Mach Translat
DOI 10.1007/s10590-007-9027-9

Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation

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

Received: 28 May 2007 / Accepted: 14 November 2007

Abstract The availability of machine-readable bilingual linguistic resources is crucial not only for rule-based machine translation 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, and, as a consequence, bilingual resources are usually more difficult to find than "shallow" monolingual resources such as morphological dictionaries or part-of-speech taggers, especially when they involve a less-resourced language. This paper describes a methodology to build automatically both bilingual dictionaries and shallow-transfer rules by extracting knowledge from word-aligned parallel corpora processed with shallow monolingual resources (morphological analysers, and part-of-speech taggers). We present experiments for Brazilian Portuguese–Spanish and Brazilian Portuguese–English parallel texts. The results show that the proposed methodology can enable the rapid creation of valuable computational resources (bilingual dictionaries and shallow-transfer rules) for machine translation and other natural language processing tasks).

Keywords Machine translation · Automatic induction · Transfer rule · Bilingual dictionary · Shallow transfer

H. M. Caseli (✉) · M. G. V. Nunes
NILC – ICMC, University of São Paulo, São Carlos, SP, Brazil
e-mail: helename@icmc.usp.br

M. G. V. Nunes
e-mail: gracan@icmc.usp.br

M. L. Forcada
Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, 03071 Alacant, Spain
e-mail: mlf@ua.es

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 (MT) 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, especially when they involve a less-resourced language.

In an attempt to overcome these challenges, several methods have been proposed to build automatically a variety of linguistic resources such as translation grammars (Kaji et al. 1992; Menezes and Richardson 2001; McTait 2003; Probst 2005) and bilingual dictionaries (Wu and Xia 1994; Fung 1995; Langlais et al. 2001; Koehn and Knight 2002; 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 induced bilingual dictionary is more than just a list of source and target word equivalences. It is a set of bilingual word and multiword entries enriched with morphological and translation direction information. Such a dictionary is an essential resource for transfer-based machine translation systems.

The induced transfer rules, in turn, associate target sequences of part-of-speech (PoS) tags to source sequences of PoS tags. These sequences can come with morphological information. Furthermore, the induced rules can contain restrictions that regulate their application. These rules are induced from blocks of pairs (source language, target language) of contiguous word-aligned items (which will be called *alignment blocks*).

The automatic building of these resources is the goal of a project called ReTraTos.[1] The ReTraTos project aims at inducing linguistic knowledge useful for machine translation—transfer rules and bilingual dictionaries—for Brazilian Portuguese (`pt`) and its translation to other two languages: Spanish (`es`) and English (`en`). To our knowledge, no studies have yet been carried out to build automatically bilingual resources for Brazilian Portuguese.

The proposed methodology uses shallow monolingual resources and parallel corpora to induce bilingual resources. The machine translation experiments carried out for the `pt–es` and `pt–en` language pairs produced reasonable results.

A number of different scenarios can benefit from the methods presented in this paper:

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

– On the one hand, many languages cannot afford bilingual data but can have "basic" or "shallow" monolingual data which would be used to build morphological analysers and part-of-speech taggers. For instance, deeper analysis tools for Portuguese are not publicly available (such as parsers) or even developed (such as wordnets, deeper semantic tools).[2] The existing monolingual data can be used to build the analysis and generation modules of a transfer system (Hutchins and Somers 1992), in particular a shallow-transfer system. The methodology presented here can be applied to generate bilingual linguistic data for such a pair of languages, provided that parallel corpora exist. These data could be used to build the transfer module, with lexical transfer being performed by the bilingual dictionaries induced and structural transfer being performed by the shallow-transfer rules inferred.
– On the other hand, it may be the case that we can use the linguistic data already available for two language pairs, say A–B and C–D. The monolingual part of these data (used in analysis and generation) would be used to generate bilingual data to build a transfer system for a new language pair, say A–D. These bilingual data are built by the method proposed here combining the analysis of A and the generation of D with a parallel corpus for languages A and D.

There is a distinct advantage in the method proposed in this paper, as compared to other learning approaches to machine translation (such as statistical machine translation). Our method 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, the data generated by our method may be easily converted to be used in the `Apertium`[3] (Armentano-Oller et al. 2006) open-source machine translation platform, since the underlying machine translation architecture is very similar to that of `Apertium`. The existing modules and publicly available linguistic data for `Apertium` may also be used to induce data for new language pairs, as is done in this paper for the `pt–en` pair.

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 Sect. 3. The experiments carried out with `pt–es` and `pt–en` language pairs are described in Sect. 4. This paper ends with some conclusions and proposals for future work (Sect. 5).

2 Related work

Several methods have been proposed in the last years aiming at automatically inducing bilingual linguistic resources useful for MT. Among these methods are those proposed to induce bilingual dictionaries and transfer rules.

[2] The Portuguese parser PALAVRAS (Bick 2000) is not publicly available.

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

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

Several automatic word aligners have been proposed (Brown et al. 1993; Och and Ney 2000; Caseli et al. 2005). They use different alignment criteria such as statistics (e.g. co-occurrence frequency) and similarity (e.g. cognate measures).

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% (complete 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 positions. Three judges evaluated 23.8% of the induced entries and the average accuracy was 73.1%.

