Towards Deeper MT: Parallel Treebanks, Entity Linking, and Linguistic Evaluation

Ankit Srivastava, Vivien Macketanz,
Aljoscha Burchardt, and Eleftherios Avramidis

German Research Center for Artificial Intelligence (DFKI Berlin),
Language Technology Lab, Alt-Moabit 91c, 10559 Berlin, Germany
firstName.lastName@dfki.de
http://www.dfki.de/web

Abstract. In this paper we investigate techniques to enrich Statistical Machine Translation (SMT) with automatic deep linguistic tools and evaluate with a deeper manual linguistic analysis. Using English–German IT-domain translation as a case-study, we exploit parallel treebanks for syntax-aware phrase extraction and interface with Linked Open Data (LOD) for extracting named entity translations in a post decoding framework. We conclude with linguistic phenomena-driven human evaluation of our forays into enhancing the syntactic and semantic constraints on a phrase-based SMT system.

Keywords: Machine Translation, Parallel Treebanks, Linked Open Data, Manual Evaluation

1 Introduction to Three Deep Language Processing Tools

Machine Translation (MT) like other language processing tasks is confronted with the Zipfian distribution of relevant phenomena. Although surface-data-driven systems have enlarged the head considerably over the last years, the tail still remains a challenge. Many approaches have therefore tried to include various forms of linguistic knowledge in order to systematically address chunks of the tail [1]. Unfortunately, todays automatic measures for MT quality are usually not able to detect these particular changes in the translations that may or may not constitute improvements. Therefore, we have argued for an evaluation approach that extends the current MT evaluation practice by steps where language experts inspect systems outputs [2]. We have started to use this extended evaluation approach in our contribution to the WMT2016 IT task [3]. In this paper, we will report more in-depth on three of the "deeper" ingredients of our work.

2 Baseline Machine Translation Systems

The experiment is based on two baseline systems: **Phrase-based SMT** follows several state-of-the-art phrase-based system settings as indicated in the Shared

task of Machine Translation in WMT [4]. As the best system UEDIN-SYNTAX [5] included several components which were not openly available, we proceeded with adopting several settings from the next best system UEDIN [6]. In our system we follow the practice of augmenting the generic training data (Europarl [7], News Commentary, MultiUN [8], Commoncrawl [9]) with domain-specific data (Libreoffice, Ubuntu, Chromium [10]), and building relevant extensive language models, interpolated on in-domain data, as described above. **Rule-based MT** (RBMT) as in the transfer-based system Lucy [11] is also part of our experiment as a baseline, due to its state-of-the-art performance in many shared tasks. In this method, translation occurs in three phases, namely analysis, transfer, and generation. All three phases consist of hand-written linguistic rules.

The set of parallel sentences for training, and the development and test sets for tuning and testing respectively were sourced from the data provided for the WMT 2016 shared task on machine translation of IT domain [12], available at `http://www.statmt.org/wmt16/it-translation-task.html`.

3 Syntax-aware Phrase Extraction

In this section we describe a syntax-aware enhancement to the phrase-based SMT baseline system described in Section 2. We extract linguistically motivated phrase pairs by obtaining phrase structure parse trees for both the source and target languages using monolingual constituency structure parsers such as the Berkeley Parser [13], and then aligning the subtrees using a statistical tree aligner [14]. These phrase pairs (illustrated with an example in Figure 1) are then merged with the phrase pairs extracted in the baseline SMT system into one translation model. Thus we are merely using syntax to constrain the phrase boundaries and enabling SMT decoder to pick syntax-aware phrases, thereby ensuring noun phrases and verb phrases remain cohesive. Through experimentation detailed in [15], we have discovered that non-linguistic (BASE) phrase-based models have a long tail (of coverage) and syntax-aware phrases underperform, if not concatenated with non-linguistic phrase pairs. We observed the syntax-aware system scored 0.8 BLEU points over the baseline system.

Example (1) illustrates how the syntax-aware system (SYN) improves over the baseline SMT system (BASE) by outputting the missing modal verb *ändern*.

(1) **Src (en):** You can **change** the screen saver settings.
Base (de): Sie können die Bildschirmschoner Einstellungen.
Syn (de): Sie können die Bildschirmschoner Einstellungen **ändern**.
Ref (de): Sie können die Bildschirmschoner-Einstellungen **ändern**.

