Domain-Specific Hybrid Machine Translation from English to Portuguese

João Rodrigues(✉), Luís Gomes, Steven Neale, Andreia Querido,
Nuno Rendeiro, Sanja Štajner, João Silva, and António Branco

Department of Informatics, Faculty of Sciences, University of Lisbon,
Lisbon, Portugal
{joao.rodrigues,luis.gomes,steven.neale,andreia.querido,
nuno.rendeiro,sanja.stajner,jsilva,antonio.branco}@di.fc.ul.pt

Abstract. Machine translation (MT) from English to Portuguese has not typically received much attention in existing research. In this paper, we focus on MT from English to Portuguese for the specific domain of information technology (IT), building a small in-domain parallel corpus to address the lack of IT-specific and publicly-available parallel corpora and then adapted an existing hybrid MT system to the new language pair (English to Portuguese). We further improved the initial version of the EN-PT hybrid system by adding various modules to address the most frequently occurring errors in the initial system. In order to assess the improvements achieved by each of these dedicated modules, we compared all versions of our MT system automatically. In addition, we conduct and report on a detailed error analysis of the initial and final versions of our system.

Keywords: Hybrid machine translation · TectoMT · Lexical semantics · IT domain · Portuguese

1 Introduction

Phrase-based statistical machine translation (PBSMT) models are generally considered to be the state-of-the-art for any language pair and domain for which large enough parallel corpora exist. For many language pairs, however, training corpora of sufficient size are limited to only a few domains. For English to Portuguese machine translation (MT), for example, large parallel corpora are available for just two particular domains – legal documents (the JRC-Acquis corpus [10]), and parliamentary discussions (the Europarl corpus [9]).

In this paper, we address the problem of English to Portuguese machine translation for the IT domain, focusing on the conversations of real users with technical support. In this scenario, users first ask a question in Portuguese which is machine translated into English, and then the answer is searched for in an English database, automatically translated back to Portuguese and presented back to the user. As there are no publicly available parallel corpora for the IT

J. Silva et al. (Eds.): PROPOR 2016, LNAI 9727, pp. 50–61, 2016.
DOI: 10.1007/978-3-319-41552-9_5

domain, we compiled a small in-domain corpus, the QTLeap Corpus[1], consisting of 4,000 utterance pairs (2,000 questions and 2,000 answers) from the IT domain [8], under the QTLeap project[2].

Following the widespread assumption that rule-based and hybrid MT systems give better results for domains and language pairs for which limited parallel data is available – a result of their capacity to make generalisations and thus better overcome data sparsity – we opted for building a hybrid MT system. Our starting point is the TectoMT system [20] which we have adapted from English-Czech to the English-Portuguese language pair. Guided by a detailed human evaluation and error analysis of our initial English-Portuguese TectoMT system, we then added four new modules to handle the most frequently occurring mistakes produced by the initial system.

2 Related Work

Our summary of related work is divided into two sections – firstly, we summarize previous studies on MT from English to Portuguese (Sect. 2.1), and secondly we introduce the hybrid MT system (TectoMT) from which our system for English-Portuguese was built (Sect. 2.2).

2.1 English to Portuguese Machine Translation

Previous studies of MT from English to Portuguese are very scarce, with most reporting on the results of phrase-based statistical MT (PBSMT) systems. Examples of this include results reported on the JRC-Acquis corpus [10] (BLEU = 55) and on the substantially smaller FAPESP corpus of scientific news texts [2] (BLEU = 46). Scores for domain-specific PBSMT systems [6] are substantially lower – trained on Europarl and tested on TED talks and the magazine of Portuguese airline *TAP*, they report BLEU scores 20 and 19 respectively. Scores achieved using Google Translate were better (although still low) for the same task – 28 and 26, respectively.

Recently, two studies were released that report the performance of a baseline hybrid MT system from English to Portuguese for the IT domain compared with a baseline PBSMT system on the same domain [17,19]. In this paper we go one step further, enhancing the baseline TectoMT system from English to Portuguese with specific modules dedicated to reducing the recurrent errors in the baseline system. Furthermore, an extensive human evaluation is performed and reported.

2.2 TectoMT - A Hybrid Machine Translation System

TectoMT is a hybrid system, incorporating elements of statistical and rule-based MT into a modular framework that can be adapted to include various NLP

[1] Available from: http://www.meta-share.org/.