Other approaches have also been proposed in the literature. Koehn and Knight (2002) propose building a bilingual dictionary from unrelated monolingual corpora. Langlais et al. (2001) build bilingual dictionaries based on simple distributional properties of n-grams and little linguistic knowledge. Schafer and Yarowsky (2002) combine two existing bilingual dictionaries to make a third one using one language as a bridge.

This paper proposes a bilingual dictionary induction method based on alignments produced by an automatic word aligner as will be explained in more detail in Sect. 3.1.

2.2 Induction of translation rules

In the literature, methods for inducing transfer rules are based on many different approaches. 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 to a rule induction module which produces a set of transfer rules. These rules can, in turn, be used by the MT rule application module 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 in at least 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,

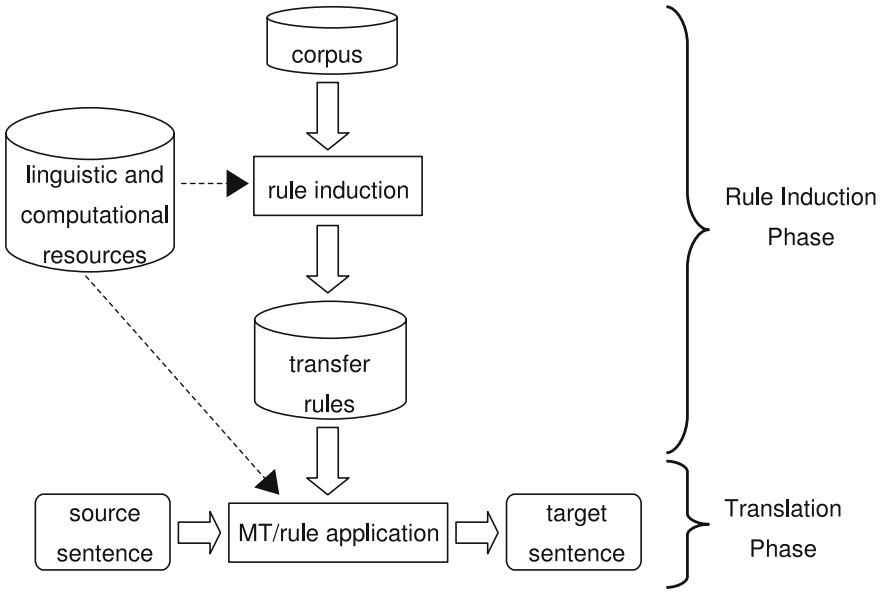

Fig. 1 Architecture of a transfer rule induction system (McTait 2003)

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 Menezes and Richardson (2001) 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 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.

In Carbonell et al. (2002); Lavie et al. (2004), 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[4] are determined from the syntactic structure, the word alignments and the source and target dictionaries.

Sánchez–Martínez and Ney (2006) used 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.

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

[4] 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).

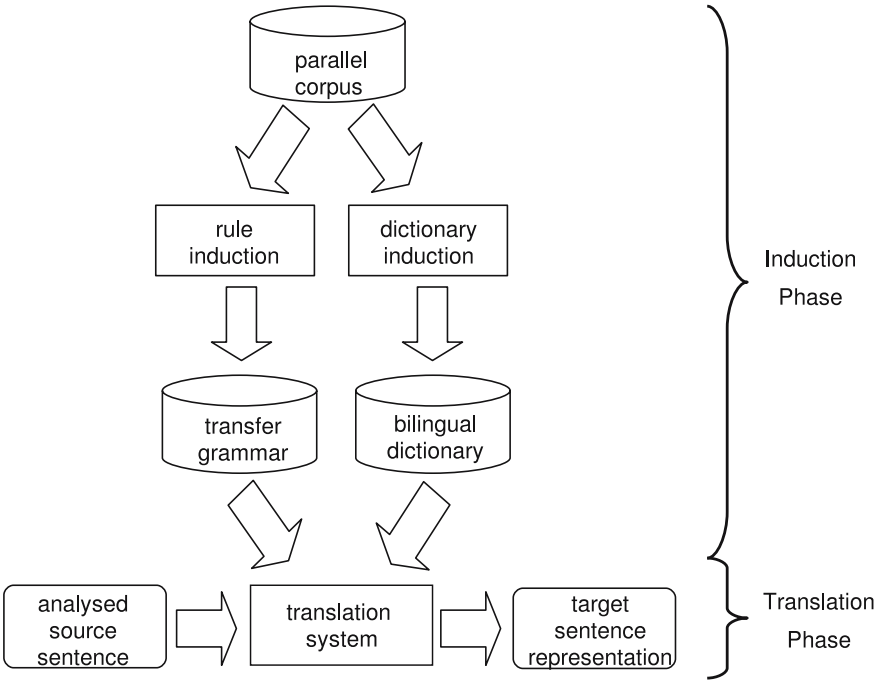

Fig. 2 Scheme of proposed induction and translation systems

3 Induction and translation in the ReTraTos environment

The general scheme of the proposed induction and translation systems is shown in Fig. 2. 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 MT system applies the induced rules that best fit an already analysed source sentence. This system also looks for the best translation for each source word in the bilingual dictionary. It produces as output a representation of a target sentence, that is, sequences of lexical forms (lemma, PoS tags and morphological attributes) before generation in the target language.

The induction systems are introduced in the next two sections.