4 Named Entity Translation Using Linked Data

Given the fact that our SMT system was not trained on data from the same domain as our testset (IT-domain), a number of technical terms (named entities) were either mistranslated or not translated consistently. One technique to

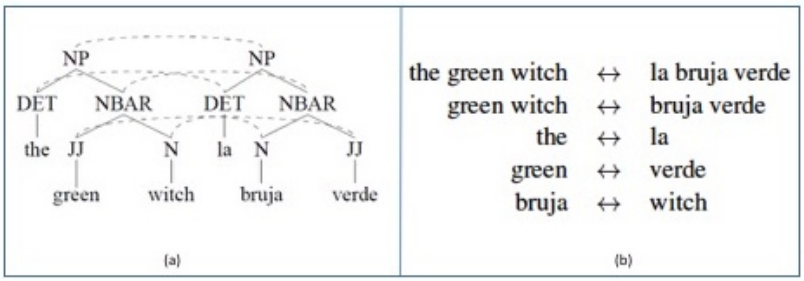

Fig. 1. Example of (a) A parallel treebank entry and (b) The associated set of extracted phrase pairs.

address this is to integrate the SMT system with a Named Entity Recognition (NER) system. In this section, we describe our approach.

We exploit multilingual terms semantically linked with each other in the form of freely available linguistic linked data on the web such as DBpedia[1] to identify named entities in our dataset in the same vein of [16]. These entities and their linked translations are then forced upon the SMT decoder such that the Moses decoder favours these translations over those from the translation model. A step-by-step procedure is detailed in [17].

Example (2) illustrates how the term *MS Paint* is wrongly identified as a person in the baseline system (BASE). On the other hand, the linked data system (LINK) correctly disambiguates the entity.

(2) **Src (en):** **MS Paint** is a good option.
Base (de): **Frau Farbe** ist eine gute wahl.
Link (de): **Microsoft Paint** ist eine gute wahl.
Ref (de): **MS Paint** ist eine gute Möglichkeit.

5 Deep Manual Evaluation

We carried out an extensive manual evaluation of the performance of our MT systems described above. To this end, we created a domain-specific test suite with the objective of validating the systems' capabilities of specific linguistic phenomena.[2] Our method consists of the following steps: A linguist identifies systematically occurring errors related to linguistic phenomena in the output of the systems. 100 segments containing the respective phenomenon are randomly extracted for each linguistic category. The total occurrences of the phenomena in the source segments are counted in the selected sets and analogous in the

[1] `http://wiki.dbpedia.org`
[2] We understand the term "linguistic phenomena" in a pragmatic sense, covering a wide range of issues that can impact translation quality.

system outputs[3]. The instances of the latter are divided by the instances of the former, giving the percentage of correctly translated phenomena.

The following example depicts the counting of instances of one of the linguistic phenomena, namely the menu item separators ">". All systems except for the RBMT system correctly transfer the ">".

(3) source: Go to Settings > iCloud > Keys. *2 inst.*
SMT: Gehen Sie zu Einstellungen > icloud > Schlüssel. *2 inst.*
RBMT: Gehen Sie zu Einstellungs-> iCloud >-Tasten. *0 inst.*
Syntax: Gehen Sie zu Einstellungen > icloud > Schlüssel. *2 inst.*
linked d.: Gehen Sie zu Einstellungen > icloud > Schlüssel. *2 inst.*

The linguistic categories that we found to be prone to translation errors in this context can be found in Table 1. For these categories, 2105 instances in 657[4] segments were found altogether. The overall average performance of the four systems at hand is rather similar, ranging from 71% to 77%.

Even though the **SMT** and the **RBMT** system have very similar overall average scores that outperform the other two systems, their scores on the phenomena are quite complimentary. The two linguistic extensions did not have strong effects on the performance of the systems on the error categories that we found error-prone (in pilot studies) and important for the given IT helpdesk domain. The **SMT-syntax** and the **linked data** system have similar overall scores and similar scores on the linguistic categories. What is particularly noteworthy is the only (negative) outlier, namely phrasal verbs. Precisely in this class, we would have hoped to see an improvement in performance of SMT-syntax. We will further investigate the reasons for this failure of dealing with phrasal verbs.

Table 1. Translation accuracy on manually evaluated sentences focusing on particular phenomena. Boldface indicates best systems on each phenomenon (row) with a 0.95 confidence level.

	#	SMT	RBMT	Syntax	linked d.
imperatives	247	68%	**79%**	68%	68%
compounds	219	55%	**87%**	55%	56%
">" separators	148	**99%**	39%	**97%**	**97%**
quotation marks	431	**97%**	94%	93%	94%
verbs	505	85%	**93%**	81%	85%
phrasal verbs	90	22%	**68%**	7%	12%
terminology	465	**64%**	50%	53%	52%
average		**76%**	**77%**	71%	72%

[3] A detailed description of how to count the occurrences of the phenomena including explicit examples will be published elsewhere.

[4] For the phrasal verbs, only 57 instead of 100 segments could be extracted as this is a rather rarely occurring phenomenon.

