Bootstrapping a Hybrid MT System to a New Language Pair

João Rodrigues, Nuno Rendeiro, Andreia Querido, Sanja Štajner, António Branco
Department of Informatics, Faculty of Sciences
University of Lisbon, Portugal
{joao.rodrigues, nuno.rendeiro, andreia.querido, stajner.sanja, antonio.branco}@di.fc.ul.pt

Abstract

The usual concern when opting for a rule-based or a hybrid machine translation (MT) system is how much effort is required to adapt the system to a different language pair or a new domain. In this paper, we describe a way of adapting an existing hybrid MT system to a new language pair, and show that such a system can outperform a standard phrase-based statistical machine translation system with an average of 10 persons/month of work. This is specifically important in the case of domain-specific MT for which there is not enough parallel data for training a statistical machine translation system.

Keywords: Hybrid Machine Translation, Tecto MT, Portuguese Synthesis

1. Introduction

Phrase-based statistical machine translation (PBSMT) is considered as state-of-the-art MT approach whenever sufficiently large parallel (or comparable) datasets for training are available. However, for many language pairs and translation directions (English to Portuguese among them) large training datasets only exists for few domains, such as parliamentary discussions (Europarl (Koehn, 2005)) or legal documents (JRC-Acquis corpus (Steinberger et al., 2006)). In such cases, it is assumed that a rule-based or a hybrid MT system lead to better results as it can better overcome the data sparsity and generalise over the unseen word forms (especially useful in the case of morphologically rich languages). The main concern is, however, that adaptation of a rule-based or a hybrid MT system (its rules) to a new language pair or a new domain may require considerable time and effort.

In this paper, we focus on a hybrid MT system (TectoMT (Popel and Žabokrtskỳ, 2010)) and show that the adaptation of the existing TectoMT system for English-Czech language pair to a new language pair (English to Portuguese translation) can be done with an average of 10 persons/month of work time and outperform the standard PBSMT system on a domain-specific MT task.

In the next section (Section 2.), we briefly describe the architecture of the TectoMT system and its main components. Section 3. describes how we adapted from the English synthesis to the required Portuguese synthesis. In Section 4. we present the results of the automatic evaluation of our English to Portuguese TectoMT system and discussion. Section 5. contains closing remarks and directions for future work.

2. Related Work

TectoMT is a modular, machine translation system (Popel and Žabokrtskỳ, 2010), based on the Prague tectogrammatics theory of Sgall (Sgall et al., 1986). It uses two layers of structural description, the shallow a-layer (syntactic analysis) and the deep t-layer (Figure 1).

The tree-to-tree transfer (Žabokrtský et al., 2008) is performed on the deep t-layer (tectogramatical layer) by using the Maximum Entropy context-sensitive translation models (Mareček et al., 2010). The analysis phase (which converts the input sentence into the a-layer and then the t-layer) and the synthesis phase (which converts the translated t-layer representation to the a-layer and then to the output surface string) are mostly rule-based. They use a modular structure, allowing for its components to be inherited, and adapted, for different languages.

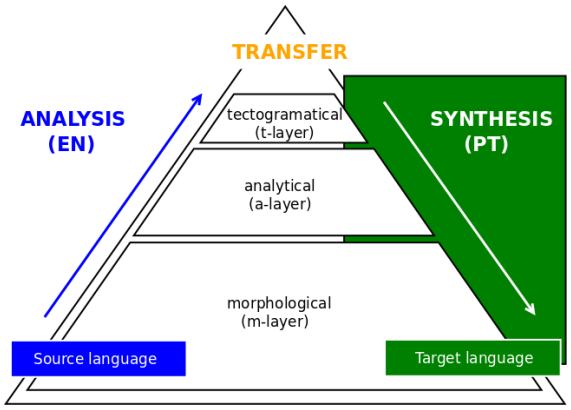

Figure 1: TectoMT architecture.

The English-Czech TectoMT system achieved very good results on the Workshop on Statistical Machine Translation 2013 shared task (Bojar et al., 2013). Those results, together with the widespread concerns that the state-of-the-art PBSMT may be reaching a performance ceiling, make us believe that the use of the linguistically rich MT approaches, such as TectoMT, could be a good way forward towards higher quality MT.

In order to adapt the English-Czech TectoMT to a new language pair, it is only necessary to adapt the synthesis phase of the original system and train a transfer model (handled by the tree-to-tree maximum entropy translation model (Mareček et al., 2010)) which takes around two weeks to train over the whole Europarl corpus.

The blocks for English analysis and synthesis phases can be inherited from the English-Czech system (Popel and

Žabokrtskỳ, 2010) directly.

