e.g. *Nas*) is followed by tagger's output (its base forms and PoS tags, e.g. *Em*<`pr`>+o<`det`>) and the alignment produced by the lexical aligner (the positions of correspondent items on the parallel sentence, e.g. 1_2).

---

and Spanish (version).

[3] Precision is the number of correct alignments divided by the total amount of alignments produced by the sentence aligner and recall stands for the number of correct alignments divided by the total amount of alignments found in the reference corpus.

[4] The `Apertium` package, together with its documentation, is available through `http://www.apertium.org`.

[5] Informations about sentence (`TCAalign`) and word (`LIHLA`) aligners used by ReTraTos, as well as their source codes, can be obtained at `http://www.nilc.icmc.usp.br/nilc/projects/aligners.htm`.

<s snum=20>Nas/Em<pr>+o<det><def><f><pl>:1_2 demais/demais<adj><mf><sp>:3 situações/situação<n><f><pl>:4 ,/,<cm>:5 se/se<cnjadv>:6 não/não<adv>:7 aponta/apontar<vblex><pri><3><sg>:8 a/o<det><def><f><sg>:9 causa/causa<n><f><sg>:10 específica/específico<adj><f><sg>:11 ,/,<cm>:12 o/o<det><def><m><sg>:13 *TECR/TECR:14 serve/servir<vblex><pri><3><sg>:15_16_17 como/como<rel><adv>:18 bússola/bússola ... </s>
<s snum=20>En/En<pr>:1 las/el<det><def><f><pl>:1 restantes/restante<adj><mf><pl>:2 situaciones/situación<n><f><pl>:3 ,/,<cm>:4 si_bien/si_bien<cnjadv>:5 no/no<adv>:6 señala/señalar<vblex><pri><3><sg>:7 la/el<det><def><f><sg>:8 causa/causa<n><f><sg>:9 específica/específico<adj><f><sg>:10 ,/,<cm>:11 el/el<det><def><m><sg>:12 *TECR/TECR:13 hace/hacer<vblex><pri><3><sg>:14 las/el<det><def><f><pl>:14 veces/ vez<n><f><pl>:14 de/de<pr>:15 brújula/brújula<n><f><sg>:16 ... </s>

Table 1. A pre-processed pt–es translation example from CorpusFAPESP

4 Inducing transfer rules in ReTraTos project

As mentioned before, the translation examples are split into three types of blocks according to alignment between their words: omission (type 0), alignment preserving word order (type 1) and reordering (type 2). By doing this it is possible to induce rules for each type separately.

Figure 2 shows examples of these three types of alignment blocks where source and target items are accompanied by their positions on source and target sentences.

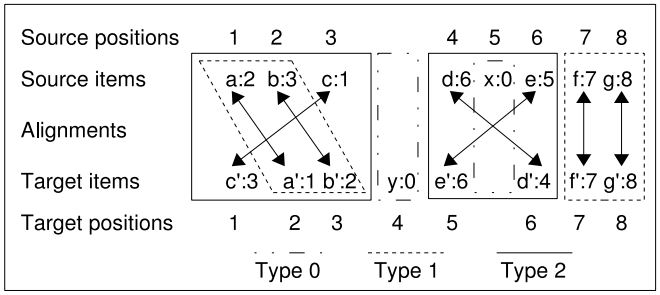

Fig. 2. Types of alignment blocks

In Figure 2, for example, the source items *a* and *b* are aligned to *a'* and *b'* in a way that preserves item order, thus, they form an alignment block of type 1. Furthermore, they are also part of an alignment block of type 2 once the source

item c has a cross-link to c'.[6]

After building these blocks, the rules are induced from each type separately following the three phases explained in the next sections: pattern identification (4.1), rule generation (4.2) and rule filtering and sorting (4.3).

4.1 Pattern identification

Similarly to [6], the pattern identification phase was performed in two steps: monolingual and bilingual. At the monolingual step, source patterns are identified by an algorithm developed based on Sequential Pattern Mining (SPM) technique and the `PrefixSpan` algorithm [9].

