Bootstrapping a hybrid deep MT system

João Silva and **João Rodrigues** and **Luís Gomes** and **António Branco**
University of Lisbon, NLX—Natural Language and Speech Group
Faculdade de Ciências, Universidade de Lisboa
Edifício C6, Piso 3, Campo Grande, 1749-016 Lisboa, Portugal
{jsilva,joao.rodrigues,luis.gomes,antonio.branco}@di.fc.ul.pt

Abstract

We present a Portuguese↔English hybrid deep MT system based on an analysis-transfer-synthesis architecture, with transfer being done at the level of deep syntax, a level that already includes a great deal of semantic information. The system received a few months of development, but its performance is already similar to that of baseline phrase-based MT, when evaluated using BLEU, and surpasses the baseline under human qualitative assessment.

1 Introduction

Data-driven phrase-based MT has been, for many years, the technique that has achieved the best results in MT, much due to the availability of huge parallel data sets. Requiring such large amounts of training data is a hindrance for languages with fewer resources. Statistical MT (SMT) as an approach, however, may have intrinsic limitations that go beyond that of data availability.

The main weakness of current SMT methods ultimately stems from the limited linguistic abstraction that is employed, which leads to difficulties in correctly handling the translation of certain phenomena, such as getting the correct word order when translating between languages with different typology and in maintaining the semantic cohesion of the translated text.

SMT has attempted to tackle these issues by making use of richer linguistic structure, such as hierarchical methods and tree-to-tree mappings, but these methods have been unable to clearly improve on the phrase-based state-of-the-art.

There is a growing opinion that the previous approaches to SMT may be reaching a performance ceiling and that pushing beyond it will require approaches that are more linguistically informed and that are able to bring semantics into the process.

The classic analysis-transfer-synthesis architecture (the Vauquois triangle) provides a promising foundation onto which such approaches can be built. Underlying this architecture is the rationale that, the deeper the level of representation, the easier transfer becomes since deeper representations abstract away from surface aspects that are specific to a language. At the limit, the representation of the meaning of a sentence, and of all its paraphrases, would be shared among all languages.

This paper reports on our work of building a deep MT system, which translates between Portuguese and English, where transfer is performed at the level of a deep syntactic representation.

Portuguese is a widespread language, with an estimated 220 million speakers, and is the fifth most used language on the Web. Despite this, it is relatively less-resourced in terms of available NLP tools and resources (Branco et al., 2012). In this respect, the current work also allowed us to determine a minimal set of NLP tools required to get a deep MT system running, which helps to assess the feasibility of building such a system for under-resourced languages.

This paper is organized as follows. Section 2 presents the translation pipeline. Section 3 evaluates the system intrinsically by comparing it with a state-of-the-art phrase-based SMT approach, and extrinsically by human assessment in the context of a cross-lingual information retrieval task. Section 4 concludes with some final remarks.

2 Translation pipeline

Our pipeline is built upon the Treex system (Popel and Žabokrtský, 2010), a modular NLP framework used mostly for MT and the most recent incarnation of the TectoMT system (Žabokrtský et al., 2008). Treex uses an analysis-transfer-synthesis architecture, with transfer being done at the deep syntactic level, where a Tectogrammatical (Tecto) formal description is used.

The choice of Treex as the supporting framework was motivated by several reasons.

Firstly, Treex is a tried and tested framework that has been shown to achieve very good results in English to Czech translation, on a par with phrase-based SMT systems (Bojar et al., 2013).

Secondly, Treex uses a modular framework, where functionality is separated into *blocks* (of Perl code) that are triggered at different stages of the processing pipeline. This modularity means that we can easily add blocks that make use of existing Portuguese NLP tools and that handle Portuguese-specific phenomena.

Thirdly, English analysis and English synthesis are already provided in Treex, from the work of Popel and Žabokrtský (2010) with Czech, and should be usable in the our pipeline with only little adjustments.

An overview of each of the steps that form the Vauquois triangle—analysis, transfer and synthesis—follows below.

2.1 Analysis

Analysis proceeds in two stages. The first stage is a shallow syntactic analysis that takes us from the surface string to what in the Treex framework is called the a-layer (analytical layer), which is a grammatical dependency graph. The second stage is a deep syntactic analysis that takes us from the a-layer to the t-layer (tectogrammatical layer).

2.1.1 Getting the a-layer

We resort to LX-Suite (Branco and Silva, 2006), a set of pre-existing shallow processing tools for Portuguese that include a sentence segmenter, a tokenizer, a POS tagger, a morphological analyser and a dependency parser, all with state-of-the-art performance. Treex blocks were created to call and interface with these tools.

After running the shallow processing tools, the dependency output of the parser is converted into Universal Dependencies (UD, (de Marneffe et al., 2014)). These dependencies are then converted into the a-layer tree (a-tree) in a second step. Both steps are implemented as rule-based Treex blocks.

