Machine Translation for Multilingual Troubleshooting in the IT Domain: A Comparison of Different Strategies

Sanja Štajner and **João Rodrigues** and **Luís Gomes** and **António Branco**
Department of Informatics, Faculty of Sciences
University of Lisbon, Portugal
{sanja.stajner, joao.rodrigues, luis.gomes, antonio.branco}
@di.fc.ul.pt

Abstract

In this paper, we address the problem of machine translation (MT) of domain-specific texts for which large amounts of parallel data for training are not available. We focus on the IT domain and on English to Portuguese machine translation, and compare different strategies for improving system performance over two baselines, the first using only large dataset of out-of-domain data, and the second using only a small dataset of in-domain data. Our results indicate that adding a domain-specific bilingual lexicon to the training dataset significantly improves the performance of both a hybrid MT system and a PBSMT system, while adding out-of-domain sentence pairs to the training dataset only improves the performance of a hybrid MT system. Furthermore, we perform a human evaluation of the sentences generated by the hybrid MT system and the standard PBSMT system built using the same training datasets. The results indicate some significant differences between those two MT approaches in this specific task.

1 Introduction

Although the problem of machine translation has been extensively studied in the last 30 years and is one of the main topics of the natural language processing (NLP), English to Portuguese MT is rarely addressed.

Our work aims to fill that gap by addressing the problem of English to Portuguese MT for a specialised domain (the IT domain) using two MT approaches: the standard PBSMT system and a hybrid MT system based on deep translation approach. We focus on translation from English to Portuguese of short sentences taken from real-usage scenarios, where user questions are followed by answers from an IT technician. The data was gathered in a continuous way during user interaction with a technical support team via chat. We explore three different strategies for enlarging the training dataset: (1) adding an in-domain bilingual terminology; (2) adding a certain portion of the out-of-domain corpus; and (3) adding both an in-domain bilingual terminology and a certain portion of the out-of-domain corpus. Our objective is to explore which of the three strategies leads to greater improvements in the system performance for each of the two MT approaches (PBSMT and hybrid MT). In order to gain a better insight into strengths and weaknesses of both MT systems, we also conduct a human evaluation and error analysis of their output sentences.

The remainder of the paper is organised as follows: Section 2 introduces studies that are relevant to our work; Section 3 describes the corpora, MT systems, experimental setup, goals and evaluation procedures; Section 4 presents and discusses the results of both automatic and human evaluation; and Section 5 summarises the findings of this study and gives directions for future work.

2 Related Work

The rule-based machine translation (MT) systems, such as Systran (Toma, 1977), ETAP-3 (Boguslavsky, 1995), and Lucy (Alonso and Thurmair, 2003), required linguistic expertise to operate and were difficult

Proceedings of the 1st Deep Machine Translation Workshop (DMTW 2015), pages 106–115,
Praha, Czech Republic, 3–4 September 2015.

to adapt to different languages. The emergence of the word-based IBM models (Brown et al., 1988; Brown et al., 1990; Brown et al., 1993) heralded a new approach to MT – statistical machine translation (SMT) systems. Later, the word-based SMT models were replaced by better-performing phrase-based (Koehn et al., 2007) or hierarchical phrase-based (Li et al., 2009) SMT systems. However, it was noticed that those shallow SMT approaches which do not use any deeper linguistic information or syntax are not able to capture long-distance dependences and may lead to problems with word order and grammatical and semantic cohesion (Fishel et al., 2012). Shallow syntax-based SMT systems tried to address those issues using three different approaches: a *tree-to-string* translation, where linguistic information is applied only on the source side (Huang et al., 2006); a *string-to-tree* translation, where linguistic information is applied only on the target side (Galley et al., 2004), and a *tree-to-tree* translation, where linguistic information is applied on both source and target side (Eisner, 2003). However, for the majority of language pairs, phrase-based SMT systems still produce better results.

The main limitation of SMT systems is that they require large amounts of parallel (or at least comparable) training data, which is hard to obtain for language pairs not covered by the Europarl corpora (Koehn, 2005). Even if Europarl contains data for a particular language pair, another problem arises if the SMT system is needed for a different domain, as the training data may not cover the specific vocabulary or sentence constructions present in the targeted domain. In order to address this problem, many domain-adaptation techniques for SMT have been proposed, ranging from simply adding out-of-domain data to the small amount of in-domain data for training (Foster and Kuhn, 2007) to more sophisticated techniques, such as selecting only particular sentences from the out-of-domain data which are most similar to the in-domain data (Axelrod et al., 2011) or are similar to the sentences with the lowest translation quality (Banerjee et al., 2015).