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 bilingual entries; and (4) the treatment of morphosyntactic differences related to entries in which the value of the target gender or number attribute has to be determined from information that goes beyond the scope of the dictionary entry itself.

In the first step, the method looks for all possible translations in the target (source) sentence for each source (target) word (its lemma, PoS tags and attributes), in each translation example. This search is guided by the word alignments (see Sect. 4.1). If more than one word is found on one or on both sides, the character "+" is used to join together the PoS information of these words to form a multiword unit. At the end of this step, the method stores all possible translations for each source (target) word or multiword unit, associated with their occurrence frequency.

The next step merges the two bilingual dictionaries: (1) by choosing the translation with the highest frequency; (2) by setting the valid translation direction (source–target or target–source), if necessary;[5] and (3) by applying a frequency threshold to constrain the creation of entries containing multiword units. An entry involving more than one word on one or both sides will be created only if it occurs at least n times in the corpus ($n = 50$ in the experiments presented in this paper). This constraint reduces the effect of incorrect multiword unit alignments, since, for this alignment category, the error rate is fairly high (11% in `pt-es` and 16% in `pt-en` parallel corpora) (Caseli 2007).

The third step tries to generalize the attribute values in bilingual entries with the same translation direction by merging the different values. During translation, the best value for the attributes will be determined by the MT system.

Finally, the fourth step deals with entries whose values of gender or number attributes on either side cannot be determined using only information in the entry. This happens when the same word is valid for both values of the gender or number attribute in one language, but corresponds to two different translations in the other language, one for each attribute value. In this step, for each word, the system looks for an entry which has the general value for either gender (`mf`) or number (`sp`) on one side and, on the other side, there is a merged value for either gender (`f|m`) or number (`pl|sg`). If such an entry is found, the system replaces it with three entries according to the translation directions: one for each attribute value and another replacing the merged value with a value representing that gender (`GD`) or number (`ND`).

3.2 Inducing the transfer rules

Based on word alignments, the translation examples are divided into *alignment blocks* (sequences of aligned items). Figure 3 shows the three types of alignment blocks: omission (type 0), alignment preserving item order in sentence (type 1) and reordering (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

[5] A bilingual entry is valid in both translation directions if the correspondence that it represents is the most frequent in both directions (this is considered to be the most general case). When the correspondence is the most frequent in one direction only, this direction has to be explicitly indicated, as this is considered to be a special case.

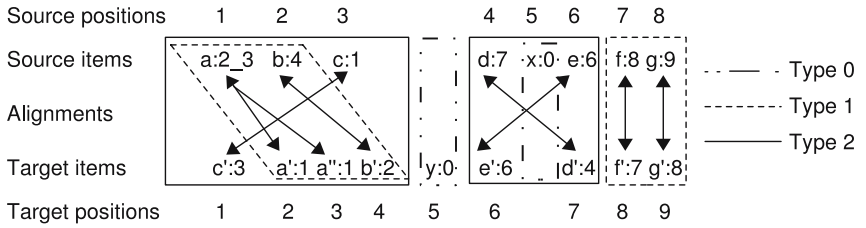

Fig. 3 Types of alignment blocks

to a', a'' and b' in a way that preserves item order, so they form an alignment block of type 1. Furthermore, they are also part of an alignment block of type 2, since the source item c has a cross-link to c'.[6] 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.

Just like alignment templates (Och and Ney 2004), 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 from the information in 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).

After building these alignment blocks, the rules are induced from each type separately, following the four phases explained in the next Sects: (3.2.1) pattern identification, (3.2.2) rule generation, (3.2.3) rule filtering and (3.2.4) ordering.

3.2.1 Pattern identification

Similarly to McTait (2003), the pattern identification phase is performed 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).

According to Pei et al. (2004), SPM identifies as patterns the sequences of items that occur at least a minimal number of times (ϵ). Once they are identified, patterns are taken as prefixes of other possible frequent sequences and the search goes on looking for new patterns with these prefixes.

In our method, however, a pattern is allowed to occur more than once in the same sequence (alignment block). This distinction is very important, since the rules are induced mainly from sequences of PoS tags which can occur several times in the same translation example. The rules can also be induced from lexicalized items, but, in this case, many occurrences of the same lexicalized items are rare.

Since pattern identification is performed for each type of alignment block separately, the frequency threshold (ϵ_t) is different for each type of alignment block

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

t (0, 1 or 2). The value of ϵ_t is calculated as $\epsilon_t = p_{\text{ident}} \times n_t$, where p_{ident} is a percentage (an input parameter) and n_t is the total amount of alignment blocks of type t. For example, suppose that we have $n_1 = 10{,}000$, $n_0 = 1{,}000$ and $n_2 = 100$ common to all block types and the $p_{\text{ident}} = 15\%$; the frequency thresholds would be $\epsilon_1 = 1{,}500$, $\epsilon_0 = 150$ and $\epsilon_2 = 15$, respectively. Using these thresholds, a relevant pattern of type 2 will be identified if it occurs at least 15 times in the alignment blocks of type 2. If we were using the same threshold for all types of alignments, very few relevant patterns coming from less frequent alignment types would be identified.

Other thresholds (also input parameters) are used to limit pattern length. Source patterns longer than a given threshold are discarded and patterns from type 0 are searched for in a window of n items on the left and n items on the right of an alignment block of type 0.

