

Use of Domain-Specific Language Resources in Machine Translation

Sanja Štajner, Andreia Querido, Nuno Rendeiro, João Rodrigues, and António Branco

Department of Informatics, Faculty of Sciences

University of Lisbon, Portugal

{sanja.stajner, joao.rodrigues, andreia.querido, nuno.rendeiro, antonio.branco}@di.fc.ul.pt

Abstract

In this paper, we address the problem of Machine Translation (MT) for a specialised domain in a language pair for which only a very small domain-specific parallel corpus is available. We conduct a series of experiments using a purely phrase-based SMT (PBSMT) system and a hybrid MT system (TectoMT), testing three different strategies to overcome the problem of the small amount of in-domain training data. Our results show that adding a small size in-domain bilingual terminology to the small in-domain training corpus leads to the best improvements of a hybrid MT system, while the PBSMT system achieves the best results by adding a combination of in-domain bilingual terminology and a larger out-of-domain corpus. We focus on qualitative human evaluation of the output of two best systems (one for each approach) and perform a systematic in-depth error analysis which revealed advantages of the hybrid MT system over the pure PBSMT system for this specific task.

Keywords: Qualitative human evaluation, domain-specific machine translation, hybrid machine translation, TectoMT

1. Introduction

Phrase-based statistical machine translation (PBSMT) systems are considered the state of the art for language pairs for which large amounts of parallel data for training are available. For vast majority of language pairs (English-Portuguese among them), however, the available corpora are usually limited on one or two particular domains, e.g. legal documents (JRC-Acquis corpus), or parliamentary discussions (Europarl corpus). In those cases, for domain-specific MT, it is believed that rule-based or hybrid MT systems have higher potential to overcome the problems of data sparsity.

In this paper, we focus on English to Portuguese MT for Information Technology (IT) domain, motivated by the following real world usage scenario: an user asks a question in Portuguese, the question is machine translated into English, the answer is searched for in an English database, automatically translated back to Portuguese and presented to the user (Figure 1).

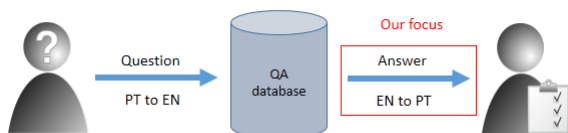


Figure 1: The task.

All experiments have been performed within the QTLeap project,¹ which aims to investigate an articulated methodology for machine translation based on deep language engineering approaches. As no freely available corpora for the EN-PT language pair exists for the IT domain, a small EN-PT corpus was compiled under the QTLeap project in order to enable in-domain examples and guide the hand-crafted rules for the synthesis phase of the hybrid MT system being

developed. This QTLeap corpus consists of recorded interactions of real users with experts to obtain technical support via chat, which were translated by professional translators into the eight languages of the project.²

We perform a series of MT experiments using two systems: (1) the TectoMT (Žabokrtský et al., 2008) adapted to English-Portuguese translation (Silva et al., 2015), as a hybrid system with hybrid analysis, rule-based synthesis and statistically based transfer, and (2) the standard PBSMT system in Moses (Koehn et al., 2007). We vary the training datasets exploring three different strategies to overcome the problem of small amount of training data: (1) adding larger out-of-domain dataset, (2) adding in-domain bilingual terminology, (3) adding both.

The main contribution of this paper is the in-depth error analysis, showing the error patterns in each of the systems (TectoMT and PBSMT) when trained on the same datasets, thus directly contrasting those two approaches and showing the main advantages of the hybrid MT system. We further propose a set of rules for improving the synthesis stage in the TectoMT system to eliminate those errors in the future versions of the system.

The next contribution lies in better understanding how different (fine-grained) types of errors influence fluency and adequacy, which relies on the combination of human assessment of fluency and adequacy and detailed error analysis of each group of sentence pairs (depending on the fluency and adequacy scores).

Finally, our results confirm the hypothesis that TectoMT achieves best improvements with adding only bilingual terminology to the small in-domain training dataset, while PBSMT achieves best improvements with adding a combination of a larger out-of-domain corpus and bilingual terminology, regardless of the out-of-domain corpora used (Europarl or scientific news).