Hybrid MT systems, in turn, aim to exploit the best of both SMT and rule-based approaches, usually either by combining rule-based transfer with statistical language models in the synthesis phase (Habash and Dorr, 2002), or by combining rule-based with statistical approaches at different points of the Vauquois triangle, as the TectoMT system (Žabokrtský et al., 2008) that we use in this study.

2.1 English-Portuguese MT

The English-Portuguese translation model built using the standard PBSMT system in the Moses toolkit (Koehn et al., 2007), trained on the largest existing parallel corpora for this language pair (the JRC-Acquis corpus (Steinberger et al., 2006)) achieves a BLEU score (Papineni et al., 2002) of 55% (Koehn et al., 2009). The standard PBSMT system in the Moses toolkit trained on the Fapesp-v2 corpus of English-Brazilian Portuguese texts from the Brazilian scientific news magazine *Revista Pesquisa FAPESP*[1] (Aziz and Specia, 2011) achieves 46.28% BLEU score (Salton et al., 2014).

To the best of our knowledge, there have been no studies reporting performances of English to Portuguese MT systems for any domain-specific tasks, neither have there been any studies comparing different MT approaches for this language pair.

3 Methodology

The next four subsections describe the corpora (Section 3.1), MT systems (Section 3.2), experimental setup and the main goal of the translation experiments (Section 3.3), as well as the human evaluation procedure (Section 3.4).

3.1 Corpora

We used four corpora in this study:

1. **EP** – Europarl corpus (Koehn, 2005) with English on the source side and Portuguese on the target side (1,960,407 sentence pairs) was used as the large out-of-domain corpus.

2. **IT1** – An in-domain IT corpus with 2,000 sentence pairs (1,000 questions and 1,000 answers) compiled under the QTLeap project[2].

[1] http://revistapesquisa.fapesp.br/

[2] http://qtleap.eu/

Corpora	Source (EN)	Target (PT)
TERM	arrow key gatekeeper Planning System Database	tecla de seta controlador de chamadas Base de Dados do Sistema de Planeamento
IT1	If your disc is not recognized, try changing the USB port. Which antivirus should I keep, MSE or AVG?	Se o disco não está a ser reconhecido, tente trocar de entrada USB. Qual antivrus devo manter, MSE ou AVG?
IT2	In the Insert menu, select Picture. In the taskbar there is an icon shaped like binoculars, click and type in what you want to search.	No menu inserir selecione Imagem. Na barra de Tarefas há um ícone em forma de binóculos, clique e escreva o que pretende procurar.
EP	Please rise, then, for this minute's silence. You have requested a debate on this subject in the course of the next few days, during this part-session.	Convido-os a levantarem-se para um minuto de silêncio. Os senhores manifestaram o desejo de se proceder a um debate sobre o assunto nos próximos dias, durante este período de sessões.

Table 1: Examples from the corpora

3. **IT2** – Another in-domain IT corpus, with 1,000 sentence pairs (answers only) compiled under the QTLeap project, and comparable with the IT1 corpus.[3]

4. **TERM** – A parallel corpus of IT terminology (unigrams or multiword expressions), which consists of the *Microsoft Terminology Collection*[4] (13,030 terms) and a small portion of LibreOffice terminology[5] (995 terms).

Examples from each corpora are presented in Table 1.

3.2 Systems

This section describes the two MT systems used for the experiments.

3.2.1 TectoMT

TectoMT (Žabokrtský et al., 2008) is a structural MT system which uses two layers of structural description, the shallow a-layer and the deep t-layer, performing the transfer on the t-layer (Figure 1). It encompasses three phases along the Vauquois triangle: analysis (which transforms the input sentence into the a-layer and t-layer in a two-step process), transfer (at the t-layer), and synthesis (which converts the translated t-layer representation to the a-layer and then to the output surface string). The analysis and synthesis phases are hybrid, while the transfer phase is mostly statistical, based on the Maximum Entropy context-sensitive translation models (Mareček et al., 2010).