In the bilingual step, the target items aligned to each source pattern are examined (in the parallel translation example) 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 (as can be seen in Fig. 3). The same frequency thresholds used for the monolingual phase (ϵ_t) are applied in the bilingual phase.

For example, a bilingual pattern induced from alignment blocks of type 1 is `det n`→`det n` as in the sequence shown in Fig. 5: *a exploração*→*la explotación*.[7]

3.2.2 Rule generation

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

In the first step, two kinds of constraints can be built—value constraints and agreement/value constraints—as in Carbonell et al. (2002).[8] A value constraint specifies which values are expected for the features on 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).

Constraints are derived from feature values (inflectional information) in translation examples and the items that they constrain are represented by source (Xi_j) or target (Yi_j) variables, where X stands for the source language and Y for the target language. In these variables, $i > 0$ stands for the position of the item in the corresponding (source or target) pattern and $j > 0$ indicates the position of a particular morphological feature in that item.

The format of a value constraint is $Vi\_j = v$ where V can be X (source variable) or Y (target variable) and v is the value set for this variable. The format of agreement/value constraints, on the other hand, can be: $Xi_1\_j_1 = Xi_2\_j_2[, Xi_3\_j_3 \ldots] = v$ (source)

[7] In this bilingual pattern, `det` stands for determiner and `n` for noun.

[8] It is worth mentioning that 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 in Carbonell et al. (2002).

or $Yk_1\_h_1 = Yk_2\_h_2[, Yk_3\_h_3 \ldots] = v$ (target) for monolingual constraints, or $Xi_1\_j_1[, Xi_2\_j_2 \ldots] = Yk_1\_h_1[, Yk_2\_h_2 \ldots] = v$ for bilingual constraints.

For example, consider 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><f><sg>, <f><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 = f` and `X1_3 = X2_2 = sg`, indicating that `det` and `n` have the same gender and number. The complete set of source constraints is shown below (Source_Set # 1).

Source_Set #1 : { (X1_1 = def),
(X1_2 = X2_1 = f),
(X1_3 = X2_2 = sg) }

A similar approach builds target and bilingual constraints. For example, considering that the target side of the previous bilingual pattern has the same feature values as the source side, the resulting target (Target_Set #1) and bilingual (Bilingual_Set #1) constraint sets are as shown below.

Target_Set #1 : { (Y1_1 = def),
(Y1_2 = Y2_1 = f),
(Y1_3 = Y2_2 = sg) }

Bilingual_Set #1 : { (X1_1 = Y1_1 = def),
(X1_2, X2_1 = Y1_2, Y2_1 = f),
(X1_3, X2_2 = Y1_3, Y2_2 = sg) }

After building the bilingual constraints, all monolingual constraints included among the bilingual ones are discarded to avoid redundancy. In the example above, only the bilingual constraints in Bilingual_Set #1 are needed.

In the second step, for each set of constraints, the method looks for another set of constraints that differs in just one value. If this set is found, the different values are merged (in alphabetical order of values and with the character '|' between them) and the new generalized constraint set replaces the first two. For example, the bilingual constraint sets Bilingual_Set #1 and Bilingual_Set #2 differ in just one value (`sg` and `pl`); these sets are replaced by the set of generalized constraints Bilingual_Set #3.

Bilingual_Set #2 : { (X1_1 = Y1_1 = def),
(X1_2, X2_1 = Y1_2, Y2_1 = f),
(X1_3, X2_2 = Y1_3, Y2_2 = pl) }

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

Bilingual_Set #3 : { (X1_1 = Y1_1 = def),
(X1_2, X2_1 = Y1_2, Y2_1 = f),
(X1_3, X2_2 = Y1_3, Y2_2 = pl|sg) }

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

3.2.3 Rule filtering

The filtering of induced rules usually has two purposes: (1) to minimize the length of the translation grammar and (2) to solve ambiguities. In our method, the minimization of grammar length is reached by means of the minimal occurrence frequency threshold (ϵ_t) used in the pattern identification phase (see Sect. 3.2.1).

Ambiguity resolution, on the other hand, is carried out for transfer rules having the same source side (sequence of source PoS tags), but different target sides (ambiguous rules). For these rules, the best target option is determined to be the most frequent one (frequency equal to $\text{freq}_{\text{best}}$). The other less frequent target possibilities will be filtered only if their frequencies are at least $p_{\text{filter}} \times \text{freq}_{\text{best}}$, where p_{filter} is a percentage (also an input parameter). For example, if $p_{\text{filter}} = 50\%$ only the target options with occurrence frequency equal or greater than half of $\text{freq}_{\text{best}}$ will be filtered.

Our filtering approach looks for feature and lexical values which can distinguish ambiguous rules. The *feature-value* filtering reduces the set of target side possibilities by keeping only those that can be distinguished from the best target side by a source or a bilingual constraint. If the feature-value filtering fails, i.e. if it does not find a value that can distinguish two target options of an ambiguous rule, then, the *lexical-value* filtering is applied. This second filtering creates a new transfer rule by adding lexical constraints to the source side of the rule being filtered.

In the current version of our rule induction method we decided to induce only non-ambiguous grammars, so when both filters fail only the best target option is kept and the remaining possibilities are discarded.