According to [9], SPM identifies as patterns the sequences of items that occurs at least a minimal number of times (ϵ).[7] Once identified, patterns are taken as prefixes of other possible frequent sequences and the search goes on looking for new patterns with these prefixes.

In ReTraTos, during monolingual step, source patterns are found following the SPM technique but, different from `PrefixSpan`, a pattern can occur more than once in the same sequence. This difference is very important since the rules are induced from sequences of PoS tags and lexicalized items which are expected to occur several times on the same translation example.

At the bilingual step, for each source pattern, the target items aligned to source ones are looked for to form the bilingual pattern; only bilingual patterns of the type being processed are accepted. This filter has to be applied since, for example, bilingual patterns different from type 2 (reordering) can be induced from alignment blocks of type 2 (see Figure 2).

Examples of bilingual patterns induced from alignment blocks of types 0, 1 and 2 (Figure 2) would be, respectively: $dx \rightarrow d'$ (`n pr`→`n` such as *alinhamento de*→*alignment*), $ab \rightarrow a'b'$ (`det n`→`det n` such as *o exemplo*→*the example*) and $dxe \rightarrow e'd'$ (`n pr n`→`n n` such as *alinhamento de palavras*→*word alignment*).[8]

4.2 Rule generation

The rule generation phase is also performed in two steps: (1) building of constraints between feature values in one (monolingual) or both (bilingual) sides of a bilingual pattern and (2) generalization of these constraints.

Based on [1], at the first step, two kinds of constraints can be built: value constraint and agreement/value constraint. A value constraint specifies which values are expected for the features in each side of a bilingual pattern. An agreement/value constraint, in turn, indicates which items on one or both sides have the same feature values (agreement constraint) and which are these values (value constraint).

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

[7] Frequency threshold is an input parameter chosen empirically as 0.15% of the total amount of alignment blocks of the type being processed.

[8] In these examples of bilingual patterns, `pr` stands for preposition, `n` for noun and `det` for determiner.

Constraints are derived from feature values (inflectional information) in translation examples and the items that them constrain are represented by source (Xi_j) or target (Yi_j) variables, where i and j ($i, j > 0$) indicate, respectively, the item and the feature position at a bilingual pattern.

For example, considering the bilingual pattern identified previously `det n`→ `det n` and the following set of source feature values in one of the several translation examples in which this pattern occurs: {<def><m><sg>, <m><sg>}.[9] A value constraint is built for the first value (`def`) indicated as `X1_1 = def` and two agreement/value constraints are built for the other two feature values (`m` and `sg`): `X1_2 = X2_1 = m` and `X1_3 = X2_2 = sg` indicating that `det` and `n` have the same gender and number.

A similar approach builds target and bilingual constraints. After building the bilingual constraints, all monolingual constraints also specified on bilingual ones are excluded to avoid redundancy. For example, considering that the target side of the bilingual pattern shown above has the same feature values than the source side, so, only the bilingual constraints shown on Set1 are needed.

Set1: {(`X1_1=Y1_1=def`), (`X1_2,X2_1=Y1_2,Y2_1=m`), (`X1_3,X2_2=Y1_3,Y2_2=sg`)}

At the second step, for each set of constraints, ReTraTos looks for another that differs in just one value. If this set is found, the different values are merged (according to a crescent alphabetical order and with the character '|' between them) and the new generalized constraint set replaces the first two. For example, considering the constraint sets Set1 and Set2 which differ in just one value (`sg` and `pl`); these sets are replaced by the set of generalized constraints Set3.

Set2: {(`X1_1=Y1_1=def`), (`X1_2,X2_1=Y1_2,Y2_1=m`), (`X1_3,X2_2=Y1_3,Y2_2=pl`)}

Set3: {(`X1_1=Y1_1=def`), (`X1_2,X2_1=Y1_2,Y2_1=m`), (`X1_3,X2_2=Y1_3,Y2_2=pl|sg`)}

After this phase, bilingual patterns with constraints are considered as transfer rules.

