The Prague Bulletin of Mathematical Linguistics
NUMBER 108 JUNE 2017 209–220

Comparing Language Related Issues for NMT and PBMT between German and English

Maja Popović

Humboldt University of Berlin

Abstract

This work presents an extensive comparison of language related problems for neural machine translation and phrase-based machine translation between German and English. The explored issues are related both to the language characteristics as well as to the machine translation process and, although related, are going beyond typical translation error classes. It is shown that the main advantage of the NMT system consists of better handling of verbs, English noun collocations, German compound words, phrase structure as well as articles. In addition, it is shown that the main obstacles for the NMT system are prepositions, translation of English (source) ambiguous words and generating English (target) continuous tenses. Although in total there are less issues for the NMT system than for the PBMT system, many of them are complementary – only about one third of the sentences deals with the same issues, and for about 40% of the sentences the issues are completely different. This means that combination/hybridisation of the NMT and PBMT approaches is a promising direction for improving both types of systems.

1. Introduction

Neural machine translation (NMT), a new paradigm to statistical machine translation (SMT), has emerged very recently and has already surpassed the performance of the mainstream approach in the field, phrase-based MT (PBMT) for a number of language pairs. In PBMT, different models (translation, reordering, target language, etc.) are trained independently and combined in a log-linear scheme in which each model is assigned a different weight by a tuning algorithm. On the contrary, in NMT all the components are jointly trained to maximise translation quality. On one side, NMT represents a simplification – a large recurrent network trained for end-to-end

Corresponding author: maja.popovic@hu-berlin.de

Cite as: Maja Popović. Comparing Language Related Issues for NMT and PBMT between German and English. The Prague Bulletin of Mathematical Linguistics No. 108, 2017, pp. 209–220. doi: 10.1515/pralin-2017-0021.

translation is considerably simpler than a PBMT system which integrates multiple components and processing steps. On the other side, the NMT process is less transparent.

So far, the translations produced by NMT systems have been evaluated mostly in terms of overall performance scores, both by automatic and by human evaluations. This has been the case of last year's news translation shared task at the First Conference on Machine Translation (WMT16). In this translation task, outputs produced by different MT systems were evaluated (i) automatically, by various evaluation metrics, and (ii) manually, by means of ranking translations or by assigning them an overall quality score. In all those evaluations, the performance of each system is measured by means of an overall score which provides useful information about general performance of the system but does not provide any additional information.

To the best of our knowledge, only two detailed analyses of the NMT approach and comparisons with PBMT approach have been carried out so far. (Bentivogli et al., 2016) conducted a detailed analysis for the English-to-German translation of transcribed TED talks and found out that NMT (i) decreases post-editing effort, (ii) degrades faster than PBMT with sentence length and (iii) results in a notable improvement regarding reordering, especially for verbs. (Toral and Sánchez-Cartagena, 2017) go further in this direction by conducting a multilingual and multifaceted evaluation and found out that (i) NMT outputs are considerably different than PBMT outputs, (ii) NMT outputs are more fluent, (iii) NMT systems introduce more reorderings than PBMT systems, (iv) PBMT outperforms NMT for very long sentences and (v) NMT performs better in terms of morphological and reordering errors across all language pairs.

In this paper, we go in slightly different direction by identifying and comparing language related issues for two German-English systems, one NMT and one PBMT, in both translation directions. Identification of language related issues for machine translation has begun relatively recently (e.g. (Popović and Arčan, 2015), (Comelles et al., 2016)) and, although related, goes beyond the standard error classification task. Definition of issues is based both on general linguistic knowledge as well as on the phenomena related to the (machine) translation process.

The issues are manually identified for 267 English-to-German source sentences and 204 German-to-English source sentences from the WMT16 News domain data and their translations by NMT and PBMT systems.

The main goals of the experiments are:

1. to compare overall distributions of issues for the NMT and the PBMT system and identify the particular strengths of the NMT approach, i.e. particular weaknesses of the PBMT approach for each translation direction;
2. to examine the overlap between issue types in two systems in order to determine if the NMT approach simply handles all the phenomena better, or there are complementary differences. This is an important question for better understanding

potentials and limits of combination and hybridisation of the two approaches which already has shown some promising results (Niehues et al., 2016).

We choose the German-English language pair in both directions because it has been known as a rather hard one for PBMT and the improvements yielded by the NMT approach are large, especially when translating into German. Our analyses are conducted on the Edinburgh University submissions of NMT and PBMT systems to the WMT16 translation task for each language direction which were (one of) the best ranked. This (i) guarantees the reproducibility of our results as all the MT outputs are publicly available, (ii) ensures that the systems evaluated are state-of-the-art, as they are the result of the latest developments at a top MT research group worldwide. If the paper is accepted, the annotated texts with issue labels will be made publicly available, too.