3.2.4 Rule ordering

Rule ordering aims at specifying the order in which transfer rules should be applied by the machine translation system. In our method, it is done implicitly by setting the frequency and weight of each rule.

The frequency of a rule is given by the number of times it occurs in the corpus of translation examples used to induce the rule. The weight of a rule stands for the probability of its occurrence, i.e. its frequency divided by the total frequency of the rules. For each rule, the frequency and the weight of each target side and constraint set are computed. Frequency and weight information can help the translation system to decide which is the best rule to be applied at each moment during the translation process.

Figure 4 shows two transfer rules induced from alignment blocks of types 1 and 0, respectively. The first one is a `pt–es` translation rule, while the second one is a `pt–en` translation rule that expresses the reordering of noun–adjective (`n-adj`) sequences. In both rules, source and target sets of constraints are not explicitly specified, since they

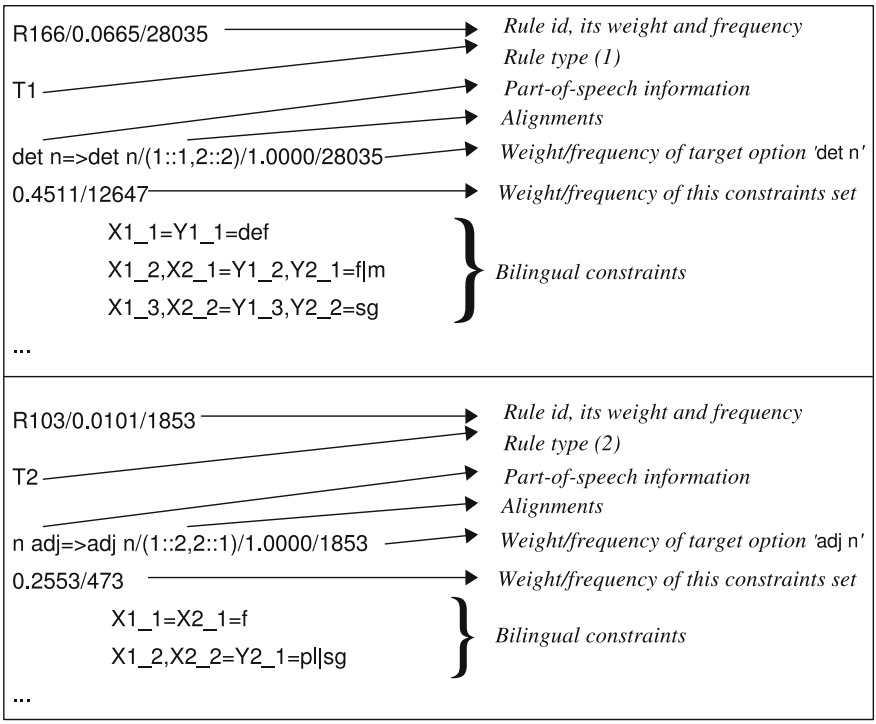

Fig. 4 Examples of transfer rules induced from alignment blocks of type 1 (from `pt` to `es`) and 2 (from `pt` to `en`), respectively

can be found in the bilingual sets. The format of our transfer rules is largely based on that of Carbonell et al. (2002); Lavie et al. (2004). This rule formalism is compatible with that used in `Apertium`, and automatic conversion from one formalism to the other will be available in the near future.

3.3 Translating sentences

The induced resources (dictionaries and rules) are used in the MT task by means of a simple translation system (see Fig. 2). The input of this system is an already analysed source sentence, i.e. a sequence of source lexical forms (lemma, PoS tag, etc.).

The implemented system has two modes of translation: word-by-word and transfer. In the first, the system translates each source word and multiword unit looking for the best translation in the bilingual dictionary. In the second, the system chooses and applies the best suitable transfer rules following a left-to-right longest-match procedure.

More specifically, the "best suitable rule" is the most frequent rule which: (a) matches the source sequence, (b) matches a set of source constraints (there can be more than one) and (c) this source constraint set is the most frequent. Therefore, the selected rule might not be the most frequent.

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 (Sect. 4.1) and the evaluation settings and results (Sect. 4.2).

4.1 Preprocessing of bilingual corpora

A corpus of PoS-tagged and word-aligned parallel sentences (translation examples) is the input of the proposed inducing methods.

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 1,049,462 tokens (503,596 in `pt` and 545,866 in `es`). The other corpus consists of 17,397 pairs of `pt–en` parallel sentences with 1,026,512 tokens (494,391 in `pt` and 532,121 in `en`). Both corpora contain articles from the online version of a Brazilian scientific magazine, Pesquisa FAPESP.[10] It contains parallel texts written in Brazilian Portuguese (original), English (version) and Spanish (version).

These corpora were PoS-tagged using two tools available in `Apertium` (Armentano-Oller et al. 2006): a morphological analyser and a PoS tagger. 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. The PoS tagger chooses the best possible analysis based on a first-order hidden Markov model (HMM).