4.3 Rule filtering or sorting

This last phase involves filtering or sorting of transfer rules. Filtering has two purposes: (1) to minimize translation grammar length —what, according to [7], speeds translation process— and (2) to solve ambiguities. Sorting, by its turn, aims at specifying the order in which transfer rules should be applied.

At ReTraTos, the minimization of grammar length is reached by means of a minimal occurrence frequency threshold (ϵ) used in the monolingual step of

[9] In this example, <def><m><sg> are the feature values of the first item (`det`) and <m><sg> are the feature values of the second one (`n`) in the source side of the bilingual pattern `det n`→`det n`. Furthermore, `def` stands for definite (type); `m` and `f` stands, respectively, for masculine and feminine (gender); and `sg` and `pl` stands, respectively, for singular and plural (number).

pattern identification phase (see section 4.1). The ambiguity resolution, on the other hand, is still under study and the followed approach, so far, is to look for feature and lexical values which can distinguish two transfer rules with the same source side but different target ones.

The sorting of transfer rules is done implicitly by setting the frequency and weight of each one. The frequency of a rule is given by the number of times it occurs in translation examples while its weight stands for the probability of its occurrence, that is, its frequency divided by the amount of rules with the same source side (non-ambiguous rules, such that on Figure 3, have weight equal to 1). Frequency and weight information can help the translation system to choose the best rule to be applied.

Figure 3 shows a transfer rule induced by ReTraTos from an alignment block of type 1. Source and target constraints are not specified since bilingual constraints already include monolingual ones. The transfer rules' format used in ReTraTos is largely based on that by [1, 4].

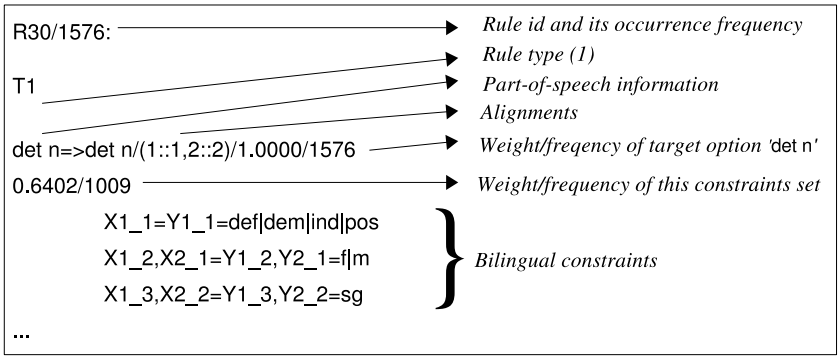

Fig. 3. Example of a transfer rule induced by ReTraTos

5 Inducing bilingual dictionary in ReTraTos project

In addition to transfer rules, a bilingual dictionary is also induced based on the lexical alignments found in translation examples. This bilingual dictionary follows the same XML[10] DTD (document type definition) used by `Apertium`.

The bilingual dictionary is induced following a five-step process: (1) creation of one bilingual dictionary to each translation direction (one source→target and another target→source); (2) merge of the two dictionaries created previously (indicating the valid translation direction, if necessary); (3) generalization of dictionary entries (an approach similar to that applied on constraint generalization presented on Section 4.2); (4) treatment of syntactic differences when

[10] `http://www.w3.org/XML/`

the value of gender or number have to be determined[11] and (5) treatment of multiword units.

An example of a bilingual entry is shown below. The `pt` noun *abordagem* (approach) has as its best translation found in translation examples the `es` noun *abordaje*.

```
<e>
 <p>
  <l>abordagem<s n="n"/><s n="f"/></l>
  <r>abordaje<s n="n"/><s n="m"/></r>
 </p>
</e>
```

In this entry, the left and right sides are tagged `<l>` and `<r>`, respectively, while `<s>` elements (with the attribute `n`) represent PoS tags (noun, `n`) and inflectional information (feminine, `f` and masculine, `m`).