We believe that our evaluation results will be of interest to the wider research community, both regarding development of NMT and PBMT systems as well as regarding development of MT evaluation and error analysis methods.

2. Related work

The first detailed analysis and comparison between the NMT and PBMT approach is carried out in (Bentivogli et al., 2016). They analysed 600 sentences from IWSLT transcriptions of TED talks (i.e. spoken language) translated from English into German. They conducted automatic analysis on manually post-edited data in terms of morphological, lexical and ordering errors together with the fine grained analysis of ordering errors and found out that the main advantage of NMT approach is better ordering, especially for verbs.

(Toral and Sánchez-Cartagena, 2017) performed a multifaceted automatic analysis based on independent human reference translations for nine language pairs from news domain. The analysis consists of output similarity, fluency measured by LM perplexity, degree of reordering as well as three broad error classes: morphological, reordering and lexical errors. The main findings confirm the results from previous publication, i.e. the reduction of morphological and reordering errors by NMT. In addition, both publications report degradation of the NMT approach for long sentences.

While both publications report results of an extensive analysis and comparison of NMT and PBMT approaches, neither of publications deals with language related issues based on the source and the target language properties and their differences.

The first step towards such analysis is reported in (Farrús et al., 2010) where a simple error scheme containing five broad classes is used for comparison of two Spanish-Catalan SMT systems. This scheme is then further expanded in (Comelles et al., 2016) by identifying and classifying relevant linguistic features for the English-Spanish language pair based on general linguistic knowledge as well as on the phenomena occurring in the given corpus. The linguistic issue taxonomy is used for development of a linguistically motivated automatic evaluation metric VERTa (Comelles et al., 2012) which enables using different combinations of the described linguistic features.

Similar analysis is conducted in (Popović and Arčan, 2015) where problematic patterns for PBMT between South Slavic languages on one side and English and German on another side were identified and analysed.

Nevertheless, none of the publications dealing with linguistically motivated issues includes analysis of an NMT system, nor the German-English pair.

3. Language related issues

Identification of language related issues has begun rather recently, so there are still no strict guidelines regarding their definition. In any case, the issues have to be linguistically motivated so that they can reflect the (un)ability of a machine translation system to translate specific linguistic phenomena. However, they should not only contain traditional linguistic categories but also categories which are related to the (machine) translation process. The issues should be clearly defined and widely understandable so that the results can be easily understood and shared.

Although issue identification task is related to error classification task, it goes beyond it: some of the issues defined so far directly correspond to some typical error classification categories, such as "verb form" or "mistranslation", however for a number of issues such relation is still hard to find.

For example, when an MT system does not handle a source German compound properly, error categories in the English output can be "mistranslation", "missing word" (components are missing), "word order" (components are in incorrect position), but the issue label for each of these cases would be "compound word".

Annotation was carried out by researchers familiar with human and machine translation process. The source language, its reference translation, and the two translation outputs in random order were given to the annotators.

The most prominent issues for both translation directions are:

- **ambiguous source word**
  The obtained translation for the given word is in principle correct, but not in the given context.
- **article**
  Rules for articles in German and English differ – therefore, some of the articles are added, missing, or incorrectly translated as (in)definite. In addition, some of the German articles are incorrectly inflected.
- literal translation
  Word-by-word translated parts.
- **mistranslation**
  Incorrect translation of words or word groups.
- **source multiword expression**
  Failing to treat a multiword expression as a whole.

- **MT phrase structure**
  Phrases/chunks are not treated properly so that the (group(s) of) words are misplaced, mistranslated and/or incorrectly inflected. "MT" refers to the fact that these are not linguistic phrases.
- **preposition**
  Mostly mistranslated, sometimes omitted or added.
- **verb**
  Problems with translation of verbs: main, auxiliary, modal, participle, formation of tenses, order, etc.
  - **form**
    Verb inflection does not correspond to the person and/or the tense.
  - **order**
    Verb or verb parts are misplaced.
  - **missing**
    Verb or verb parts are missing.

For English-to-German translation:

- **noun collocation**
  English sequence consisting of a head noun and additional nouns and adjectives is incorrectly translated, often into an unintelligible construction.
  - **noun collocation + compound**
    English noun collocation which corresponds to an incorrectly formed German compound word. The German compound word is mistranslated, or there are problems with components: missing, added or separated.