¹www.qt Leap.eu

²The corpus is freely available through METASHARE:
<http://metashare.metanet4u.eu/>.

2. Related Work

We divide the related work into two sections. The first section (Section 2.1.) gives a brief introduction to the hybrid MT system used in our study (TectoMT) providing the necessary background for a better understanding of the results and discussion. The second section (Section 2.2.) summarises previous studies on English to Portuguese machine translation task.

2.1. TectoMT – A Hybrid MT System

TectoMT (Žabokrtský et al., 2008) is a modular hybrid MT system which uses two levels of structural representation: a shallow analytical layer (a-layer), and a deep tectogrammatical layer (t-layer). It consists of three phases: analysis, transfer, and synthesis.

The analysis phase consists of two steps. In the first step, standard dependency parsers are used to construct the a-layer. In the second step, the dependency trees of the a-layer (a-trees) are converted into the dependency trees of the t-layer (t-trees) using a set of rule-based modules. The a-layer is represented by a labeled dependency tree (a-tree) which contains all tokens of the sentence as its nodes. Each node contains several types of information: *word form*, *lemma*, *part-of-speech tag* and *morphological information*, and *afun*, a surface dependency label denoting syntactic function (subject, object, predicate, etc.). The t-layer is a deep syntactic/semantic layer represented by a dependency tree (t-tree) which describes the linguistic meaning of the sentence according to the FGD (Functional Generative Description) theory (Sgall et al., 1986). The nodes of the t-tree contain only content words of the sentence and added information (not contained in the sentence) as pro-dropped subject personal pronouns. Each t-node has four attributes: *t-lemma* (“deep lemma” which is in most cases identical to surface lemma), *functor* (a semantic role label based on the FGD theory), *grammateme* (contains information such as person, number, tense, modality, etc.), and *formeme* (morpho-syntactic form information, such as *v:to+inf* for infinitive verbs or *n:into+X* for a prepositional phrase). Auxiliary words are not represented as nodes in the t-tree but rather influence the attributes of the t-nodes.

The transfer phase in TectoMT is performed on the t-layer by translating t-lemmas and conversion of formemes and grammatemes (Bojar and Týnovský, 2009; Žabokrtský et al., 2008). This phase is mostly statistical, based on maximum entropy (MaxEnt) model (Mareček et al., 2010), enhanced with translation dictionaries and a small number of handcrafted rules for handling out-of-vocabulary words.

Finally, the synthesis phase consists of series of small, mostly rule-based modules which have a goal of transforming the translated t-tree into an a-tree and then linearise the a-tree into a plain surface form of the output sentence. These modules are language-specific and take care of word order, agreement (e.g. subject-predicate agreement or noun-adjective agreement), insertion of grammatical words (such as prepositions, articles, particles, etc.), inflections, and capitalisation.

2.2. English to Portuguese Machine Translation

The studies concerning EN-PT MT are very scarce and mostly report on results of the PBSMT systems. The best results (BLEU = 0.55) were obtained on the JRC-Acquis corpus (Koehn et al., 2009), followed by the results obtained using a significantly smaller FAPESP corpus (Aziz and Specia, 2011) of scientific news texts (BLEU = 0.46). The PBSMT systems trained on Europarl corpus (and interpolated with models trained on datasets from the same domain as the test datasets) and tested on domain-specific corpora – TED talks and TAP (Portuguese airline) magazine – achieved significantly lower BLEU scores, 0.20 and 0.19 respectively (Costa et al., 2014). Google Translate achieved better, but still not very high, BLEU scores (0.28 and 0.26, respectively) on the same task (Costa et al., 2014). There have been two studies comparing a hybrid and a PBSMT system for EN-PT language pair (Silva et al., 2015; Štajner et al., 2015), reporting the performances of the two approaches as comparable. None of those studies, however, performs an error analysis to directly compare the errors made by those systems.

3. Machine Translation Experiments

The corpora used in MT experiments, experimental setup and the results of the automatic evaluation of all MT systems are presented in the next three subsections.