[2] http://www.qtleap.eu.

tasks in a single pipeline [20]. The system handles translation over three phases: analysis (of the source language), transfer (of information from source to target language), and synthesis (into the target language). The analysis and synthesis phases are primarily modular – allowing for independent, statistical and/or rule-based NLP tools and processes to be wrapped as 'blocks' and combined to form scenarios (combinations of blocks) specific to required tasks – while the transfer phase that links the two is primarily statistical.

TectoMT is based on two levels of structural representation – a shallow analytical layer (a-layer) and a deep tectogrammatical layer (t-layer) that describes the linguistic meaning of a sentence according to functional generative description (FGD) theory [16]. The translation process goes thorough these two levels of representation, both of which represent input sentences as labeled dependency trees of varying complexity:

- *a-trees*, with each token in the sentence being represented as an *a-node* constructed from:
  - original word forms
  - lemmas
  - part-of-speech (POS) tags
  - morphological information
- *t-trees*, with each token in the sentence being represented as a *t-node* constructed from:
  - deep lemmas (usually identical to the surface lemma)
  - functors (FGD theory-based semantic role labels)
  - grammatemes (person, number, tense, modality etc.)
  - formemes (morphosyntactic information such as `v:to+inf` for infinitive verbs or `n:into+X` for a prepositional phrase).

In a typical example, the analysis phase will involve input sentences being parsed and processed by different scenarios of blocks to construct a-layer trees, which are then propagated upwards to construct t-layer trees. The transfer phase then carries on, whereby t-lemmas (lemmas from the t-layer) are translated and formemes and grammatemes converted from source to target language [3,20] – this phase is mostly statistical, and based on maximum entropy (MaxEnt) models enriched with specific translation dictionaries and a small number of handcrafted rules for handling out-of-vocabulary words. Finally, primarily rule-based scenarios in the synthesis perform the reverse of the analysis phase, transforming translated t-trees into a-trees and then linearizing these into output sentences in surface form. For Portuguese, many of the modules in the analysis and synthesis phases are language-specific and handle problems such as word order, agreement (e.g. subject-predicate agreement or noun-adjective agreement), insertion of grammatical words (such as prepositions, articles, particles, etc.), inflections, and capitalization.

3 English-Portuguese TectoMT Systems

In this section, we describe our initial, baseline EN-PT TectoMT system (Sect. 3.1), its improved version (Sect. 3.2), and four modifications to the improved version (Sects. 3.3, 3.4, 3.5 and 3.6) that each focus on addressing the different problems highlighted in a detailed human evaluation of the initial system.

3.1 First EN-PT TectoMT System (System 1)

Building on the original English-Czech TectoMT system to produce our initial English-Portuguese version was primarily focused on adapting the rule-based modules used in the synthesis phase scenario. In the analysis phase, the conversion of source sentences in English to a-trees was already handled by various blocks of NLP tools that perform sentence splitting, tokenization, morphological tagging and dependency parsing. We followed the existing English-Czech annotation pipeline developed for the CzEng 1.0 parallel corpus [4] – using the Morče tagger [18] and the Maximum Spanning tree parser [11] trained on the CoNLL-2007 conversion of the Penn Treebank [13] – and kept the same rule-based blocks for creating a-trees and then t-trees as were used in the original English-Czech version of TectoMT [20].

When translating the English t-trees into Portuguese t-trees in the transfer phase, the transfer of t-lemmas and formemes is handled simultaneously by producing an n-best list of translation variants using t-lemma and formeme translation models (TM). For each t-lemma or formeme for a given source (English) t-tree, the translation model estimates the probability of different translation variants given the source t-lemma or formeme and any additional context. This probability is calculated as a linear combination of:

- *Discriminative Translation Models* – a prediction based on features extracted from the source tree using a MaxEnt model.
- *Dictionary Translation Models* – a dictionary of possible translations with relative frequencies (these models, which do not take contextual features into account, are called *static* models in TectoMT's source code).

