Neural Machine Translation for Literary Texts

Henny Sluyter-Gäthje, Heike Zinsmeister, Fabian Barteld

henny.sluyter-gaethje@uni-potsdam.de, {Heike.Zinsmeister, Fabian.Barteld}@uni-hamburg.de

MOTIVATION

Specifics of literary texts

- Poeticity
- Rich vocabulary
- Large set of syntactic constructions
- Variable domains
- Long sentences

(van Cranenburgh & Bod 2017)

Examples from Uncle Tom's Cabin

- "They an't pop'lar, and they an't common; but I stuck to "em, sir; I've stuck to "em, and realized well on "em […]"
- In fact, if not exactly a believer in the doctrine of the efficiency of the extra good works of saints, he really seemed somehow or other to fancy that his wife had piety and benevolence enough for two to indulge a shadowy expectation of getting into heaven through her superabundance of qualities to which he made no particular pretension.

Neural MT versus statistical MT

Neural Machine Translation newly proposed in 2014, improved over state of the art statistical methods

NMT
- Fewer lexical, morphological and reordering errors
- More fluent output
- Better handling of rare words

SMT
- Better performance on longer sentences
- Copes better with small amount of training data

Compare: Bentivogli et al. (2016), Toral & Sánchez-Cartagena (2017), Koehn, P. & Knowles, R. (2017)

CORPORA

In-domain

- Bilingual Formal/Informal Address Corpus (Faruqui&Pado, 2012)
- Books corpus (Tiedemann, 2012)

→ 114 texts

(English vocabulary size 117492, German vocabulary size: 222089)

Mainly originally English (55) and French (34), some German texts (16)

- Corpus of German Language Fiction (Fischer& Strötgen, 2017)

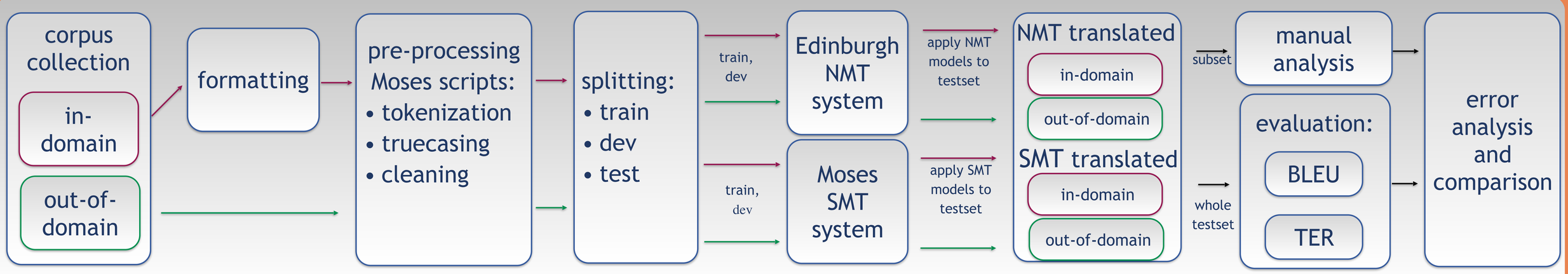
→ 30 texts (Vocabulary size: 115312)

Out-of-domain

- Europarl (Koehn et al., 2007)

→ 1,920,209 sentences

PIPELINE



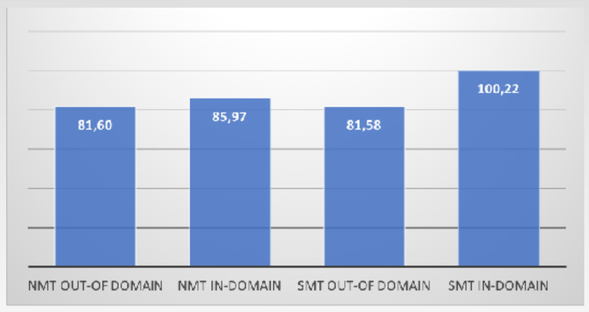
RESULTS

Original Accordingly, on entering the room, we found him present, in the uniform of an officer of his rank, about to commence a march in the forests of America.