For German-to-English translation:

- **German compound**
  German compound is mistranslated or remained untranslated, or there are problems with components: missing, added or in incorrect order.
- **English continuous verb tenses**
  Continuous verb tenses do not exist in German, so that English present/past continuous tense is often substituted by simple present/past tense, or there are problems with verb parts.

4. Data sets

The texts used in the described experiments consist of 267 English-to-German source sentences and 204 German-to-English source sentences from the WMT16 News domain data and their NMT and PBMT translations. The annotation process is still fully manual, so that annotating the whole test sets each consists of about 3000 sentences would be too intensive. Therefore the smaller subsets were extracted from the set of the sentences which participated in human ranking, in order to also enable future experiments concerning relationship between issues and ranks. For the same

direction	system	BLEU	chrF
en→de	NMT	35.0	61.9
	PBMT	31.5	58.5
de→en	NMT	42.5	66.5
	PBMT	38.9	66.2

Table 1. Overall automatic scores BLEU and chrF on analysed texts for both systems and both translation directions.

reason, only two systems were analysed, one NMT and one PBMT. (Partial) automatisation of the annotation process should be certainly part of the future work.

The NMT system (Sennrich et al., 2016) is based on attentional encoder-decoder and operates on subword units. In addition, back-translations of the monolingual News corpus is used as additional training data. This system is ranked as the best for both translation directions.

The PBMT system (Williams et al., 2016) is a Moses based system which follows the standard PBMT approach of scoring translation hypotheses using a weighted linear combination of features. The core features are 5-gram LM model, phrase translation and lexical translation scores, word and phrase penalties and a linear distortion score. Tuning of model weights is performed by k-best batch MIRA.

Although other systems were ranked better in the WMT16 task, we decided to use this one because it has been developed by the same group, and we believe that therefore the comparison is more reliable.

5. Results

5.1. Overall automatic scores

First, in Table 1 we report the overall BLEU (Papineni et al., 2002) and chrF (Popović, 2015) scores for the analysed texts. The NMT system clearly outperforms the PBMT system for both translation directions and by both scores. It can be noted that the absolute chrF improvement is larger for translation into German, indicating that NMT introduces morphological improvements.

5.2. Comparison of issue distributions

The frequencies of the most prominent issues for the NMT and the PBMT system are presented in Table 2. Since the issues are defined on the sentence level, the numbers in tables represent raw issue counts normalised by the total number of sentences. For example, the verb form issues for English→German translation are interpreted as follows: from 100 English source sentences, verb form problems occur in 4.9 sentences translated by NMT and in 9.4 sentences translated by PBMT.

In addition, percentages of correct sentences (”no issues”) as well as of sentences for which it was difficult to define any particular issue (”difficult to analyse”) are shown.

First, it can be seen that the percentage of correct sentences[1] is significantly higher for the NMT system than for the PBMT system. As for ”difficult” sentences, there is almost no difference between the systems, only between the translation directions – there are more for English-to-German.

As for the issue types, for both translation directions the NMT system clearly outperforms the PBMT system for:

- verbs in the following aspects: form, order and omission
- articles
- English noun collocations and German compounds
- phrase structure

These findings, while shedding different kind of light on the strengths and weaknesses of the two approaches, also confirm the results reported in previous work, namely that one of the main advantages of the NMT approach is better dealing with morphology and ordering, especially for verbs. Verb forms and German compounds clearly represent morphological challenges, whereas both morphology and order are implicitly related to phrase structure and treatment of noun collocations. Since all these issues are strongly related to fluency, the fluency improvements reported in related work are corroborated, too.

The results also show that for some issue types the behaviour depends on the translation direction, so that NMT outperforms PBMT for:

- ambiguous words and literal translations for German to English
- mistranslation and multiword expressions for English to German

but for the opposite translation direction these issues are better handled by the PBMT system.

Furthermore, target English continuous tenses are slightly better handled by PBMT, and represent the most frequent obstacle for German-to-English NMT translation (11.7%).

Finally, it can be observed that the prepositions are rather problematic for both systems. They are the most frequent issue for the English-to-German NMT system and second frequent (after continuous tenses) for the other translation direction, so the future work on NMT improvement should take this into account.

Sentence length

Previous work reported significance of the sentence length, namely that the PBMT approach outperforms NMT for longer sentences. Therefore we also investigated issue distributions for different sentence lengths. Nevertheless, we have found neither

[1] About 8% of sentences is identical to the corresponding reference translation.