After English t-trees have been translated into Portuguese t-trees the synthesis phase begins, for which Portuguese-specific rule-based blocks were written (in Perl) to handle tasks such as word ordering, insertion of negations, prepositions, conjunctions, agreement, formation of compound verbs, and so on. Where possible, existing tools for Portuguese [5] have been used to construct the scenario for synthesis, owing to their greater level of accuracy over the tools available in the original TectoMT system, with new rule-based blocks being created in order to integrate these tools into the TectoMT pipeline [15].

This initial, baseline version of the TectoMT system for English-Portuguese was trained on the whole Europarl corpus [9]. The synthesis scenario was improved iteratively, controlling for both the MT output (as BLEU) and a human error analysis of 1,000 sentences from a small in-domain corpus in each step. This set of 1,000 sentences was obtained from the same corpus as the training and test sets, without any overlapping between them. After each iteration – usually involving the addition of new blocks in the synthesis scenario – the MT output (as BLEU) was checked and a human error analysis performed by two linguistic experts. These experts – both native speakers of Portuguese – analyzed the most frequently missing n-grams (up to 3-grams) and the t-trees at the starting point of the synthesis phase, using their analysis to suggest rules for enforcing better synthesis – the transformation of t-trees to output sentences in Portuguese.

3.2 Second EN-PT TectoMT System (System 2)

The second version of the EN-PT TectoMT system saw the introduction of some improvements over the initial, baseline system. Building on the first version of the system, tokenization, lemmatization, morphological analysis, part-of-speech (PoS) tagging and dependency parsing were improved. For this second version of the EN-PT TectoMT system, improvements were also made to the analysis phase firstly by adding missing lemmas for use with the POS-tagger and by adding extra rules for tokenization.

Improvements were made in the synthesis phase of the second version of the EN-PT TectoMT system, namely by adding missing lemmas to the LX-Inflector component of the LX-Suite in order to handle nominal expressions and to the LX-Conjugator component in order to handle verbal expressions. An additional block for handling the insertion of quotation marks in quoted expressions was also added to the synthesis scenario. Next follows an example of the resulting translation using system 1 (a) and system 2 (b) with this block:

(a) *No separador de Slides em [...]*
(b) *No separador de 'Slides' em [...]*

Over the next few subsections, we describe the implementation of additional modules built to improve the second version of the EN-PT TectoMT system to address various problems discovered in the human evaluation of error analysis on the first system.

3.3 EN-PT TectoMT with Word Sense Disambiguation (System 2 + WSD)

The transfer phase in TectoMT is based on lemma-to-lemma translation models, but lemmas themselves are often ambiguous, and can be represented by multiple meanings. We thus experimented with using additional information from source language (English) word sense disambiguation (WSD) – the computational task of determining the correct meaning of a word in a particular context – in the

TectoMT transfer. For each a-layer node created in the analysis phase, we add additional contextual features containing word sense information from both the current node and its parent node to the Discriminative (MaxEnt-based) translation model. This work has been described in greater detail in previous work [12]. Next follows an example of the resulting translation using system 1 (a) and system 2 (b) with the WSD embedded information:

(a) *No domínio de notificação de Windows há o ícone de Panda.*
(b) *Na **área** de notificação de Windows há o ícone de Panda.*

English word senses were obtained using the UKB system [1], a collection of tools and algorithms for performing graph-based WSD over a pre-existing knowledge base. For a given word, UKB is able to query a graph-based representation of WordNet [7] and return the appropriate synset identifier that represents the meaning of the given word, using its surrounding words as context. In addition to synset identifiers, we also provide supersenses to the translation model as features – supersenses are the 45 semantic files by which synset identifiers are organized in WordNet, allowing senses to be generalized across semantic classes like PEOPLE, GROUP or ARTIFACT.

3.4 EN-PT TectoMT with Hidden Entities (System 2 + HideIT)

The error analysis of the initial, baseline version of the EN-PT TectoMT system suggested that a substantial number of translation errors originate from the incorrect handling of named entities (NEs), especially those that are domain-specific (IT) and thus cannot be successfully captured by named entity recognition and classification (NERC) tools. To address this, we experimented with the implementation of a rule-based component called *HideIT* to account for domain-specific entities that do not require translation such as URLs, shell commands, and code snippets. Next follows an example of the resulting translation using system 1 (a) and system 2 (b) with the HideIT block:

(a) *Envie um correio qualidade@pcmedic. PT.*
(b) *Envie um correio a **qualidade@pcmedic.pt.***

The HideIT component consists of two blocks. The first block is applied at the very start of the translation pipeline – just after the tokenization of the source text and before any meaningful linguistic processing takes place – and attempts to recognize such entities using manually gathered heuristics from 2,000 sentences from the in-domain development corpus. Recognized entities are then replaced with an appropriate placeholder (e.g. `xxxCMDxxx` or `xxxURLxxx` for shell command and URL, respectively), while the original values are stored as metadata. The second block is applied at the very end of the translation pipeline, and extracts the values that were recognized earlier and stored as metadata and swaps them with the placeholders that were introduced by the first block to hide the entities from the core processing components of the translation pipeline.