3. Adaptation of the Synthesis Phase

As a modular framework, TectoMT contains the different components separated into *blocks* (of Perl code) that are triggered at each stage of the processing pipeline.

The synthesis stage starts by converting the result of the transfer stage, represented as a t-tree, into an a-tree. In the last step/block of the synthesis stage, the a-tree nodes are converted to surface forms.

Unlike the a-tree, the t-tree does not contain nodes which represent functional words, such as prepositions or conjunctions. The information about auxiliary words is instead preserved in a form of an attribute to a node, or in the relations between the nodes. The t-tree uses the *lemma* to express the lexical meaning of the node and the *formeme* to help separating the lexical from the syntactic information. For example, the formeme of the semantic noun in a subject position is represented as *n:sub*. For the morphological categories the grammateme tectogrammatical representation is used (Žabokrtský et al., 2008).

For the Portuguese synthesis adaptation, the process begun by adapting the existing English blocks (Popel and Žabokrtskỳ, 2010), with the corresponding Portuguese linguistic phenomena. The next step was to create new blocks for Portuguese-specific phenomena.

We resorted to the existing tools developed by (Branco and Silva, 2006) whenever possible, due to the higher accuracy over the available tools within the original TectoMT system. For verbal conjugation and for nominal inflection, we used the LX-Conjugator and LX-Inflector to generate surface forms (described in more details in Section 3.1.). We created new TectoMT blocks to call and incorporate these tools into the TectoMT pipeline.

The English to Portuguese TectoMT system was being improved iteratively, controlling for the BLEU score and human error analysis of 1,000 sentences in each step. The set of 1,000 sentences (QTLEAP-IT1) was compiled under the QTLeap project by collecting the sentences from a real-usage scenario where a user chats with the Information Technology (IT) support and translating them by professional translators. After each iteration (adding of new synthesis blocks), we checked the BLEU score and performed a human error analysis by two linguists, native speakers of Portuguese. The linguists were analyzing the most frequently missing n-grams (up to three-grams) and the t-trees at the starting point of the synthesis phase, and suggesting rules which would enforce better synthesis (transformation of t-trees to the surface form).

3.1. Portuguese-Specific Synthesis Blocks

In this subsection, we describe the blocks used for the synthesis pipeline for Portuguese (the names of the blocks are presented in bold).

3.1.1. Add Blocks

AddAuxVerbCompoundPassive adds the Portuguese auxiliary verb *ser* to a passive verb and creates a child node before the current node, setting as default the third person in the indicative.

AddConditional creates an auxiliary node for the Portuguese conditional conjunction *se*. These situations are present when the corresponding t-node functor (semantic values of syntactic dependency relations which express the functions of individual modifications in the sentence) is a conditional. The new auxiliary node is placed preceding the reference node.

AddArticles adds articles as child nodes of the noun nodes according to their definiteness grammatemes, and person and gender (information about person and gender is contained in the Interset).

AddAuxVerbModalTense adds an auxiliary expression for combined modality and tense according to the verbal mood (indicative, imperative or conditional), the tense (simultaneous, preceding or subsequent event) and the deontic modality.

AddVerbNegation creates the particle corresponding to the negation of the negated verb.

AddPrepos adds the prepositional nodes. In Portuguese, the adverbs are also included as a possible prepositional formation.

AddSubconjs adds the subordinating conjunctions.

AddCoordPunct handles the commas in the coordinations.

AddComparatives adds the nodes for the comparative degree (*mais*).

AddParentheses adds the parenthesis according to the t-nodes attribute (is_parenthesis).

AddSentFinalPunct adds the end-of-sentence punctuation mark.

3.1.2. Remove Blocks

DropSubjPersProns removes the nodes that identify the personal pronouns (#PersPron).

DropPersPronSb deletes personal pronouns in subject position (in pro-drop languages such as Portuguese).

DeleteSuperfluousAuxCP deletes superfluous nodes which may appear by default in some combinations of prepositions or subordinate conjunctions. For example, it replaces "for X and *for* Y" with "for X and Y".

3.1.3. Reorder Blocks

MoveRhematizers shifts rhematizers before articles and prepositions. Rhematizers are expressions whose function is to signal the topic-focus articulation categories in the sentence.

3.1.4. Repair Blocks

ImposeLemma fixes erroneous lemmas produced during the transfer phase (a few lemmas were repeatedly wrongly selected during the transfer phase and for those cases, the suggested rules impose the choice of the correct lemma).