In the analysis stage, all tokens from the input English sentence are first transformed into nodes in a labeled dependency tree (a-tree) to form a surface syntax layer (analytical layer or a-layer). This is achieved using various NLP tools that perform sentence splitting, tokenisation, morphological tagging, and dependency parsing. We follow the annotation pipeline used for the CzEng 1.0 parallel corpus (Bojar et al., 2012), using the Morče tagger (Spoustová et al., 2007) and the Maximum Spanning Tree parser (McDonald et al., 2005) trained on the CoNLL-2007 conversion of Penn Treebank (Nilsson et al., 2007). Dependencies are further transformed by the rule-based blocks into the a-layer which contains

[3]The decision to test the systems only on the answers is the result of the nature of the task in the QTLeap project.

[4]`https://www.microsoft.com/Language/en-US/Terminology.aspx`

[5]We would like to thank Eleftherios Avramidis and Lukas Poustka for making the LibreOffice corpus available to us.

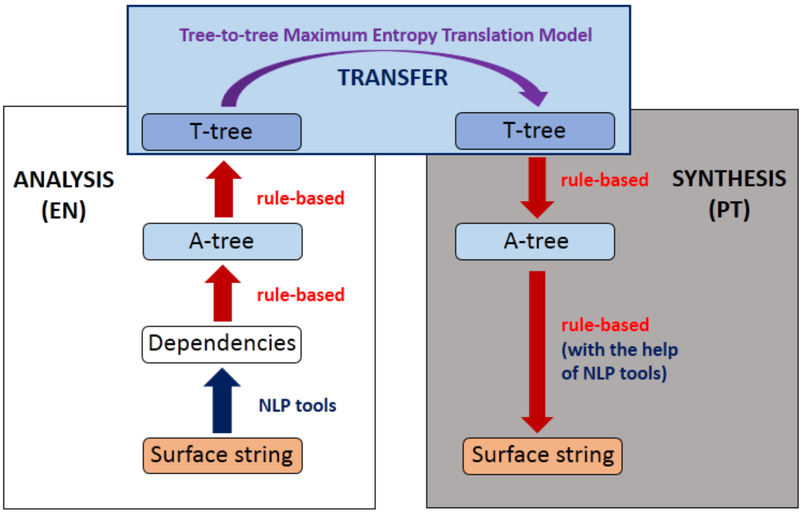

Figure 1: Schema of the TectoMT system

the corresponding *word forms*, *lemmas*, *morphological tags* and *afun* labels (which denote syntactic functions such as subject, predicate, object and attribute).

The next step in the analysis stage is performed using another rule-based block that converts a-trees into t-trees (tectogrammatical layer or t-layer). The t-layer describes the input sentence according to the Functional Generative Description (GFD), and unlike the a-layer (which contains all input tokens), the t-layer only contains content words as nodes (t-nodes). Auxiliary words, such as prepositions, subordinating conjunctions or auxiliary verbs, become attributes of the t-nodes. This is illustrated in an example of the a-layers and t-layers in Figure 2. The t-layer can also introduce new nodes (which did not exist in the a-layer), as for example, in the case of pro-dropped subject personal pronouns which do not correspond to any token in the input sentence.