3.5 EN-PT TectoMT with Added Gazetteer (System 2 + Gazetteer)

We also focused on trying to obtain correct translations and localizations of NEs in the IT domain – such as menu items, button names, sequences and messages – that are expected to appear in a fixed inflectional form. The fact that such NEs are fixed allowed us to match expressions from a specialized lexicon (gazetteer) in the source text and replace them with their equivalent expressions in the target language. The English-Portuguese gazetteer was collected from four sources: localization files of VLC,[3] LibreOffice,[4] KDE,[5] and IT-related Wikipedia articles.

Following the tokenization of text in the analysis phase, expressions in the gazetteer are searched for in the source sentence. Matched expressions – which can span several neighbouring tokens – are then replaced by a single-word placeholder. Then, in the transfer phase, these placeholders are replaced in the t-trees by the corresponding expressions stored in the gazeteer from the target language. Note that this step is performed before translating any of the other words of the source sentence.

3.6 EN-PT TectoMT with Domain Adaptation (System 2 + DomAdapt)

The error analysis of the first version of the EN-PT TectoMT system also highlighted many incorrectly translated domain-specific words and phrases that still could not be addressed using the new HideIT and Gazetteer implementations. To address this problem, we experimented with domain adaptation during the transfer phase by interpolating translation models from a general domain (Europarl) and the IT domain (2,000 utterances from the QTLeap corpus described in the Introduction to the paper enriched with parallel terminology from both the Microsoft Terminology Collection,[6] and LibreOffice localization data[7]). This interpolation helps to account for some of the errors in the output of the initial system originating from a lack of in-domain training data.

The interpolation was not applied only to the lexical transfer (of lemmas, as in the experiments with HideIT and Gazetteer), but also to the transfer of formemes. It had been noticed that the IT domain formeme Translation Models (TMs) had a different distribution of probabilities to the general domain (Europarl) TMs, and so it was ventured that the interpolation of formeme TMs could also be beneficial. The EN-PT TectoMT system trains four standard TMs from parallel training data – a Dictionary formeme TM, a Discriminative formeme TM, a Dictionary t-lemma TM, and a Discriminative t-lemma TM. For the interpolation of these

[3] http://downloads.videolan.org/pub/videolan/vlc/2.1.5/vlc-2.1.5.tar.xz.

[4] http://download.documentfoundation.org/libreoffice/src/4.4.0/libreoffice-translations-4.4.0.3.tar.xz.

[5] svn://anonsvn.kde.org/home/kde/branches/stable/l10n-kde4/pt/messages.

[6] Available from: http://www.microsoft.com/Language/en-US/Terminology.aspx.

[7] Available from: https://www.libreoffice.org/community/localization/.

TMs, each of the four TMs was assigned an interpolation weight (1.0 for the Dictionary formeme and Discriminative t-lemma TMs, and 0.5 for the Discriminative formeme and Dictionary t-lemma TMs). Next follows an example of the resulting translation using system 1 (a) and system 2 (b) with domain adaptation:

(a) *No menu de desempenhar escolhe voltar a celeridade normal.*
(b) *No menu de* ***Reproduzir*** *escolhe voltar a* ***velocidade*** *normal.*

4 Evaluation

Our evaluation of all of the systems described in the previous Section has consisted of two methods – an automatic evaluation of MT output and a manual evaluation of error analysis performed by linguistic experts.

4.1 Automatic Evaluation

We performed an automatic evaluation of MT output (as BLEU [14]) for all of the described EN-PT TectoMT systems: System 1, System 2, System 2 + WSD, System 2 + HideIT, System 2 + Gazetteer, System 2 + DomAdapt, and System 2+ (System 2 enriched with all four additional modules – WSD, HideIT, Gazetteer, and DomAdapt). The results are presented in Table 1.