6 Conclusion

We have described several ways of making machine translation more linguistically aware. We have attempted to introduce linguistically aware phrases in the models as well as show improvements in the translation of named entities by linking with semantic web resources such as the DBpedia. Our detailed evaluation of relevant linguistic phenomena has shown that the performance of several MT systems differs as do several ways of system combinations. Given this detailed method and results, it is now possible to select/improve systems with respect to a given task. The deep linguistic evaluation we have shown is task-based. In other domains and settings, other issues would need to be inspected. While reference-based automatic evaluation treats requirements of the task only very indirectly and measures "improvement on average", this direct, source-driven evaluation makes it possible to evaluate the performance and measure improvement on task-specific aspects. One obvious way for improving statistical systems would be to create targeted training material focussing on the relevant aspects such as imperatives starting from the test items.

7 Acknowledgment

This article has received support from the EC's Horizon 2020 research and innovation programme under grant agreements no. 645452 (QT21), from FP7 (2007-2013) under grant agreement number 610516: "QTLeap: Quality Translation by Deep Language Engineering Approaches", and by Project Digitale Kuratierungstechnologien (DKT), supported by the German Federal Ministry of Education and Research (BMBF), "Unternehmen Region", instrument "Wachstumskern-Potenzial" (No. 03WKP45). We also thank the two anonymous reviewers for their valuable comments.

References

1. Steedman, M.: Romantics and Revolutionaries: What Theoretical and Computational Linguistics need to know about each other. In: Linguistic Issues in Language Technology: 6(11) (2011)
2. Burchardt, A., Harris, K., Rehm, G., Uszkoreit, H.: Towards a Systematic and Human-Informed Paradigm for High-Quality Machine Translation. In: Proceedings of the LREC 2016 Workshop Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem, Portoro, Slovenia (2016)
3. Avramidis, E., Burchardt, A., Macketanz, V., Srivastava, A.: DFKI's system for WMT16 IT-domain task, including analysis of systematic errors. In: Proceedings of the First Conference on Machine Translation, Berlin, Germany (2016)
4. Bojar, O., Buck, C, Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Radu, S., Specia, L.: Findings of the 2013 Workshop on Statistical Machine Translation. In: Proceedings of the Eighth Workshop on Statistical Machine Translation, pp. 1–44. Association for Computational Linguistics, Sofia, Bulgaria (2013)

5. Nadejde, M., Williams, P., Koehn, P.: Edinburghs Syntax-Based Machine Translation Systems. In: Proceedings of the Eighth Workshop on Statistical Machine Translation, pp. 170–176. Association for Computational Linguistics, Sofia, Bulgaria (2013)
6. Durrani, D., Haddow, B., Heafield, K., Koehn, P.: Edinburghs Machine Translation Systems for European Language Pairs. In: Proceedings of the Eighth Workshop on Statistical Machine Translation, pp. 114–121. Association for Computational Linguistics, Sofia, Bulgaria (2013)
7. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of the tenth Machine Translation Summit, Volume 5, pp. 7986. Phuket, Thailand (2005)
8. Eisele, A., Chen, Y.: MultiUN: A Multilingual Corpus from United Nation Documents. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation, pp. 2868–2872. European Language Resources Association (ELRA), La Valletta, Malta (2010)
9. Buck, C., Heafield, K., van Ooyen, B.: N-gram Counts and Language Models from the Common Crawl. In: Proceedings of the Language Resources and Evaluation Conference, Reykjavk, Iceland (2014)
10. Tiedemann, J.: News from OPUS A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In: N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, Advances in Natural Language Processing, volume V, chapter V, pp. 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria (2009)
11. Alonso, J.A., Thurmair, G.: The comprendium translator system. In: Proceedings of the Ninth Machine Translation Summit. International Association for Machine Translation (IAMT) (2003)
12. Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A.J., Koehn, P., Logacheva, V, Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., Zampieri, M.: Findings of the 2016 Conference on Machine Translation. In: Proceedings of the First Conference on Machine Translation at ACL 2016, pp. 131–198. Association for Computational Linguistics, Berlin, Germany (2016)
13. Petrov, S., Klein, D.: Improved Inference for Unlexicalized Parsing. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics, Rochester, New York (2007)
14. Zhechev, V.: Unsupervised Generation of Parallel Treebank through Sub-Tree Alignment. In: Prague Bulletin of Mathematical Linguistics, Special Issue on Open Source Tools for MT, Volume 91 , pp. 89–98 (2009)
15. Srivastava, A.: Phrase Extraction and Rescoring in Statistical Machine Translation. PhD Thesis, Dublin City University (2014)
16. McCrae, J.P., Cimiano, P.: Mining Translations from the Web of Open Linked Data. In: Proceedings of the Joint Workshop on NLP, LOD, and SWAIE, pp. 8–11. Hissar, Bulgaria (2013)
17. Srivastava, A., Sasaki, F., Bourgonje, P., Schneider, J.M., Nehring, J., Rehm, G.: How to Configure Statistical Machine Translation for Linked Open Data. In: Proceedings of the 38th Annual Conference on Translating and Computer, London, United Kingdom (2016)