3.1. Corpora

We used six corpora in this study:

1. **EP1** – Full Europarl corpus (Koehn, 2005) with English on the source side and Portuguese on the target side (1,960,407 sentence pairs).
2. **EP2** – A smaller portion of Europarl corpus which has the same size as the FAPESP corpus (162,350 sentence pairs).
3. **FAPESP** – A Portuguese-English bilingual collection of the online issue of the scientific news Brazilian magazine “Revista Pesquisa FAPESP”³ (Aziz and Specia, 2011).
4. **IT1** – An in-domain IT corpus with 2,000 sentence pairs (1,000 questions and 1,000 answers) compiled under the QTLeap project (QTLeap corpus, batch 1). This corpus was used as the training dataset (or a part of the training dataset) in our translation experiments.
5. **IT2** – Another in-domain IT corpus, with 1,000 sentence pairs (answers only) compiled under the QTLeap project (QTLeap corpus, batch 2), and comparable with the IT1 corpus. This corpus was used as the test dataset in all our translation experiments.
6. **TERM** – A parallel corpus of IT terminology (unigrams or multiword expressions), which consists of the *Microsoft Terminology Collection*⁴ (13,030 terms)

³<http://revistapesquisa.fapesp.br/>

⁴<https://www.microsoft.com/Language/en-US/Terminology.aspx>

Corpora	Source (EN)	Target (PT)
EP	I am pleased that the Council and the Commission have responded to this concern of Parliament.	Congratulo-me por ver que o Conselho e a Comissão atenderam a esta preocupação do Parlamento.
FAPESP	Do you make use of these procedures at Alellyx? At this stage Simpson had already began to get involved in the project.	Na Alellyx utilizam-se esses procedimentos? Nessa época, o Simpson já estava começando a se envolver no projeto.
IT1	Remove the wireless network and try reconnecting again using valid credentials. How can I delete Skype’s conversation history?	Remova a rede sem fios e ligue novamente com dados válidos. Como posso apagar o histórico de conversas do Skype?
IT2	In the Insert menu, select Table. Select the cells you want to format, then click the right mouse button on one that is selected and choose Format Cells.	No menu inserir selecione Tabela. Selecione as células que pretende formatar, depois clique com o botão direito do rato numa que esteja selecionada e escolha Formatar Células.
TERM	shift key file system command line options	tecla shift sistema de ficheiros opções da linha de comandos

Table 1: Examples from the corpora

Experiment	Training dataset					BLEU score	
	EP1/EP2	FAPESP	IT1	TERM	Total	TectoMT	PBSMT
EP1	1,960,407	/	/	/	1,960,407	19.34	18.99
EP2	162,350	/	/	/	162,350	17.76	17.53
FAPESP	/	162,350	/	/	162,350	19.43	19.20
IT1	/	/	2,000	/	2,000	20.77	21.55
IT1+EP2	162,350	/	2,000	/	164,350	19.77	20.79
IT1+FAPESP	/	162,350	2,000	/	164,350	20.53	*22.64
IT1+TERM	/	/	2,000	14,025	16,025	*21.89	*22.73
IT1+EP2+TERM	162,350	/	2,000	14,025	178,375	*21.04	*22.25
IT1+FAPESP+TERM	/	162,350	2,000	178,375	14,025	*21.63	*23.53

Table 2: Number of sentence pairs (terms and MWE in the case of the TERM corpus) used for the training and the achieved BLEU score. Best results of each system are shown in bold. Results of the systems which significantly (using paired bootstrap resampling (Koehn, 2004)) outperformed all four baselines are shown with an ‘*’.

and a small portion of LibreOffice terminology (995 terms).⁵

Several examples from each corpus are given in Table 1.

3.2. Experimental Setup

We performed a series of MT experiments for English to Portuguese translation using two different approaches: a hybrid MT system (TectoMT) and a PBSMT system in the Moses toolkit. All models were tested on the same dataset (IT2).

For the TectoMT system, we used the English to Portuguese TectoMT system (Silva et al., 2015) developed under the QTLeap project.

For the PBSMT system, we used the GIZA++ implementation of IBM word alignment model 4 (Och and Ney,

2003), and the refinement and phrase-extraction heuristics described further by Koehn et al. (2003). All PBSMT systems were tuned on the IT1 corpus using minimum error rate training (MERT) (Och, 2003). Their language models were built on a 2,121,382 sentence corpus (target side of the full EP+FAPESP corpora) using the KenLM (Heafield, 2011) 5-gram language model. The stack size was limited to 100 hypotheses during decoding.