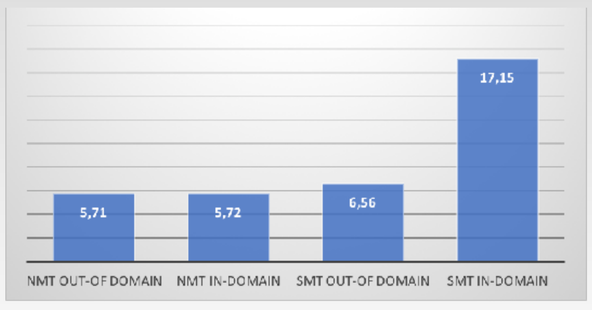
NMT Als er in die Zimmer trat, fanden wir ihn, in der Uniform eines Offiziers seines UNK zes, einen Marsch in den Wäldern Amerikas zu beginnen.

SMT gleich beim Eintritt in das Zimmer , wo wir ihn gefunden , in der Uniform eines Offiziers seines Ranges , im Begriff , anzufangen , einen Marsch in die Wälder von Amerika .

Automatic Comparison

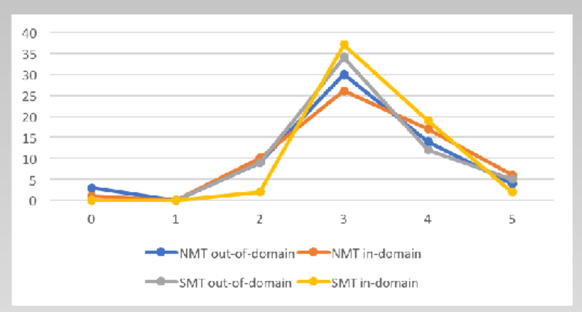


Translation Error rate (TER), a low score is desirable. Measures the amount of steps needed for post-processing.

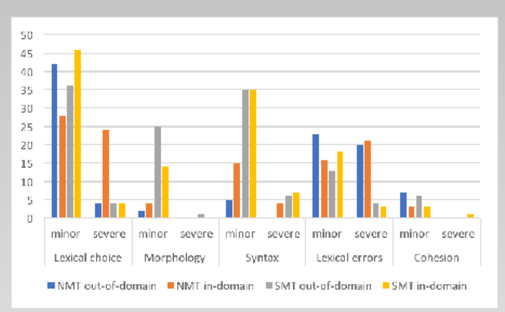


BLEU score, a high score is desirable. Matches ngrams from output to the reference output.

Manual error analysis and score



Manual scores, 0: No connection to sources sentence. 1: Not understandable, 2: Information can be gathered, 3: Sentence clearly is translation of source, 4: Only few mistakes, 5: No errors.



Manual analysis is based on Popović et al. (2013).
Lexical choice: Wrong translation
Lexical errors: Omitted/ repeated word

Summary

All systems perform equally poorly.

The SMT systems produce more syntactical and more severe morphological errors. The NMT systems cannot produce better word choice, probably due to small amount of (in-domain) training data.

FUTURE WORK

Training Data for NMT

- Size of training data: Need bigger corpus of in-domain data, best would be direct translations and lower domain variability.
- Alignment: Need fully reviewed alignments, reason for very low scores often due to misalignment. One-to-one alignment would be optimal. Automatic scoring metrics would be more reliable.
- Annotation: Edinburgh NMT system allows for POS-tagged input, the problem of unknown words in the output could be reduced as many are proper nouns.

REFERENCES

Cranenburgh, A. v. & R. Bod. 2017. A Data-Oriented Model of Liter- ary Language. In Proceedings of EACL 2017.

Faruqui, M. & Pado, S. (2012), Towards a model of formal and informal address in english, in 'Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics', Association for Computational Linguistics, pp. 623–633.

Fischer, F. & Strötgen, J. (2017), 'Corpus of German-Language Fiction (txt)'

Koehn, P. & R. Knowles. 2017. Six challenges for neural machine translation. In Proceedings of the 1st Workshop on Neural Machine Translation.

Koehn, P. et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th ACL 2007.

Popović, M. et al. 2013. Learning from human judgments of machine translation output. In Proceedings of MT Summit.

Sennrich, R., B. Haddow & A. Birch. 2016. Edinburgh Neural Machine Translation Systems for WMT 16. In Proceedings of the 1st Conference on Machine Translation.

Tiedemann, J. (2012), Parallel data, tools and interfaces in opus, in N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk & S. Piperidis, eds, 'Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)', European Language Resources Association (ELRA), Istanbul, Turkey.

Toral, A. & A. Way. 2015. Translating Literary Text between Related Languages Using SMT. In Proceedings of CLfL@NAACL-HLT.

Toral, A. & Sánchez-Cartagena, V. M. (2017), 'A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 1063–1073, Valencia, Spain, April 3-7, 2017.