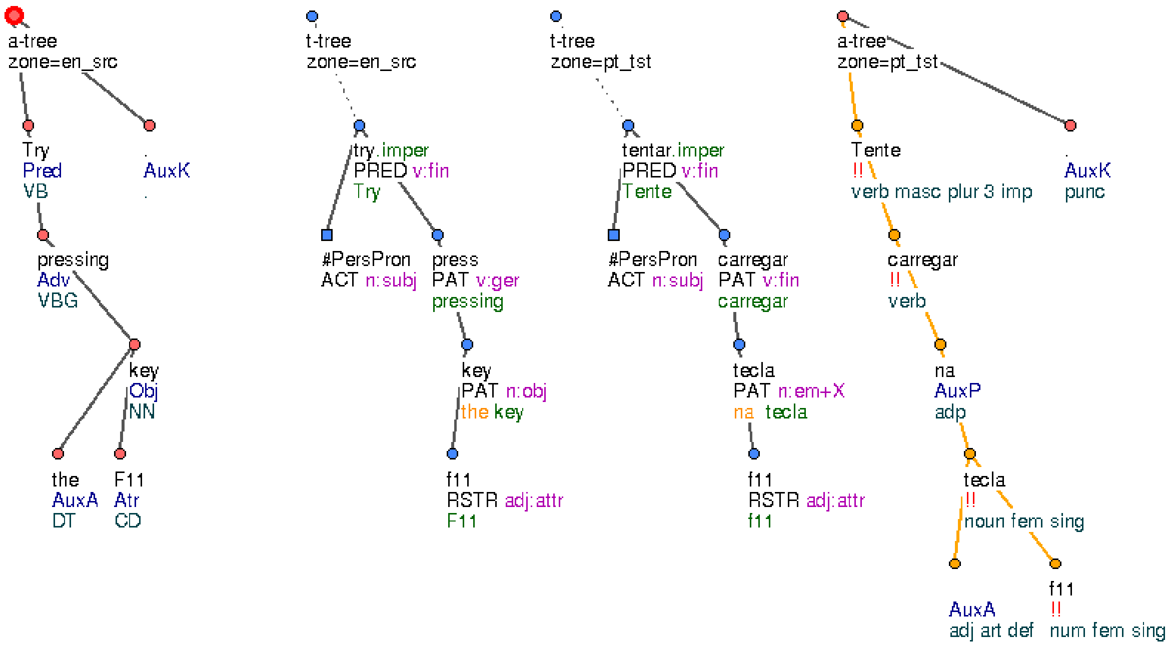

Figure 2: An example of the a-trees and t-trees in the TectoMT system (the input EN sentence: “Try pressing the F11 key.” translated into the output PT sentence: “Tente carregar na tecla f11.”)

After the transfer of the English t-trees into Portuguese t-trees, the synthesis phase constructs a flat surface form of the sentence from the Portuguese t-tree. This is achieved using additional rule-based blocks which take care of word reordering, insertion of negations, prepositions, conjuctions, correct agreement, compound verb forms, etc. The synthesis stage for Portuguese uses the LX-Suite (Branco and Silva, 2006) to perform such tasks.

The expected advantage of the TectoMT system over the standard PBSMT system is that the TectoMT translates t-tree nodes (and not the inflected forms) and should thus be able to generalise over the unseen morphological forms. This is particularly important for translation into morphologically rich languages (such as Portuguese) where data sparseness presents a problem for a purely statistically driven MT systems.

3.2.2 PBSMT

In all experiments, we use the same PBSMT model (Koehn et al., 2007), GIZA++ implementation of the IBM word alignment model 4 (Och and Ney, 2003), and the refinement and phrase-extraction heuristics as described by Koehn *et al.* (2003). We tune the systems using MERT (Minimum Error Rate Training (Och, 2003)) and build a 5-gram language model with Kneser-Ney smoothing trained with SRILM (Stolcke, 2002) on the whole target side (Portuguese) of the English to Portuguese Europarl corpus (Koehn, 2005), which contains 1,960,407 sentences.

3.3 Experiments

In all experiments, the PBSMT system uses the in-domain IT1 corpus for tuning, and the language model (LM) is trained on all sentences in the Portuguese side of the Europarl corpus (EP)[6]. All experiments (in both TectoMT and PBSMT systems) are evaluated on the same test dataset (IT2). In order to obtain two baselines for each MT approach (TectoMT and PBSMT) we train both systems on: (1) the full Europarl corpus (EP) as the out-of-domain large corpus (BaselineEP), and (2) the IT1 as the in-domain small corpus (BaselineIT).

In the next four experiments (IT+TERM, IT+EP1, IT+EP10, IT+EP10+TERM), we use the in-domain IT1 corpus as the basis for the training. As this corpus is very small (2,000 sentence pairs only), we explore three different strategies for enlarging the training dataset:

(S1) Adding an in-domain bilingual terminology (the TERM corpus in the IT+TERM experiment);

(S2) Adding a certain portion of the out-of-domain EP corpus (1,000 sentence pairs in the IT+EP1 experiment, and 10,000 sentence pairs in the IT+EP10 experiment);

(S3) Adding both an in-domain bilingual terminology and a certain portion of the out-of-domain EP corpus (10,000 sentence pairs from the EP corpus and the TERM corpus in the IT+EP10+TERM experiment)

3.4 Human Evaluation

In order to better assess strengths and weaknesses of both approaches (TectoMT and PBSMT), we also conduct a human evaluation of the sentences generated by both systems for 100 sentence pairs from the test set for the IT+TERM experiments (which led to the highest BLEU score for the PBSMT approach and the second highest BLEU score for the TectoMT approach).

3.4.1 Fluency and Adequacy

We ask two native speakers of Portuguese (both employed as linguists) to evaluate the fluency and adequacy of the machine translation obtained by the TectoMT and PBSMT systems trained on the IT+TERM dataset. We follow the TAUS guidelines[7], which suggest a 1–4 scale for both aspects.