An example of a multiword bilingual entry induced by ReTraTos is shown below. In this example, the best translation found for the `pt` verb *encontrar* (to find) was the `es` multiword verb *deparar con*.

```
<e>
 <p>
  <l>encontrar<s n="vblex"/></l>
  <r>deparar<g><b/>con<b/><g/><s n="vblex"/></r>
 </p>
</e>
```

Aiming at reducing the influence of wrong alignments produced by the lexical aligner (mainly on multiword units), a frequency threshold was applied to constrain the amount of induced multiword entries to that occurring at least 50 times.

6 Conclusions and future work

This paper has briefly described the transfer rule and bilingual dictionary induction procedures developed at ReTraTos project. Some experiments have already been carried out resulting in several entries for the bilingual dictionary and some rules induced for specific alignment types and PoS tags. Although these experiments have given good insights about what can be induced, new experiments are being carefully designed and will be carried out in a near future.

So, the next steps of ReTraTos project are: (1) to induce transfer rules for specific alignment block types and PoS tags separately, (2) to induce new bilingual entries for the bilingual dictionary and (3) to evaluate the induction procedures

[11] It happens when, in one language, a word varies in gender/number (`f`, `m` or `sg`, `pl`) but its translation has a value valid for both gender/number (`mf` or `sp`) like the `es` word *análises* wich is singular-plural (`sp`) and its possible translations in `pt`: the singular (`sg`) word *análise* and the plural (`pl`) word *análises*. In this case, when translation is performed from the more general entry to the less general one the value of gender/number will have to be dertemined based on the neighbouring items.

by using induced transfer rules and bilingual entries for translating between pt and es.

We intend to reproduce the induction process of ReTraTos for pt–en parallel corpus still this year once the project is expected to be completed on the beginning of 2007. Since the procedures were designed to be language-neutral we think pt–en parallel corpus pre-processing is the task that will take more time and, thus, have already been initiated.

The expected results of ReTraTos project are (1) sets of transfer rules and (2) bilingual dictionaries for two language pairs: pt–es and pt–en.

Acknowledgements

We thank FAPESP, CAPES and CNPq for financial support.

References

1. Carbonell, J., Probst, K., Peterson, E., Monson, C., Lavie, A., Brown, R., Levin, L.: Automatic Rule Learning for Resource-Limited MT In Proceedings of the 5th Conference of the Association for Machine Translation in the Americas (AMTA 2002) (2002) 1–10
2. Caseli, H. M., Nunes, M. G. V., Forcada, M. L.: Evaluating the LIHLA lexical aligner on Spanish, Brazilian Portuguese and Basque parallel texts In Proceedings of the XXI Congreso de la Sociedad Espaola para el Procesamiento del Lenguaje Natural (SEPLN) (2005) 1–8
3. 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
4. Lavie, A., Probst, K., Peterson, E., Vogel, S., Levin, L., Font-Llitjos, A., Carbonell, J.: A Trainable Transfer-based Machine Translation Approach for Languages with Limited Resources In Proceedings of the 9th Workshop of the European Association for Machine Translation (EAMT-04) (2004)
5. Lavrac, N., Flach, P., Zupan, B. Rule Evaluation Measures: A Unifying View In Dzeroski, S., Flach, P. (eds) Lecture Notes in Artificial Inteligence **1634** (1999) 174–185
6. McTait, K.: Translation patterns, linguistic knowledge and complexity in an approach to EBMT In Carl, M., Way, A. (eds) Recent Advances in Example-Based Machine Translation (2003) 1–28
7. Menezes, A., Richardson, S. D.: 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 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001) (2001) 39–46
8. Paula, M. F., Rezende, S. O.: Implementao do Ambiente para Explorao de Regras RulEE Technical Report (ICMC-USP) (2003)
9. 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)** (2004) 1–17