ImposeFormeme fixes some of the errors in formemes. If the lemma does not correspond to the correct formeme, the block will search for the closest lemma with the correct part-of-speech. In some specific lemmas, for example, this block forces the use of the Portuguese proposition *de*, and

checks in every following child node (in the sentence order) for a formeme which is a verb and adds the preposition.
FixPossessivePronouns fixes the inflection of possessive pronouns, e.g. *seu, teu* and *meu*.

3.1.5. Agreement Blocks

AddGender sets the gender for every noun and adjective a-node. It resorts to the LX-Suite tagger, passes the form, lemma, part-of-speech tags and other attributes to the tagger and returns the best gender.
ImposeSubjpredAgr sets the gender, number and person of verbs according to their subjects. In Portuguese, verbs have no gender but a noun-complement is needed for the gender concordance.
ImposeAttrAgr sets the gender, number and person according to their governing nouns.
SecondPersonPoliteness sets politeness person for Portuguese (third person). This occurs when the lemma is represented as a *PersPron* (all personal and possessive pronouns, including reflexive pronouns) and the corresponding t-node is in the second person.

3.1.6. Inflect Blocks

GenerateWordforms generates the corresponding word forms for each lemma by using the LX-Conjugator for the verb nodes and the LX-Inflector for the adjective and nouns nodes. This block also uses the Interset (Zeman, 2008) part-of-speech, number, mood, tense, person and lemma. This block also handles the inflection for the superlative degree.
GeneratePronouns generates pronouns by using the formeme and the Interset person, gender and number. This block also handles the possessive, dative, accusative and oblique case pronouns.

3.1.7. Other Blocks

CopyTtree performs a deep-copy of the current transfer t-trees into an a-tree.
CliticExceptions handles the clitics in Portuguese.
MarkSubject fills the *Afun* label (analytical functions which correspond to syntactic functions such as subject, predicate, object and attribute) with the *Sb* attribute marking a subject. The values are read from the formeme. For Portuguese, if the formeme is a possessive noun or a noun marked as a subject, then the *Afun* will get the *Sb* value.
InitMorphcat fills the Interset (Zeman, 2008) morphological categories based on the corresponding grammateme and formeme. It also takes into account the morphological values for the gender, number and person in the case of a possessive pronoun.
ProjectClauseNumber sets the number of the clause for the finite verb clauses.
PrepositionContraction handles preposition contractions in Portuguese.
CapitalizeSentStart capitalises the first letter of the first (non-punctuation) token in the sentence.
ConcatenateTokens creates the final sentence as a concatenation of the a-nodes.

An example of the a-trees and t-trees is given in the Figure 2. This example also illustrates the use of *AddGender* and *AddPrepos* blocks. From the AddGender block the node '*imagem*' gets the tag '*fem*' which will make possible to the block AddPrepos to add the preposition '*a*'.

4. Evaluation and Error Analysis

In this section, we report on the automatic evaluation of the full translation pipeline (which contains analysis, transfer and synthesis phases) for the translation from English to Portuguese, and discuss the findings of the conducted error analysis.
As the test corpus, we use $1,000$ parallel sentences from the IT domain (QTLEAP-IT2), compiled under the QTLeap project in a similar way as the $1,000$ parallel sentences used for the development of the synthesis phase (none of the $1,000$ test sentences can be found among the $1,000$ sentences used for development).
For the automatic evaluation, we use the BLEU score (Papineni et al., 2002) and compare our TectoMT system with the standard phrase-based SMT system built using the Moses toolkit (Koehn et al., 2007). Both systems were trained over the full Europarl corpus (Koehn, 2005), consisting of 1,960,407 sentence pairs. The PBSMT was tuned using the QTLEAP-IT1 dataset (the same dataset which was used for the development of the rules for the synthesis phase of the TectoMT system).
As can be seen from the results presented in Table 1, our adaptation of TectoMT system to EN→PT translation task significantly outperforms the standard PBSMT system on the IT test set (the difference is statistically significant at a 0.05 level of significance, using the paired bootstrap resampling (Koehn, 2004)).
Comparing to other languages pairs translation (Rosa et al., 2015) using TectoMT, the translation from English to Portuguese falls within the same range of score values.

System	EN→PT
PBSMT	0.1899
TectoMT	0.2254

Table 1: Comparison of BLEU scores

5. Conclusions

We have presented an English→Portuguese hybrid deep MT system (TectoMT) which has been adapted from the existing English-Czech TectoMT system. Due to the TectoMT system having highly modular structure, the adaptation of the system to a new language pair only required the adaptation of the synthesis phase (blocks of Perl code) and the training of the transfer models. Both tasks required a total effort of 10 persons/month.
Some of the blocks were inherited from already existing synthesis blocks and modified to Portuguese, the others were written especially for the EN→PT system in order to cover the linguistic phenomena which are specific for Portuguese. In this paper, we described this process in details, additionally showing that it is possible to integrate the existing language tools (state-of-the-art tools for each specific language) in the TectoMT pipeline.