[6]Note that TectoMT does not need a development dataset and language model.

[7]`https://www.taus.net/think-tank/best-practices/evaluate-best-practices/adequacy-fluency-guidelines`

Fluency rates “the extent to which the translation is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker”:

4 – Flawless
3 – Good
2 – Disfluent
1 – Incomprehensible

Adequacy rates “how much of the meaning expressed in the source is also expressed in the target translation”:

4 – Everything
3 – Most
2 – Little
1 – None

3.4.2 Error Analysis

Following the error classification proposed by Costa-jussà and Farrús (2015) for evaluation of MT from Spanish to Catalan, we asked human evaluators to classify errors of each sentence into four classes:

1. **Orthographic:** punctuation marks, accents, upper- and lowercase, letters, joined/split words, extra spaces, apostrophe;

2. **Morphologic:** gender concord, number concord, verbal morphology (tense, aspect), lexical morphology (POS);

3. **Semantic:** polysemy, homonymy, incorrect meaning, untranslated words (left in the source language), missing words;

4. **Syntactic:** prepositions, relative pronouns, verbal periphrasis, clitics, articles, reorderings.

4 Results

The next two subsections present the results of the automatic evaluation of all experiments (Section 4.1), and the human evaluation and error analysis of the selected pair of experiments (Section 4.2).

4.1 Automatic Evaluation

The experimental setup for each experiment (the type and the size of the corpora used) and the obtained BLEU scores on the whole test set are presented in Table 2.

All four experiments (IT+TERM, IT+EP1, IT+EP10, and IT+EP10+TERM) of the TectoMT system significantly outperformed both baselines indicating that in the TectoMT approach both strategies (adding different portions of the out-of-domain corpus, and adding bilingual terminology) lead to significant improvements over the BaselineIT. The combination of both strategies (IT+EP10+TERM) resulted in the highest achieved BLEU score (significantly better than all others for the TectoMT system).

For the PBSMT approach, the only two experiments which significantly outperformed the BaselineIT were those trained on the IT+TERM and on the IT+EP10+TERM corpora. This suggests that, for a PBSMT system, adding terminology has a greater impact than adding the out-of-domain corpus. In fact, adding a small portion of out-of-domain corpus (1,000 sentence pairs from EP) to the training dataset negatively influenced the system’s performance, resulting in a BLEU score significantly lower than the BaselineIT. Adding a larger portion of the out-of-domain corpus (10,000 sentence pairs from EP) seems not to influence the system’s performance significantly.

Experiment	Training			Dev.	Test	Results (BLEU score)	
	EP	TERM	IT1	IT1	IT2	TectoMT	PBSMT
BaselineEP	all	/	/	2,000	1,000	19.34	18.99
BaselineIT	/	/	2,000	2,000	1,000	20.77	21.55
IT+TERM	/	14,025	2,000	2,000	1,000	**21.89**	**22.73**
IT+EP1	1,000	/	2,000	2,000	1,000	**20.97**	*21.08
IT+EP10	10,000	/	2,000	2,000	1,000	**21.16**	21.66
IT+EP10+TERM	10,000	14,025	2,000	2,000	1,000	**22.20**	**22.16**

Table 2: Translation experiments setup – type and the size of the corpora used (the number of sentence pairs for the IT1, IT2, and EP corpora, and the number of unigram or multiword expression pairs in the case of the TERM corpus), and the results of the automatic evaluation (the results of the systems which significantly outperformed both baselines are shown in bold; the '*' marks the result which is significantly lower than the result for the BaselineIT; statistical significance is calculated using paired bootstrap resampling (Koehn, 2004))

4.2 Human Evaluation Results

The results of our human evaluation of the fluency and adequacy of the output are presented in Table 3. For each sentence we additionally calculate the *Total* score (for each annotator separately) as the rounded arithmetic mean of its *Fluency* and *Adequacy* scores. The TectoMT system achieved significantly higher adequacy score and total score than the PBSMT system. The mean and median value of the fluency score in the TectoMT system was higher than in the PBSMT system, but the reported difference was not statistically significant (at a 0.05 level of significance using the marginal homogeneity test).