English→German	system	
issue type	NMT	PBMT
no issues	**35.7**	20.2
difficult to analyse	5.6	6.4
(src) ambiguous word	15.4	**10.5**
article	**8.5**	15.8
literal	6.7	**6.0**
mistranslation	**5.6**	7.5
(src) multiword expression	**4.9**	5.2
(src) noun collocation	**4.5**	7.1
+ *(tgt) compound*	**1.9**	7.1
(MT) phrase structure	**1.1**	5.6
preposition	17.5	**17.2**
verb – *form*	**4.9**	9.4
– *order*	**1.5**	10.9
– *missing*	**1.5**	24.0

German→English	system	
issue type	NMT	PBMT
no issues	**39.0**	26.3
difficult to analyse	3.9	3.9
(src) ambiguous word	**9.3**	10.7
article	**7.8**	13.7
compound	**4.4**	7.8
literal	**4.4**	9.8
mistranslation	9.3	**8.3**
(src) multiword expression	4.4	**3.4**
(MT) phrase structure	**2.0**	6.8
preposition	11.2	**10.2**
verb – *form*	**2.0**	2.9
– *order*	**0.5**	5.8
– *missing*	**1.0**	5.8
– *continuous tense*	11.7	**8.3**

Table 2. Percentage of issues (raw counts normalised over the total number of sentences) for English-to-German (above) and German-to-English (below) translation.

overlap degree	% of sentences	
	en→de	de→en
complete (100%)	27.3	31.9
high (>50%)	9.7	13.7
low (≤50%)	20.6	16.2
none (0%)	42.4	38.2

Table 3. Percentage of sentences with four distinct overlap degrees between NMT and PBMT issues: complete overlap (100%), high overlap (>50%), low overlap (≤50%) and no overlap (0%).

a relation between issue types and sentence length, nor advantages of the PBMT system for longer sentences. It should be noted that the maximal sentence length in our data set was 36 words, whereas the results reported in previous work show that important changes start for sentences longer than 40 words. Therefore this aspect should be investigated thoroughly in future work.

5.3. Overlap between PBMT and NMT issues

The results described in previous section have shown that the NMT system does not simply outperform the PBMT system by having less issues of all types, but that there are certain complementary differences. In order to explore overlapping and complementary issues, we carried out the following experiments.

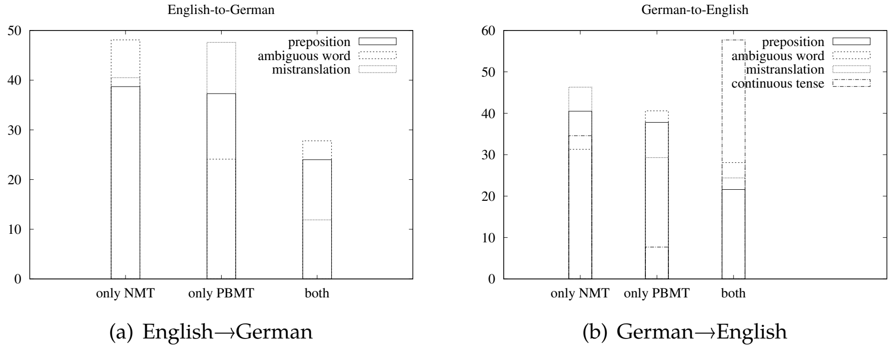
As a first step, we calculated overall overlap of the issues for each translation direction in the form of the F-score. For English-to-German this score is 37.9%, and for German-to-English 44.6%. These scores are not very high, indicating that there is a number of complementary issues.

The next step was to calculate the overlap F-score for each sentence and then divide the sentences into four groups: 1) complete overlap, same issues (100%), 2) high overlap (between 50 and 100%), 3) low overlap (between 0 and 50%) and 4) no overlap, completely different issues (0%).

The distributions of sentences over these four overlap degree groups are shown in Table 3 for both translation directions, and it can be seen that:

- only about one third of the sentences has identical issues;
- the majority (about 40%) of sentences have completely different issues;
- there are more sentences with low overlap than those with high overlap.

These findings show that, although the NMT approach surely performs better than the PBMT approach, there are complementary problems and errors. We believe this is an important finding because it means that there is a room for improvement of both systems in terms of combination and hybridisation.



(a) English→German (b) German→English

Figure 1. Distribution (%) of complementary and identical issues.

The last step in this direction was to examine which are the most frequent overlapping issues as well as how much of the prominent NMT issues is complementary with the PBMT ones.

First part of the analysis showed that the majority of the identical sentences are either correct, or are sentences for which it was hard to define issues.