For each of the two systems (TectoMT and PBSMT), we performed four baseline experiments (using the EP1, EP2, FAPESP, or IT1 corpus as the training dataset) and five experiments which exploited three different strategies for enlarging the small in-domain training dataset (IT1) by adding: (1) a larger out-of-domain dataset (IT1+EP2 or IT1+FAPESP), (2) quasi in-domain data (IT1+TERM), and (3) a combination of both (IT1+EP2+TERM or IT1+FAPESP+TERM).

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

Group		Description	Fluency	Adequacy
TectoMT better rated than PBSMT	1a	Both good (TectoMT = 4, PBSMT = 3)	0	0
	1b	TectoMT good (3 or 4), PBSMT bad (1 or 2)	10	20
	1c	Both bad (TectoMT = 2, PBSMT = 1)	17	10
TectoMT and PBSMT rated the same	2a	Both good (both = 3 or both = 4)	4	11
	2b	Both bad (both = 2 or both = 1)	55	40
PBSMT better rated than TectoMT	3a	Both good (TectoMT = 3, PBSMT = 4)	0	3
	3b	TectoMT bad (1 or 2), PBSMT good (3 or 4)	8	12
	3c	Both bad (TectoMT = 1, PBSMT = 2)	6	4

Table 3: Number of sentence pairs analysed in each group.

Code	Error	Example
SENS	Incorrect sense	“phones” (with the meaning “headphones”) translated as “telemóveis” (mobile phones) instead of correct “auscultadores”
WRTR	Wrong translation	“Unidade de Google” instead of “Google Drive”
UNTR	Left untranslated	“format”, “connect”...
ADDW	Added words	“Sign into the YouTube account.” translated as “Num ISP na conta no youtube.”
MISW	Missing words	“double-click” translated as “clique” (click)
CAP	Capitalisation	that says, “Starred” → que diz estrela
COM	Missing commas	that says, “Starred” → que diz estrela
QUO	Missing quotation marks	that says, “Starred” → que diz estrela
REL	Missing relative pronoun	“escolha a língua quer” instead of “escolha a língua que quer”
DEF	Missing definite articles	“em seu teclado” instead of “em o seu teclado”
WRA	Wrong article	“a pesquisa” instead of “uma pesquisa”
VMD	Wrong verb mood	“seleciona voltar a” instead of “selecione voltar a”
WVM	Wrong verb morphology	wrong tense, person or number
INP	Inserted preposition	“Alt key” → “tecla de alt”
MIP	Missing preposition	“carregar botão” instead of “carregar no botão”
WRP	Wrong preposition	“ir em de codificação” instead of “ir a codificação”
GCA	Gender concordance (article)	“escolha a canal” instead of “escolha o canal”
GCP	Gender concordance (pronoun)	“junto de sua perfil” instead of “junto ao seu perfil”
NUC	Number concordance	“botões que diz ” instead of “dizem”
WWO	Wrong word order	Open Network Broadcast → “Difusão de rede de abrir ”, instead of “Abrir Difusão de Rede”

Table 4: Classification of adequacy and fluency errors.

3.3. Results of the Translation Experiments

The training datasets of each of the nine experiments (performed for both MT systems) are presented in Table 2, together with the results of the automatic evaluation of both systems (TectoMT and PBSMT) in terms of the BLEU score (Papineni et al., 2002).

Our results indicated that addition of a larger out-of-domain corpus only improves the performance of the PBSMT system, while it deteriorates the performance of the TectoMT system. The TectoMT achieves best improvements with addition of the bilingual terminology only.

4. Hybrid vs Statistical MT System

It is known that BLEU scores cannot be used for comparing two MT systems with different architectures (TectoMT and PBSMT in our case), but only for comparing different versions of the same system (either PBSMT or TectoMT in our case). Therefore, we focused on the results of the IT1+TERM experiments – which led to the best BLEU

score for the TectoMT system, and the second best BLEU score for the PBSMT system (Table 2) – and performed human evaluation and in-depth error analysis in order to compare the performances of the TectoMT and PBSMT systems on the same datasets.