Aspect	Mean		Median		Mode		Sign.	IAA
	TectoMT	PBSMT	TectoMT	PBSMT	TectoMT	PBSMT		
Fluency	**1.78**	1.74	**2**	1.5	2	2	0.054	0.52
Adequacy	**2.28**	2.24	2	2	2	2	**0.047**	0.55
Total	**2.27**	2.23	2	2	2	2	**0.048**	0.55

Table 3: Results of the human evaluation of the fluency and adequacy on a 1–4 scale where higher score denotes better output (IAA is calculated as the squared Cohen's κ, and the statistical significance is calculated in SPSS using the marginal homogeneity test which represent the extension of McNemar test from binary to multinominal response for two related samples)

Errors	Mean		Median		Mode		Sign.	IAA
	TectoMT	PBSMT	TectoMT	PBSMT	TectoMT	PBSMT		
Orthographic	1.15	**0.95**	1.25	**1**	1.5	**1**	**0.001**	0.50
Morphologic	0.97	**0.74**	1	**0.5**	1	**0**	**0.000**	0.54
Syntactic	1.31	**1.26**	1.5	1.5	1.5	1.5	**0.045**	0.49
Semantic	**1.37**	1.5	1.5	1.5	2	2	**0.009**	0.53

Table 4: Results of the error analysis on a 0–2 scale where 0 – no errors, 1 – one error, and 2 – two or more errors (IAA is calculated as the squared Cohen's κ, and the statistical significance is calculated in SPSS using the marginal homogeneity test which represent the extension of McNemar test from binary to multinominal response for two related samples)

The results of the error analysis of the output sentences are presented in Table 4. The number of orthographic, morphologic, and syntactic errors was found to be significantly higher in the output of the TectoMT system than in the output of the PBSMT system, while the number of semantic errors was significantly higher in the PBSMT system.

Comparison	Scores			Number of errors			
	Fluency	Adequacy	Total	Ortho.	Morpho.	Synt.	Sem.
TectoMT>PBSMT	47	55	55	69	81	58	98
TectoMT=PBSMT	117	96	96	96	77	85	102
TectoMT<PBSMT	36	49	49	35	42	57	60

Table 5: Comparison of the outputs of the TectoMT and PBSMT systems on a sentence level (TectoMT>PBSMT for *Scores* signifies better output of the TectoMT than PBSMT system, while TectoMT>PBSMT for *Number of errors* signifies worse output of the TectoMT than PBSMT system)

In order to achieve sentence-to-sentence comparison between the two systems, we calculate:

1. How many times was the output of the TectoMT system rated as better (TectoMT>PBSMT), equal (TectoMT=PBSMT), or worse (TectoMT<PBSMT) than the output of the PBSMT system; and

2. How many times did the output of the TectoMT system contain more (TectoMT>PBSMT), equal number (TectoMT=PBSMT), or less (TectoMT<PBSMT) errors of each of the four types (orthographic, morphologic, semantic, and syntactic) than the output of the PBSMT system.

In this calculation, we compare the outputs of the TectoMT and PBSMT for each original sentence and each annotator separately, a total of 200 comparisons. The results are presented in Table 5. It seems that the sentences generated by the TectoMT system tend to represent more fluent and adequate translation than those generated by the standard PBSMT system. However, the results also show that the number of cases in which the output of the TectoMT system contains more errors than the output of the PBSMT system is greater than the number of cases in which the output of the PBSMT system contains more errors than the output of the TectoMT system. These results indicate that either: (1) the fluency of a sentence cannot be well captured by counting its orthographic, morphological, and syntactic errors, and the adequacy of a sentence cannot be well captured by counting its semantic errors, or (2) the errors produced by the TectoMT system are not as severe as the errors produced by the standard PBSMT system, and thus were, not as severely penalised in terms of fluency and adequacy scores.

5 Conclusions and Future Work

The experiments presented in this paper address the problem of English to Portuguese machine translation of the domain-specific texts (text of the IT domain in this particular case), and report on results obtained using three different techniques to enlarge the training datasets for two MT approaches: the standard PBSMT approach, and the hybrid deep MT approach employed in the TectoMT system.

Our results indicate that adding in-domain bilingual terminology, as well as adding a combination of in-domain bilingual terminology and out-of-domain sentence pairs, significantly improves the performance of both systems. Adding only some portion of out-of-domain sentence pairs, however, only improves the performance of the TectoMT system, while it either impairs or does not significantly change the performance of the standard PBSMT system.