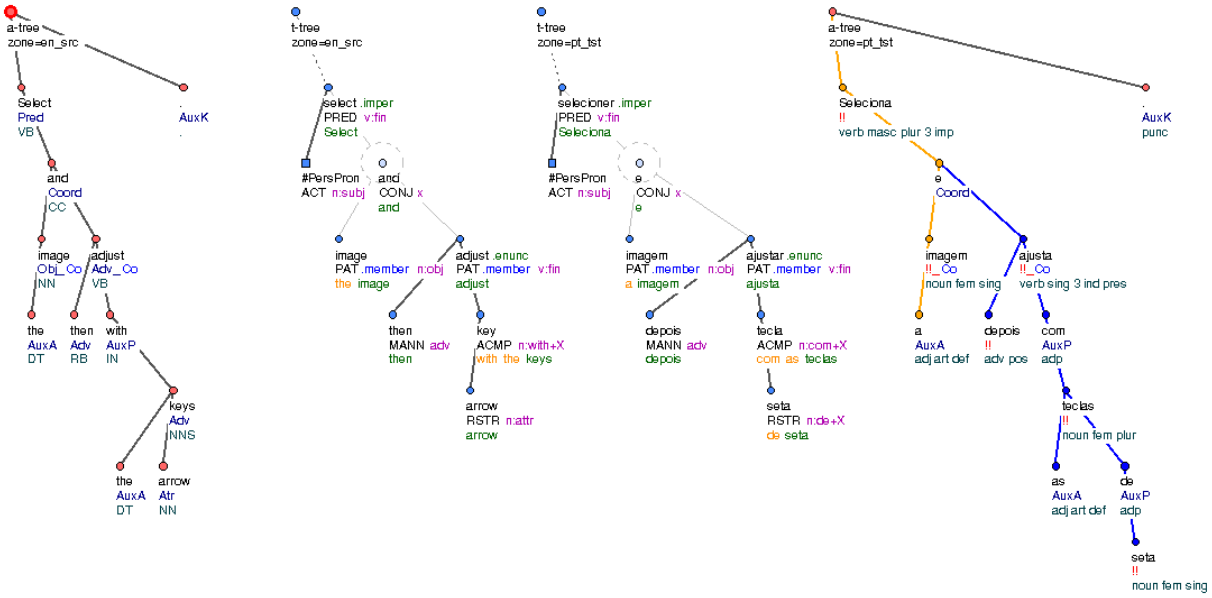

Figure 2: An example of a-trees and t-trees (EN: "Select the image and then adjust with the arrow keys." → PT: "Seleciona a imagem e depois ajusta com as teclas de seta.")

Our TectoMT system adapted to the EN→PT translation task (full system consisting of analysis, transfer, and synthesis phases) significantly outperformed the standard PBSMT system for a domain-specific translation task. These results showed that TectoMT is a promising machine translation system which can easily be adapted to a new language pair and obtain the results comparable to the standard PBSMT systems.

Acknowledgements

This work has received support by the EC's FP7 (FP7/2007-2013) under grant agreement number 610516: "QTLeap: Quality Translation by Deep Language Engineering Approaches".

6. Bibliographical References

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 workshop on statistical machine translation. In *Proceedings of the 8th Workshop on Statistical Machine Translation*, pages 1–44.

Branco, A. and Silva, J. R. (2006). A suite of shallow processing tools for portuguese: Lx-suite. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, pages 179–182. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C. J., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, pages 79–86.

Mareček, D., Popel, M., and Žabokrtský, Z. (2010). Maximum entropy translation model in dependency-based MT framework. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 201–206.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Popel, M. and Žabokrtskỳ, Z. (2010). Tectomt: modular nlp framework. In *Advances in natural language processing*, pages 293–304. Springer.

Rosa, R., Dušek, O., Novák, M., and Popel, M. (2015). Translation model interpolation for domain adaptation in tectomt. In *1st Deep Machine Translation Workshop*, page 89.

Sgall, P., Hajicová, E., and Panevová, J. (1986). *The meaning of the sentence in its semantic and pragmatic aspects*. Springer Science & Business Media.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*.

Žabokrtský, Z., Ptáček, J., and Pajas, P. (2008). TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170.

Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In *LREC*.