4.1. Human Evaluation

We randomly selected 100 original sentences from the test set and asked two linguists, native speakers of Portuguese, to rate their corresponding outputs produced by the TectoMT and PBSMT systems trained on IT1+TERM dataset (a total of 200 output sentences) in terms of their Fluency and Adequacy on a 1–4 scale (where 1 denotes very bad, and 4 very good output). The output sentences of the TectoMT and PBSMT systems were presented to the annotators in random order. The annotators favoured the output of the TectoMT system (Fluency = 1.78, Adequacy = 2.28) over the output of the PBSMT system (Fluency = 1.74, Ad-

Error	1b		1c		2a		2b		3a		3b		3c	
	Tecto	PBSMT	Tecto	PBSMT	Tecto	PBSMT	Tecto	PBSMT	Tecto	PBSMT	Tecto	PBSMT	Tecto	PBSMT
SENS	1	4	2	1	1	1	8	4	0	0	0	0	0	0
WRTR	5	7	3	4	2	3	14	20	1	1	3	2	2	1
UNTR	2	16	6	6	2	5	33	44	1	1	5	9	2	4
ADDW	0	1	0	4	0	4	0	10	0	0	0	1	0	1
MISW	4	3	0	2	1	2	1	8	0	0	0	2	0	1

Table 5: Distribution of adequacy errors.

Error	1b		1c		2a		2b		3b		3c	
	Tecto	PBSMT	Tecto	PBSMT	Tecto	PBSMT	Tecto	PBSMT	Tecto	PBSMT	Tecto	PBSMT
CAP	9	3	14	9	6	2	50	27	5	1	5	5
COM	7	2	11	1	1	0	32	10	3	2	0	0
QUO	2	0	1	1	0	0	0	0	0	0	0	0
REL	0	1	2	3	0	0	5	6	0	0	1	0
DEF	3	1	9	0	2	0	23	26	0	0	1	0
WRA	0	2	0	1	0	1	0	5	0	1	0	0
VMD	2	8	4	10	0	2	40	13	0	1	1	2
WVM	4	0	3	1	2	1	5	1	1	0	0	0
INP	3	9	4	8	1	1	26	21	1	3	2	4
MIP	3	3	7	4	2	0	27	35	5	0	3	0
WRP	0	3	4	1	0	0	3	8	1	0	0	2
GCA	1	3	0	4	0	1	6	21	0	0	1	0
GCP	0	0	0	1	0	0	0	8	0	1	0	0
NUC	0	1	4	3	0	0	14	8	0	0	1	0
WWO	0	9	6	20	0	0	39	45	6	0	5	4

Table 6: Distribution of fluency errors.

equacy = 2.24).⁶

In order to gain more insights into problems of each approach and assess the possibility to overcome them in the future, we divided those 100 sentence pairs in eight groups (Table 3) according to the scores obtained by the annotators. The sentences which were not classified in the same group by both annotators were additionally annotated by a third annotator (again a linguist and native speaker of Portuguese). The sentences were finally assigned to the group chosen by the majority of annotators.

4.2. Error Analysis

We first analysed the sentences in each group in more details searching for the repetitive error patterns and made a classification of most frequent adequacy and fluency errors (Table 4). Next, we quantified each error type, for fluency and for adequacy separately (Tables 5 and 6).

4.2.1. Adequacy Errors

In terms of adequacy, it seems that the main difference in the performance of the TectoMT system and the PBSMT system lies in the number of untranslated words – UNTR (Table 5). In the cases where the output of both systems was rated as equally good (group 2a), the PBSMT system was found to have a slightly higher number of untranslated words (UNTR) and additional words (ADDW).

In the cases where the output of both systems was rated as equally bad (group 2b), the output of the PBSMT system had a significantly higher number of wrongly translated words (WRTR), untranslated words (UNTR), added words (ADDW), and missing words (MISW), while the output of the TectoMT system had two times higher number of word sense errors (SENS).