A human evaluation of the output generated by the PBSMT and TectoMT systems revealed better meaning preservation (adequacy score) in the TectoMT system. However, the error analysis showed that the TectoMT system led to a higher number of sentences that had a greater number of orthographic, morphological, syntactic and semantic errors.

We acknowledge that both systems have room for improvement, and thus this work should only be regarded as preliminary. We used only the basic domain-adaptation technique for the PBSMT system, and no domain-adaptation techniques for the TectoMT. In future, the focus will be on implementing the state-of-the-art domain-adaptation techniques for the PBSMT system, as well as on exploring the possibilities of domain adaptation in the TectoMT.

Acknowledgements

This research was funded by the EC's QTLeap project (FP7-ICT-2013-10-610516) and the Portuguese DP4LT project (PTDC/EEI-SII/1940/2012).

References

Juan A. Alonso and Gregor Thurmair. 2003. The Comprendium Translator system. In *Proceedings of the Ninth Machine Translation Summit*.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, pages 355–362.

Wilker Aziz and Lucia Specia. 2011. Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*, Cuiabá, MT, Obtober.

Pratyush Banerjee, Raphael Rubino, Johann Roturier, and Josef van Genabith. 2015. Quality estimation-guided supplementary data selection for domain adaptation of statistical machine translation. *Machine Translation*, 29(2):77–100.

Igor Boguslavsky. 1995. A bi-directional Russian-to-English machine translation system (ETAP-3). In *Proceedings of the Fifth Machine Translation Summit*.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2012. The joy of parallelism with CzEng 1.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 3921–3928.

António Branco and João Silva. 2006. A Suite of Shallow Processing Tools for Portuguese: LX-Suite. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Peter F. Brown, John Cocke, Stephen A. Della-Pietra, Vincent J. Della-Pietra, Frederick Jelinek, Robert L. Mercer, and Paul Rossin. 1988. A statistical approach to language trans- lation. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Peter F. Brown, John Cocke, Stephen A. Della-Pietra, Vincent J. Della-Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul Rossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):76–85.

Peter F. Brown, Stephen A. Della-Pietra, Vincent J. Della-Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–313.

Marta R. Costa-jussà and Mireia Farrús. 2015. Towards human linguistic machine translation evaluation. *Digital Scholarship in the Humanities*, 30(2):157–166.

Jason Eisner. 2003. Learning non-isomorphic tree mappings for machine translation. In Yuji Matsumoto, editor, *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, pages 205–208.

Mark Fishel, Ondrej Bojar, and Maja Popovic. 2012. Terra: a collection of translation error-annotated corpora. In *Proceedings of LREC*, pages 7–14.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation (StatMT)*, pages 129–135.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. Whats in a translation rule? In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*.

Nizar Habash and Bonnie J. Dorr. 2002. Handling translation divergences: Combining statistical and symbolic techniques in generation-heavy machine translation. In Stephen D. Richardson, editor, *Machine Translation: From Research to Real Users, 5th Conference of the Association for Machine Translation in the Americas (AMTA)*, volume 2499 of *Lecture Notes in Computer Science*.

Liang Huang, Kevin Knight, and Aravind Joshi. 2006. A syntax-directed translator with extended domain of locality. In *Proceedings of the Workshop on Computationally Hard Problems and Joint Inference in Speech and Language Processing*, pages 1–8.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pages 48–54. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej 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 (ACL)*. Association for Computational Linguistics.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 2009. 462 machine translation systems for europe. In *Proceedings of the MT Summit XII*.

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

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

Zhifei Li, Chris Callison-Burch, Chris Dyer, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. 2009. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139.

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.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (EMNLP)*, pages 523–530.

Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL shared task session of EMNLP-CoNLL*, pages 915–932.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Giancarlo D. Salton, Robert J. Ross, and John Kelleher. 2014. An Empirical Study of the Impact of Idioms on Phrase Based Statistical Machine Translation of English to Brazilian-Portuguese. In *Proceedings of the third workshop on Hybrid Approaches to Transla- tion (HyTra), EACL*.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. The best of two worlds: Cooperation of statistical and rule-based taggers for czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, pages 67–74.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufi s, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the international conference on Language Resources and Evaluation (LREC)*.

Andreas Stolcke. 2002. SRILM - an Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 901–904.

Peter Toma. 1977. Systran as a multilingual machine translation system. In *Proceedings of the Third European Congress on Information Systems and Networks, Overcoming the language barrier*, pages 569–581.

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.