Table 1. BLEU scores for all systems.

Experiment	BLEU	BLEU-BLEU(System 2)
System 1	19.34	−0.48
System 2	19.82	0.00
System 2 + WSD	20.07	+0.25
System 2 + HideIT	20.16	+0.34
System 2 + Gazetteer	20.76	+0.94
System 2 + DomAdapt	21.80	+1.98
System 2+	22.42	+2.60

Table 2 shows the number of errors found by the linguists in each system multiplied by four (an estimate of the likely number of errors that would occur in 100 sentences, this was due to the interest in the “density” of each error type rather than the total number, notice that the values are a mean value of errors found by two annotators), as well as the absolute difference and the relative difference between number of errors found in System 2+ and System 1.

The results in Table 1 show that the largest improvements to the system are achieved by making use of Gazetteers (specialized lexicons) and by interpolating general and IT-domain TMs, while the addition of WSD and HideIT modules also yield slight improves of the system. Changes in the analysis and synthesis phases from System 1 to System 2 also led to substantive improvements. The full system (System 2+) – which incorporates all of the previously described improvements and additional modules – achieves good results (BLEU = 22.42).

4.2 Error Analysis

To gain better insight into the translation quality achieved by System 1 and by System 2+, we asked two linguistic experts (both native speakers of Portuguese) to analyze the specific errors made by each system on a subset of 25 sentences. The errors they discovered were then classified according to the Multidimensional Quality Metrics (MQM) framework[8] (with some slight modifications):

1. Accuracy
    - (a) Addition
    - (b) Mistranslation
    - (c) Omission
    - (d) Overly literal
    - (e) Untranslated
2. Fluency
    - (a) Grammatical register
    - (b) Spelling
    - (c) Typography
    - (d) Grammar
        - i. Word form
            - A. Part of speech
            - B. Agreement
            - C. Tense/aspect/mood
        - ii Word order
        - iii Function words
            - A. Extraneous
            - B. Incorrect
            - C. Missing
    - (e) Unintelligible
3. Locale convention
4. Terminology

The results shown in the Table 2 demonstrate that there are less Accuracy errors (−43 %) in the output of System 2+, particularly errors classified as overly literal translation or mistranslation. In terms of Fluency, the output of System 2+ showed fewer spelling errors, agreement errors, word order problems and incorrect translations of function words than were present in the output of System 1. However, the number of missing function words and tense, aspect and mood errors increased from System 1 to System 2+. Taken as a whole and in context, these results suggest that translation of terminology in particular has indeed been improved in System 2+.

[8] http://www.qt21.eu/launchpad/content/multidimensional-quality-metrics.

Table 2. Number of errors in each system (System 1 and System 2+), and their relative difference.

Error type	System 1	System 2+	%
Accuracy	0	0	0 %
-Addition	8	6	−25 %
-Mistranslation	178	82	−54 %
-Omission	46	36	−22 %
-Overly literal	30	16	−47 %
-Untranslated	22	22	0 %
Accuracy subtotal	284	162	−43 %
Fluency	2	2	0 %
-Grammatical register	0	0	0 %
-Spelling	48	40	−17 %
-Typography	46	54	17 %
-Grammar	0	0	0 %
–Word form	0	0	0 %
—Part of speech	34	34	0 %
—Agreement	56	52	−7 %
—Tense/aspect/mood	56	100	79 %
–*Word form subtotal*	146	186	27 %
–Word order	74	66	−11 %
–Function words	0	0	0 %
—Extraneous	112	110	−2 %
—Incorrect	52	32	−38 %
—Missing	210	244	16 %
–*Function words subtotal*	374	386	3 %
-Unintelligible	0	0	0 %
Fluency subtotal	690	734	6 %
Locale convention	0	0	0 %
Terminology	12	10	−17 %

5 Conclusions

Previous research addressing MT from English to Portuguese has been scarce thus far, with the few studies that do describe this language pair generally focusing on phrase-based SMT systems. In this paper, we have described our implementation of an MT pipeline from English to Portuguese for a specific domain (IT), also creating a small, in-domain corpus to account for the lack of publicly-available parallel corpora for the domain in question. Part of this corpus was used for development of our hybrid EN-PT MT system, and the other part used for testing.