4.2.2. Fluency Errors

Overall, it seems that the hybrid system more often fails in punctual cases like capitalisation (CAP), missing punctuation (COM, QUO), while the PBSMT system leads to more reordering errors (WWO) by inverting the verb+object order. In those cases where the output of both systems was rated as bad, but the output of the hybrid system was rated as slightly less bad (group 1c), it seems that human evaluators put more weight on the errors in word order (WWO) and wrong verb mood (WVM) made by the PBSMT system than on the capitalisation errors (CAP), punctuation errors (COM), and errors in prepositions (MIP and WRP) made by the hybrid system.

5. Discussion

As this error analysis was performed with the goal of discovering the shortcomings of the current version of the hybrid system in order to improve the system in its next version, we further focused on the most frequent fluency errors made by the TectoMT system (capitalisation, wrong verb

⁶The difference in adequacy scores was significant at a 0.05 level.

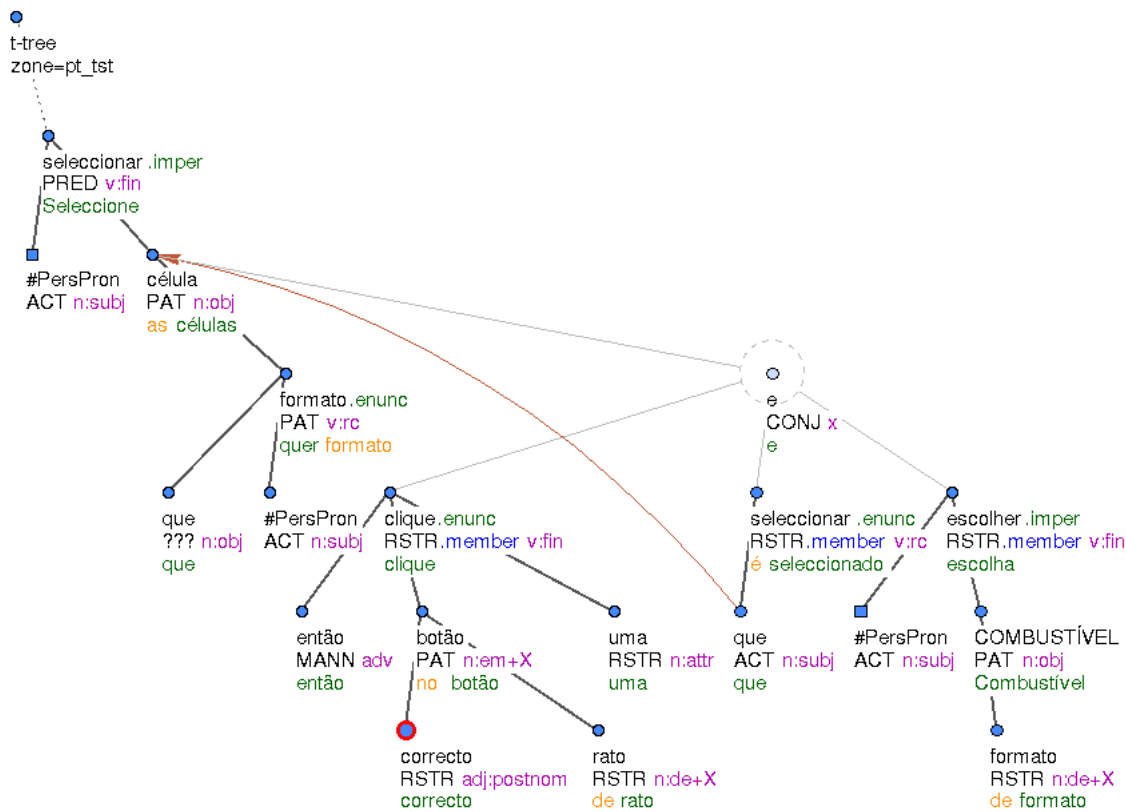


Figure 2: The target t-tree with a wrong word order (the node *correcto* should be placed after the node *botão*).

mood, and wrong word order) in order to search for their origin (whether they are the result of mistakes made during the analysis, transfer, or synthesis phase).

5.1. Capitalisation errors (CAP)

By inspecting the a-trees and t-trees of the sentences containing capitalisation errors, e.g. Examples (1) – (3), we noticed that this type of errors originates from the transfer phase. In the current version of the TectoMT system, the capitalisation is maintained only for the tokens at the beginning of a sentence.