Taking this two-tiered approach to getting the a-tree—first to UD, then from UD to a-tree—has two benefits: (i) it allows us to partly reuse the existing Treex code for converting UD to a-tree, and (ii) it provides us with a way of converting our dependencies into UD, giving us a de facto standard format that may be useful for other applications.

2.1.2 Getting the t-layer

Converting the a-tree into a t-layer tree (t-tree) is done through rule-based Treex blocks that manipulate the tree structure.

The major difference between these two trees is that the a-tree, being surface oriented, has a node for each token in the sentence, while the t-tree, being semantically oriented, includes only content words as nodes. Accordingly, the t-tree has no nodes corresponding to auxiliary words, such as prepositions and subordinating conjunctions, but conversely has nodes that do not correspond to any surface word, such as nodes used for representing pro-dropped pronouns.[1]

2.2 Transfer

Transfer is handled by a tree-to-tree maximum entropy translation model (Mareček et al., 2010) working at the deep syntactic level of Tecto trees.

This transfer model assumes that the source and target trees are isomorphic. This limitation is rarely a problem since at the Tecto level, as one would expect from a deep syntactic representation, the source and target trees are often isomorphic.

Since the trees are isomorphic, the model is concerned only with learning mappings between t-tree nodes.

The model was trained over 1.9 million sentences from Europarl (Koehn, 2005). Each pair of parallel sentences, one in English and one in Portuguese, are analyzed by Treex up to the t-layer level, where each pair of trees are fed into the model.

2.3 Synthesis

Similarly to what was done in analysis, we create new Treex blocks, but resort to pre-existing tools when possible.

The pre-existing tools, for verbal conjugation and for nominal inflection, are rule-based and are used to handle the generation of surface forms.

The rule-based Treex blocks search for patterns over the trees and are used, for instance, to generate to correct word order, to enforce agreement, and to insert the auxiliary words (such as preposition and subordinating conjunctions) that were collapsed when building the t-tree.

[1] Some nodes are removed, but information is preserved as attributes of other nodes or in the relations between nodes.

Question I was typing in something and then a blank page appeared before my text, and I do not know how to remove it

Answer Move the mouse cursor to the beginning of the blank page and press the DELETE key as often as needed until the text is in the desired spot.

Figure 1: Question-Answer pair

3 Evaluation

This Section reports on both an intrinsic and an extrinsic evaluation, the latter made possible by embedding the system into a helpdesk application that provides technical support through an online chat interface. In this regard, the application can be seen as a Question Answering (QA) system.

Since most user questions address issues that have been dealt with previously, they are matched against a database of prior questions-answer pairs. If a matching question is found, the pre-existing answer is returned, thus avoiding the need for the intervention of a human operator.

The questions and the answers in the database are stored in English (see Figure 1 for an example). An MT component enables cross-lingual usage by automatically translating non-English queries into English prior to searching the database, and by automatically translating the answer from English into the language of the user of the application.

The MT component may then impact the QA application in two ways: (i) when translating the question (PT→EN), and consequently affect the ability of the QA system to retrieve the correct answer; and (ii) when translating the retrieved answer (EN→PT), and consequently affect properties of the translated retrieved answer such as its grammaticality, readability and fluency.

Given the workings of this QA application, we are concerned with evaluating translation in the PT→EN direction, for questions, and in the EN→PT direction, for answers.

The test corpus has been developed in the scope of the QTLeap Project. Each question is paired with an answer, both in English, and each of these question-answer pairs has a corresponding reference pair in Portuguese.[2]

[2]The QTLeap project also involves Basque, Bulgarian, Czech, Dutch, German and Spanish, each being paired with English in the same QA application.

	questions PT→EN	answers EN→PT
SMT (Moses)	0.2265	0.1899
Treex pipeline	0.1208	0.1943

Table 1: Comparison of BLEU scores

3.1 Intrinsic evaluation

The intrinsic evaluation is itself broken down into an automatic and a manual evaluation.

In the automatic evaluation, the standard BLEU metric is used to compare the Treex pipeline against a system built with Moses (Koehn et al., 2007) that represents the state-of-the-art SMT phrase-based approach. Like the transfer module in the translation pipeline, the SMT model is trained over 1.9 million sentences from Europarl.

The test set consists of $1,000$ question-answer pairs. The results of the automatic intrinsic evaluation are summarized in Table 1.

BLEU scores are low, though we note that the domain of the test corpus (technical support) is very different from the domain of Europarl. For questions, the BLEU score of the Treex pipeline is fairly worse than the score of Moses. Given the application we envisage, this is to be expected. The translated question is meant to be used as database query, and not for human eyes. As such, we have so far placed relatively little effort in improving the synthesis rules for English, since issues like word order errors, agreement mismatches and missing functional words often do not prevent the query from being successful.