The morphological dictionaries available in `Apertium` were enlarged in the ReTraTos project. The `pt` and `en` dictionaries were enlarged with entries extracted from `Unitex`[11] (Paumier 2006) dictionaries. The `es` dictionary was enlarged with entries from the linguistic data used in the Spanish–Catalan (`ca`) machine translation system `InterNOSTRUM`[12] (Canals-Marote et al. 2001) provided by the Transducens machine translation group from the Universitat d'Alacant. The morphological dictionaries for `pt` and `es` available 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 `Apertium en–ca` linguistic data package (version 0.8) was enlarged to cover 61,601 surface forms.[13]

[10] *Pesquisa FAPESP* is available at http://revistapesquisa.fapesp.br .

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

[12] http://www.internostrum.com/.

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

pt	<s snum=87>Embora/Embora<cnjadv>:1 o/o<det><def><m><sg>:3 *piquiá/piquiá:4 não/não<adv>:5 esteja/estar<vblex><prs><p3><sg>:6_7 sob/sob<pr>:8 risco/risco<n><m><sg>:9 de/de<pr>:10 ser/ser<vbser><inf>:11 extinto/extinto<adj><m><sg>:11 ,/,<cm>:12 a/o<det><def><f><sg>:13 exploração/exploração<n><f><sg>:14 descontrolada/descontrolado<adj><f><sg>:15 pode/poder<vbmod><pri><p3><sg>:16 levar/levar<vblex><inf>:17 ao/a<pr>+o<det><def><m><sg>:18 desaparecimento/desaparecimento<n><m><sg>:19 dessa/de<pr>+esse<det><dem><f><sg>:0 árvore/árvore<n><f><sg>:0 em/em<pr>:20 algumas/algum<det><ind><f><pl>:21 regiões/região<n><f><pl>:22 ./.<sent>:23 </s>
es	<s snum=87>Pese_a/Pese_a<pr>:1 que/que<cnjsub>:0 el/el<det><def><m><sg>:2 *piquiá/piquiá:3 no/no<adv>:4 se/se<prn><pro><ref><p3><mf><sp>:5 encuentra/encontrar<vblex><pri><p3><sg>:5 bajo/bajo<pr>:6 riesgo/riesgo<n><m><sg>:7 de/de<pr>:8 extinción/extinción<n><f><sg>:9_10 ,/,<cm>:11 la/el<det><def><f><sg>:12 explotación/explotación<n><f><sg>:13 desmesurada/desmesurado<adj><f><sg>:14 puede/poder<vbmod><pri><p3><sg>:15 ocasionar/ocasionar<vblex><inf>:16 su/suyo<det><pos><mf><sg>:17 desaparición/desaparición<n><f><sg>:18 en/en<pr>:21 algunas/alguno<det><ind><f><pl>:22 regiones/región<n><f><pl>:23 ./.<sent>:24 </s>

Fig. 5 A pt–es translation example (a PoS-tagged and word-aligned sentence pair)

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 AER (alignment error rate) 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 2007). 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).

Figure 5 shows a pt–es translation example in which each surface form (the word as it appears in the text, e.g. the underlined pt word *a*) is followed by the output of the tagger (its lemma and PoS tags, e.g. *o*<det><def><f><sg>) and the alignment produced by the word aligner (the position of the corresponding token on the other side, e.g. 13).

Omission alignments are indicated by 0, such as the alignment of the second es token *que* (underlined in the figure), which does not have any correspondence in the pt sentence. Multiword unit alignments are expressed by concatenating (with "–") the positions of corresponding tokens, as in the underlined alignment between the pt word *esteja* (the 5th source token) and two es words: *se* and *encuentra* (the 6th and 7th target tokens). This 1:2 alignment connects a source word with a target multiword unit. Multiword units can also be output by the analyser and then selected by the PoS tagger, such as the first es token *Pese_a*. The morphological analyser is also responsible for marking unknown words with a "*" as in **piquiá*.

4.2 Evaluation settings and results

The linguistic resources induced from the two training parallel corpora described in Sect. 4.1 consist of two bilingual dictionaries and different configurations of transfer rules.

One bilingual dictionary was induced for each language pair: one with 23,450 pt–es entries and another with 19,191 pt–en entries. Some sets of transfer rules were induced considering distinct values for the identification (p_{ident}) and filtering (p_{filter}) percentages (input parameters). The best of all tested configurations used $p_{\text{ident}} = 0.07\%$ and $p_{\text{filter}} = 0.50\%$. With this configuration, 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 length threshold used in the source pattern identification step limited the rules to five source items at most.

The corpus used for testing 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). For the evaluation of the translations, a reference corpus was created, consisting of the corresponding parallel sentences in the test corpus. For example, the reference corpus used for evaluating the translation from pt to es is composed by the es sentences in the test corpus.

The automatically translated sentences were compared with those in the reference corpus by means of five automatic MT evaluation metrics: BLEU (Papineni et al. 2002), NIST (Doddington 2002) and the well-known precision (P), recall (R) and F-measure (F) (Melamed et al. 2003).

The first two measures compute, in different ways, the n-grams which are in the automatically translated sentence and also in the reference sentence. These two measures estimate the similarity in terms of length, word choice and order.

For the other three measures, the word choice, but not the word order, is considered. That is, the precision and recall of an automatically translated sentence is calculated as the number of tokens which are in this sentence and also in the reference sentence, divided by the number of proposed tokens (precision) and reference tokens (recall).