We first built an initial, baseline EN-PT hybrid MT system by adapting the existing hybrid TectoMT system from English-Czech to English-Portuguese. After performing an initial error analysis, we further improved the analysis and synthesis phases of the system and added four new modules to address most common mistakes of the initial system. Automatic evaluation of the output of the revised system using each of the newly-created module showed that each of them helps to improve the overall performance of the system, suggesting that the addition of a gazetteer (specialized lexicon) and the interpolation of general and domain-specific translation models as the most promising strategies for improving MT output. Finally, a detailed human error analysis of the initial and the final systems confirmed that the additional modules and improvements of analysis and synthesis phases implemented in the second version of the EN-PT hybrid MT system do contribute to improved MT output.

Acknowledgements. The results reported in this paper were partially supported by the Portuguese Government's P2020 program under the grant 08/SI/2015/3279: ASSET-Intelligent Assistance for Everyone Everywhere, and by the EC's FP7 program under the grant number 610516: QTLeap-Quality Translation by Deep Language Engineering Approaches.

References

1. Agirre, E., Soroa, A.: Personalizing PageRank for word sense disambiguation. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2009, pp. 33–41. Association for Computational Linguistics, Athens (2009)
2. Aziz, W., Specia, L.: 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 2011
3. Bojar, O., Týnovský, M.: Evaluation of tree transfer system. Technical report, Charles University in Prague (2009)
4. Bojar, O., Žabokrtský, Z., Dušek, O., Galuščáková, P., Majliš, M., Mareček, D., Maršík, J., Novák, M., Popel, M., Tamchyna, A.: The joy of parallelism with CzEng 1.0. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), pp. 3921–3928 (2012)
5. Branco, A., Silva, J.R.: 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) (2006)
6. Costa, A., Luís, T., Coheur, L.: Translation errors from english to portuguese: an annotated corpus. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC) (2014)
7. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
8. Gaudio, R.D., Burchardt, A., Branco, A.: Evaluating machine translation in a usage scenario. In: Proceedings of LREC (2016). (to appear in print)
9. Koehn, P.: Europarl: a parallel corpus for statistical machine translation. In: Proceedings of the Tenth Machine Translation Summit, pp. 79–86 (2005)

10. Koehn, P., Birch, A., Steinberger, R.: 462 machine translation systems for Europe. In: Proceedings of the MT Summit XII (2009)
11. McDonald, R., Pereira, F., Ribarov, K., Hajič, J.: 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), pp. 523–530 (2005)
12. Neale, S., Gomes, L., Branco, A.: First steps in using word senses as contextual features in maxent models for machine translation. In: Proceedings of the First Workshop on Deep Machine Translation, DMTW-2015, pp. 64–72 (2015)
13. Nilsson, J., Riedel, S., Yuret, D.: The CoNLL 2007 shared task on dependency parsing. In: Proceedings of the CoNLL shared task session of EMNLP-CoNLL, pp. 915–932 (2007)
14. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of ACL (2002)
15. Rodrigues, J., Rendeiro, N., Querido, A., Štajner, S., Branco, A.: Bootstrapping a hybrid MT system to a new language pair. In: Proceedings of LREC (2016). (to appear in print)
16. Sgall, P., Hajicová, E., Panevová, J.: The Meaning of the Sentence in its Semantic and Pragmatic Aspects. Springer Science & Business Media (1986)
17. Silva, J., Rodrigues, J., Gomes, L., Branco, A.: Bootstrapping a hybrid deep MT system. In: Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra), pp. 1–5. ACL (2015)
18. Spoustová, D., Hajič, J., Votrubec, J., Krbec, P., Květoň, P.: 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, pp. 67–74 (2007)
19. Štajner, S., Rodrigues, J., Gomes, L., Branco, A.: Machine translation for multilingual troubleshooting in the IT domain: a comparison of different strategies. In: Proceedings of the Deep Machine Translation Workshop (DMTW), pp. 106–115 (2015)
20. Žabokrtský, Z., Ptáček, J., Pajas, P.: TectoMT: highly modular MT system with tectogrammatics used as transfer layer. In: Proceedings of the Third Workshop on Statistical Machine Translation, pp. 167–170 (2008)