BLEU does not necessarily correlate with human judgments. This points us towards manual evaluation as a better way to measure translation quality. Recall that the translation of the retrieved answer, unlike the translation of questions, is meant to be read by humans. As such, the manual evaluation that follows is done only for answers (EN→PT).

The intrinsic manual evaluation consists of a detailed manual diagnosis of the types of translation errors found. Translation errors are classified in a hierarchy of issues, following the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014), with the help of the open-source editor translate5.[3] The classification is done by two annotators. Each annotator analyzed the same 100 answers.

[3]http://www.translate5.net/

	SMT	Treex
top-1	72.8%	71.6%
top-2	84.3%	83.1%
top-3	87.8%	87.2%

Table 2: Answer retrieval

better variant	
Treex pipeline	30.8%
SMT (Moses)	13.0%
(no difference)	56.2%

Table 3: Variant ranking

Almost two-thirds of the errors fall under the top-level category Fluency, with nearly 80% of these being classified as Grammar errors, the MQM category that includes issues such as word order, extra words, missing words, agreement problems, among others. The remaining third of the errors are in the top-level category Accuracy, which covers issues where meaning is not preserved, such as mistranslations of domain-specific terminology.

3.2 Extrinsic evaluation

The extrinsic evaluation consists of comparing two variants of the cross-lingual QA application, one using the baseline SMT for translation and another using the Treex translation pipeline.

For a given query, the QA system returns a list of answers, each associated with a confidence score.[4] For each variant, we measure if the correct answer is the first result (top-1) or among the top-2 or top-3 returned results. The summary in Table 2 shows that there is little difference between the variants. The Treex pipeline has a lower BLEU for questions, but this does not negatively impact answer retrieval.

While retrieval using the translated question is working well, the quality and usefulness of the helpdesk application ultimately hinges on the quality of the answer that is presented to the user and whether it is correct and clear enough to help the user solve their technical problem.

To evaluate this, a total of six human evaluators were asked to assess the quality of the translated answer. Their task was, given a reference question-answer pair, to compare both translated answers (anonymized and in random order) with the reference answer and pick the best translation, allowing for ties.

While in most cases there is not a clearly better variant, the output of the Treex pipeline is better than the output of the SMT system in 30.8% of the cases and worse in only 13.0% of the cases, as shown in Table 3. Inter-annotator agreement, as a ratio of matched annotations, was 0.628.

[4]The confidence score is based on several factors, such as lexical similarity and the number of times a given answer was used. In the current study, the QA engine is used as a black box and its details are outside the scope of this paper.

4 Conclusion

We have presented a Portuguese↔English hybrid deep MT system that, though still under development, achieves a BLEU score similar to that of a SMT system using the state-of-the-art phrase-based approach and, more importantly, is deemed by human evaluators to produce a text with better quality than the SMT system when embedded as part of a QA application.

The system uses an analysis-transfer-synthesis architecture, with transfer being done at the level of deep syntactic trees. This level is oriented towards semantic information, abstracting away auxiliary words while including nodes that do not correspond to any surface word.

Analysis begins by using a set of pre-existing statistical shallow processing tools for Portuguese to produce a grammatical dependency graph. This level of linguistic annotation can be seen as the minimal requirement for bootstrapping a similar deep MT system for other languages. The final step of analysis is rule-based, converting dependency graph into a deep representation. Following statistical transfer, the generation of the target surface form is also a rule-based process.

Evaluation results are very promising and the analysis-transfer-synthesis approach that is used allows much room for improvement apart from just adding more parallel data.

For instance, ongoing research is working towards enriching the pipeline with additional semantic information by plugging in tools for word sense and named-entity disambiguation into the analysis phase, thus providing the transfer phase with disambiguated terms.

Acknowledgments

This work was partly funded by the EU project QTLeap (EC/FP7/610516) and the Portuguese project DP4LT (PTDC/EEI-SII/1940/2012).

References

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 workshop on statistical machine translation. In *Proceedings of the 8th Workshop on Statistical Machine Translation*, pages 1–44.

António Branco and João Silva. 2006. A suite of shallow processing tools for Portuguese: LX-Suite. In *Proceedings of the 11th European Chapter of the Association for Computational Linguistics*, pages 179–182.

António Branco, Amália Mendes, Sílvia Pereira, Paulo Henriques, Thomas Pellegrini, Hugo Meinedo, Isabel Trancoso, Paulo Quaresma, and Vera Lúcia Strube de Lima. 2012. *A Língua Portuguesa na Era Digital / The Portuguese Language in the Digital Age*. White Paper. Springer. ISBN 978-3-642-29592-8.

Marie-Catherine de Marneffe, Natalia Silveira, Timothy Dozat, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the 9th Language Resources and Evaluation Conference*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher J. Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 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.

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

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0(12):455–463.

David Mareček, Martin Popel, and Zdeněk Žabokrtský. 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.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP framework. In *Proceedings of the 7th International Conference on Natural Language Processing*, pages 293–304.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 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.