In these experiments, we evaluated the sentences translated by the ReTraTos MT system (see Sect. 3.3) by applying the induced resources in the word-by-word translation (ReTraTos_word-by-word) and in the transfer translation

Table 1 Evaluation of pt–es–pt MT

Lang.	System	BLEU	NIST	P	R	F
pt–es	ReTraTos_transfer	0.6513	10.8516	0.7991	0.7944	0.7968
	ReTraTos_word-by-word	0.6490	10.8188	0.7971	0.7932	0.7952
	Apertium	0.6382	10.6379	0.8080	0.7964	0.8021
	Apertium-P	0.6387	10.6438	0.8082	0.7966	0.8024
es–pt	ReTraTos_transfer	0.6666	10.9756	0.8003	0.8068	0.8035
	ReTraTos_word-by-word	0.6649	10.9503	0.7991	0.8074	0.8033
	Apertium	0.6098	10.3057	0.7714	0.7853	0.7783
	Apertium-P	0.6288	10.5073	0.7841	0.7969	0.7904

(ReTraTos_transfer). The former uses only the induced bilingual dictionary, while the latter uses both the induced dictionary and the set of induced transfer rules. 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 about by using transfer rules (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.[14] For pt–en–pt, we used the MT systems FreeTranslation,[15] BabelFish[16] and Google.[17]

Table 1 shows the results of pt–es–pt translation. From these values, it is noteworthy that the ReTraTos MT system using only one (ReTraTos_word-by-word) or both (ReTraTos_transfer) of the induced linguistic resources performed a little better than Apertium's versions, with a more significant difference in the es–pt direction.

In the pt–es direction, when compared to Apertium-P, ReTraTos_transfer had an improvement of around 2% in BLEU and NIST; while in the es–pt direction, this improvement was 6% in BLEU and 4% in NIST. Although for the pt–es direction Apertium's versions had slightly higher values for precision, recall and *F*-measure, this fact indicates only that they made a better word choice than ReTraTos but do not say anything about word order, for instance.

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

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

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

[16] http://babelfish.altavista.com.

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

Table 2 Evaluation of pt–en–pt MT

Lang.	System	BLEU	NIST	P	R	F
pt–en	ReTraTos_transfer	0.2832	7.0869	0.6132	0.5986	0.6058
	ReTraTos_word-by-word	0.2606	6.7712	0.5964	0.5885	0.5924
	FreeTranslation	0.3294	7.6509	0.6670	0.6586	0.6628
	BabelFish	0.3161	7.4648	0.6517	0.6438	0.6477
	Google	0.3295	7.6112	0.6609	0.6470	0.6539
en–pt	ReTraTos_transfer	0.2400	6.1133	0.4707	0.4942	0.4822
	ReTraTos_word-by-word	0.2324	6.0173	0.4640	0.4973	0.4800
	FreeTranslation	0.3053	6.8454	0.5367	0.5846	0.5596
	BabelFish	0.3666	7.6799	0.6064	0.6419	0.6237
	Google	0.3121	6.8767	0.5379	0.5805	0.5584

such as pt and es, a greater coverage of the bilingual dictionary has a stronger impact in translation than the transfer rules.

Table 2 shows the results of pt–en–pt translation. In the evaluation for this 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, and it may also be attributed to the 5-word limit used in source pattern identification. These changes are 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 (ReTraTos_transfer) compared to the word-by-word (ReTraTos_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, although simple (in the sense that they perform only shallow changes), the induced rules can significantly improve word-by-word translation between more distant languages.

5 Conclusions and future work

We have described a methodology to build bilingual dictionaries and translation rules automatically. These linguistic resources are built by extracting knowledge from parallel corpora processed using word aligners and shallow monolingual resources such as morphological analysers and part-of-speech taggers.

We describe experiments for the Brazilian Portuguese–Spanish and Brazilian Portuguese–English language pairs. The bilingual resources inferred and the monolingual resources used to infer them are combined to build a promising first version of a shallow-transfer machine translation system.

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

of very intuitive linguistic concepts (parts of speech, local agreement, etc.), which are accessible even to non-experts.

In particular, the rules can be easily converted to the formats used by the open-source machine translation platform `Apertium` (Armentano-Oller et al. 2006), and the bilingual dictionary entries are already induced in the formalism used by `Apertium`. Thus, new machine translation systems can easily be built by combining the induced transfer rules and the bilingual dictionary entries with the modules and linguistic data distributed in `Apertium`.

This approach may be very useful for developing translation systems for less-resourced languages, for which machine-readable bilingual resources do not exist at all, but may have "shallow" or "cheap" monolingual resources, and for which experts are not so readily available.

The methodology presented here may be used as part of a *tool chain* to build shallow-transfer machine translation systems including: aligners such as LIHLA (Caseli et al. 2005) or GIZA++ (Och and Ney 2000), morphological analysers such as those found in Unitex (Paumier 2006) or Apertium (Armentano-Oller et al. 2006), and the dictionary and rule induction methods developed here. Such a tool chain would be very useful when tackling the "transfer problem", i.e. the need to generate bilingual data for every possible translation pair, when interlingua-based systems are not available or feasible, which is often the case.

As future work, we intend to implement this idea of a tool chain for machine translation using the already existing free resources from `Apertium` and from ReTraTos. Other future work includes the evaluation of different configurations of ReTraTos to determine to what extent each of the optional modules (rule filtering and rule ordering) contributes to translation quality.