- (1) Source: You have to go to **Format** > Uppercase.
- (2) Translation: Deverá ir a **formato** > Uppercase.
- (3) Reference: Deverá ir a **Formatar** > Maiúsculas.

Our examination revealed that the correctly capitalised target words (within the sentence) appear only in the cases of the wrong transfer of the original source lemmas (when the transfer is done by simply cloning the source lemma). This error is fairly easy to avoid. A possible solution would be to implement a block which forces all the nodes of the source tree with a capitalised letter to be capitalised in the corresponding node in the target tree.

5.2. Wrong verb mood (VMD)

This type of errors, e.g. Examples (4) – (6), appear due to mistakes in the analysis and transfer phases which lead to a

target tree with a missing #perspron (personal and possessive pronouns) node.

- (4) Source: In the Google Drive site **go** to the tab Recent.
- (5) Translation: No site de condução de Google **vai** ao separador recente.
- (6) Reference: No site do Google drive **vá** até ao separador Recente.

This prevents the identification of the grammatical values of the subject and the correct agreement between subject and verb.

5.3. Wrong word order (WVO)

In most cases, the wrong word order in the output of the TectoMT system, e.g. Examples (7) – (9), originates from the wrong analysis phase which leads to an incorrect part-of-speech tagging or from the missing block (currently being implemented) for handling the :postnom formeme.

- (7) Source: Click the **right** mouse button.
- (8) Translation: Clique no **correcto** botão de rato.
- (9) Reference: Clique no botão **direito** do rato.

In this case, the node **correcto** with the formeme :postnom should be reordered and placed after de **botão** node (Figure 2). Note that in this example the word **correcto** is also badly translated (the lemma **correct** should have been translated into **direito**).

6. Conclusions

We addressed the problem of domain-specific MT for a language pair for which only very small amount of training data is available. The results of our experiments showed that, in such case, performance of a hybrid MT system significantly improves with addition of a small amount of in-domain bilingual terminology (approx. 14,000 entries) to the very small in-domain training corpus (2,000 sentence pairs). They also indicated that addition of a larger out-of-domain corpus only improves the performance of the PBSMT system and not the hybrid MT system.

The extensive in-depth error analysis of the output of the hybrid and PBSMT systems trained on the same dataset (small in-domain corpus with added bilingual terminology), directly compared and analysed sentence by sentence, reported the output of the hybrid system as better than the output of the PBSMT system. The hybrid system mostly failed in capitalisation and retaining punctuation marks, while the PBSMT system had a greater number of reordering errors, untranslated tokens and missing words.

Acknowledgements

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

7. References

- Aziz, W. and Specia, L. (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, October.
- Bojar, O. and Týnovský, M. (2009). Evaluation of Tree Transfer System. Euromatrix project (statistical and hybrid machine translation between all european languages), deliverable 3.4, Charles University in Prague.
- Costa, A., Luís, T., and Coheur, L. (2014). Translation Errors from English to Portuguese: an Annotated Corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 1231–1234.
- Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July.
- Koehn, P., Och, F. J., and Marcu, D. (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.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., 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 (ACL)*. ACL.
- Koehn, P., Birch, A., and Steinberger, R. (2009). 462 Machine Translation Systems for Europe. In *Proceedings of the MT Summit XII*, pages 65–72.
- Koehn, P. (2004). Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395.
- 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 Metrics-MATR*, pages 201–206.
- Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.
- Och, F. (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.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.
- Sgall, P., Hajicová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Springer Science & Business Media.
- Silva, J., Rodrigues, J., Gomes, L., and Branco, A. (2015). Bootstrapping a hybrid deep MT system. In *Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, pages 1–5. ACL.
- Štajner, S., Rodrigues, J., Gomes, L., and Branco, A. (2015). Machine Translation for Multilingual Troubleshooting in the IT Domain: A Comparison of Different Strategies. In *Proceedings of the Deep Machine Translation Workshop (DMTW)*, pages 106–115.
- Žabokrtský, Z., Ptáček, J., and Pajas, P. (2008). TectoMT: Highly modular MT system with tectogrammatcs used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170.