As for the most prominent NMT issues, namely prepositions, ambiguous words, mistranslations and English continuos tenses, the percentages of complementary and overlapping occurrences is shown in Figure 1 for both translation directions. It can be seen that about 20-50% of total occurrences of the particular issues are complementary, i.e. do not overlap. The only exception is the verb continuous tense where the overlap is large. These results indicate that the combination of NMT and PBMT approach could "help" dealing with prepositions and lexical issues (mistranslations and ambiguous words).

6. Summary and outlook

We have conducted an extensive comparison between NMT and PBMT language related issues for the German-English language pair in both translation directions. Our aim has been to shed additional light on the strengths and weaknesses of both approaches, as well as to explore if there are complementary issues.

Following the two main goals of our experiments presented in Introduction, our main findings are:

1. The particular strengths of the NMT approach are better handling of (i) verb order, forms and avoiding verb omissions, (ii) English noun collocations and German compound words, (iii) articles and (iv) phrase structure. All these is-

sues are completely or strongly related to morphology and word order, and to fluency as well, which corroborates the results reported in previous work.

2. Although the NMT approach in total has less issues, there is a number of sentences with complementary issues. This finding can help improvement of both systems by means of combination and/or hybridisation.

Additional important findings are:

- dominant problems for the NMT system are prepositions, translation of English ambiguous words into German and forming English verb continuous tenses;
- most occurrences of prepositions, ambiguous words and mistranslations are complementary.

It should also be noted that translating prepositions represents an important obstacle for both systems and it should be addressed in future work. Apart of this, there is a number of other directions for future work, such as (i) improvement of one or both systems by addressing some of the most prominent issues, (ii) exploring combination of two approaches, (iii) investigating other language pairs, (iv) working towards (partial) automatisation of the annotation process in order to achieve scalability.

We believe that our evaluation results will be of interest both for development of NMT and PBMT systems as well as for development of MT evaluation and error analysis methods. We conducted all experiments on publicly available data, and the annotated texts are also publicly available[2].

Acknowledgements

This research has received funding from the European Union's Horizon 2020 research and innovation programme TraMOOC under Grant Agreement No. 644333.

Bibliography

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pages 257–267, Austin, Texas, November 2016.

Comelles, Elisabet, Jordi Atserias, Victoria Arranz, and Irene Castellón. VERTa: Linguistic Features in MT Evaluation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, May 2012.

Comelles, Elisabet, Victoria Arranz, and Irene Castellón. Guiding Automatic MT Evaluation by Means of Linguistic Features. *Digital Scholarship in the Humanities*, September 2016.

Farrús, Mireia, Marta Ruiz Costa-Jussà, José Bernardo Mariño, and José Adrián Rodríguez Fonollosa. Linguistic-based Evaluation Criteria to Identify Statistical Machine Translation Errors. In *Proceedings of the 14th Annual Conference of the European Asso ciation for Machine Translation (EAMT 2010)*, pages 167–173, Saint-Raphael, France, May 2010.

[2] `https://github.com/m-popovic/german-english_pbmt-nmt-issues`

Niehues, Jan, Eunah Cho, Thanh-Le Ha, and Alex Waibel. Pre-Translation for Neural Machine Translation. In *Proceedings of the 26th International Conference on Computational Linguistics (CoLing 2016)*, pages 1828–1836, Osaka, Japan, December 2016.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wie-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computa tional Linguistics (ACL 2002)*, pages 311–318, Philadelphia, PA, July 2002.

Popović, Maja. chrF: Character n-gram F-score for Automatic MT Evaluation. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT 2015)*, pages 392–395, Lisbon, Portugal, September 2015.

Popović, Maja and Mihael Arčan. Identifying Main Obstacles for Statistical Machine Translation of Morphologically Rich South Slavic languages. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT 2015)*, Antalya, Turkey, May 2015.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. Edinburgh Neural Machine Translation Systems for WMT16. In *Proceedings of the 1st Conference on Machine Translation (WMT 2016)*, pages 371–376, Berlin, Germany, August 2016.

Toral, Antonio and Víctor Manuel Sánchez-Cartagena. A Multifaceted Evaluation of Neural versus Statistical Machine Translation for 9 Language Directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, Valencia, Spain, April 2017.

Williams, Philip, Rico Sennrich, Maria Nadejde, Matthias Huck, Barry Haddow, and Ondřej Bojar. Edinburgh's Statistical Machine Translation Systems for WMT16. In *Proceedings of the 1st Conference on Machine Translation (WMT 2016)*, pages 399–410, Berlin, Germany, August 2016.

Address for correspondence:
Maja Popović
`maja.popovic@hu-berlin.de`
Humboldt University of Berlin
Unter den Linden 6, Berlin, Germany