Acknowledgements We thank the financial support of the Brazilian agencies FAPESP (02/13207-8 and 04/06707-0), CAPES (053804-3) 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). We also thank Mônica Saddy Martins, Élen Tomazela, Gema Ramírez-Sánchez, Carmen Dayrell and the Transducens group at the Universitat d'Alacant for their contributions to this work.

References

Armentano-Oller C, Carrasco RC, Corbí-Bellot AM, Forcada ML, Ginestí-Rosell M, Ortiz-Rojas S, Pérez-Ortiz JA, Ramírez-Sánchez G, Sánchez–Martínez F, Scalco MA (2006) Open-source Portuguese–Spanish machine translation. In: Proceedings of the VII Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada. Itatiaia, RJ, Brazil, pp 50–59

Bick E (2000) The parsing system Palavras, automatic grammatical analysis of Portuguese in a constraint grammar framework. Ph.D. Thesis, Aarhus University Press, Denmark

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

Canals-Marote R, Esteve-Guillén A, Garrido-Alenda A, Guardiola-Savall M, Iturraspe-Bellver A, Montserrat-Buendia S, Ortiz-Rojas S, Pastor-Pina H, Pérez-Antón P, Forcada M (2001) The Spanish–Catalan machine translation system interNOSTRUM. In: MT Summit VIII: Machine Translation in the Information Age, Proceedings Santiago de Compostela, Spain, pp 73–76

Carbonell J, Probst K, Peterson E, Monson C, Lavie A, Brown R, Levin L (2002) Automatic rule learning for resource-limited MT. In: AMTA'02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas: From Research to Real Users. Lecture Notes In Computer Science, vol 2499, London, UK, pp 1–10

Caseli HM (2007) Indução de léxicos bilíngües e regras para a tradução automática. Ph.D. Thesis, ICMC-USP, São Paulo, Brazil

Caseli HM, Nunes MGV (2007) Automatic induction of bilingual lexicons for machine translation. Int J Transl 19: 29–43

Caseli HM, Nunes MGV, Forcada ML (2005) Evaluating the LIHLA lexical aligner on Spanish, Brazilian Portuguese and Basque parallel texts. Procesamiento del Lenguaje Natural 35:237–244

Doddington G (2002) Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of ARPA Workshop on Human Language Technology, San Diego, CA, pp 128–132

Fung P (1995) A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, MA, pp 236–243

Hutchins J, Somers H (1992) An introduction to machine translation. Academic Press, London

Kaji H, Kida Y, Morimoto Y (1992) Learning translation templates from bilingual text. In: Proceedings of the fifteenth [sic] International Conference on Computational Linguistics, COLING-92. Nantes, France, pp 672–678

Koehn P, Knight K (2002) Learning a translation lexicon from monolingual corpora. In: Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX), Philadelphia, PA, pp 9–16

Langlais P, Foster G, Lapalme G (2001) Integrating bilingual lexicons in a probabilistic translation assistant. In: MT Summit VIII: Machine Translation in the Information Age, Proceedings, Santiago de Compostela, Spain, pp 197–202

Lavie A, Probst K, Peterson E, Vogel S, Levin L, Font-Llitjós A, Carbonell J (2004) 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), Valletta, Malta, pp 1–8

McTait K (2003) Translation patterns, linguistic knowledge and complexity in an approach to EBMT. In: Carl M, Way A (eds) Recent advances in example-based machine translation. Kluwer Academic Publishers, Dordrecht, The Netherlands pp 307–338

Melamed ID, Green R, Turian JP (2003) Precision and recall of machine translation. In: Proceedings of the Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT/NAACL 2003), Edmonton, Canada, pp 61–63

Menezes A, Richardson SD (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 39th Annual Meeting of the ACL and 10th Meeting of the European Chapter, Toulouse, France, pp 39–46

Och FJ, Ney H (2000) Improved statistical alignment models. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, China, pp 440–447

Och FJ, Ney H (2003) A systematic comparison of various statistical alignment models. Comput Linguist 29(1):19–51

Och FJ, Ney H (2004) The alignment template approach to statistical machine translation. Comput Linguist 30(4):417–449

Papineni K, Roukos S, Ward T, Zhu W (2002) BLEU: a method for automatic evaluation of machine translation. In: ACL-02: the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, pp 311–318

Paumier S (2006) Unitex 1.2 user manual. Université Paris-Est, Marne-la-Vallée, France

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

Probst K (2005) Learning transfer rules for machine translation with limited data. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA

Sánchez–Martínez F, Ney H (2006) Using alignment templates to infer shallow-transfer machine translation rules. In: Pyysala S, Salakoski T, Ginter D, Pahikkala T (eds) Advances in natural language processing, Proceedings of 5th International Conference on Natural Language Processing FinTAL, vol. 4139 of Lecture Notes in Computer Science, Turku, Finland, pp 756–767

Schafer C, Yarowsky D (2002) Inducing translation lexicons via diverse similarity measures and bridge languages. In: Proceedings of CoNLL-2002, Taipei, Taiwan, pp 1–7

Wu D, Xia X (1994) Learning an English–Chinese lexicon from parallel corpus. In: Proceedings of the 1st Conference of the Association for Machine Translation in the Americas (AMTA-1994), Columbia, MD pp 206–213