The Sources of Text Complexity for NMT

Anna Iankovskaia
University of Wolverhampton

Original idea of the project: Prof Dr Ruslan Mitkov

Structure of Presentation

- Goals of the research
- Research questions
- Previous work
- Methodology
- Limitations
- Preliminary results
- Acknowledgements
- References

Goals of the Research

- Although being the state of the art, NMT is still prone to errors
- The study aims to:

1) Identify typical lexical, syntactic, and grammatical patterns which could lead to errors

2) Develop a program capable of detecting some of them before the source language is processed by NMT

Research Questions

- What are the sources of complexity at lexical and syntactic level?
- What types of MWEs are most likely to be mistranslated by NMT?
- Is a transformer-based program able to predict where NMT is most likely to fail?

Previous Work (I)

- First attempts to determine the sources of complexity were made during the era of rule-based machine translation (RBMT)
  - “Translatability indicators“ i.e. text features able to degrade the quality of MT output (Underwood and Jongejan, 2001)
  - Lists of linguistic features contributing to lexical, syntactic, and semantic ambiguity as a set of rules to follow when authoring a text for MT (Bernth and Gdaniec, 2001)

Previous Work (II)

- Controlled language — “a restricted version of a natural language which has been engineered to meet a special purpose“ (Kittredge, 2016, p. 13)
- Confidence Index measuring the level of confidence of an MT system about the quality of its translation (Bernth, 1999)
- Tool able to determine whether a text in English is suitable for MT based on the averaged translatability index which is calculated from all translatability indicators and their weights (Underwood and Jongejan, 2001)

Previous Work (III)

- Several studies consider the correlation between MT and post-editing:
  - Correlation between the quality of MT output and the product analysis and the effort spent on the post-editing (Daems et al., 2017)
  - Correlation between the difficulty of the source text and the cognitive and technical effort of post-editors (O’Brien, 2005; O’Brien, 2006)

Frequent sources of complexity for MT

- Pronominal anaphora (Mitkov and Schmidt, 1998)
- Multi-word expressions (Barreiro et al., 2013)
- Lexical ambiguity i.e. polysemy (Carpuat and Wu, 2007; Ngueng et al., 2018)
- Sentence length (Hung, Ngueng and Shimazu, 2012)
- Difference in sentence structuring between the source and the target (Birch, Osborne and Koehn, 2008; Popović and Arčan, 2015)

Methodology: Investigation

Lexical and syntactic complexities:

- English-Russian NMT of 20 texts from the News Commentary Parallel Corpus (Tiedemann, 2012) by means of DeepL[1] and ModernMT[2]

↓

- Manual analysis of errors

1 https://www.deepl.com/translator

2 https://www.modernmt.com/translate/

Methodology: Implementation

Hybrid approach

Deep Learning:

BERT (Devlin et al., 2019)

One of the MWE patterns

Rule-based

Syntactic patterns

Methodology: Evaluation

- 2 X F-1 Score will be used

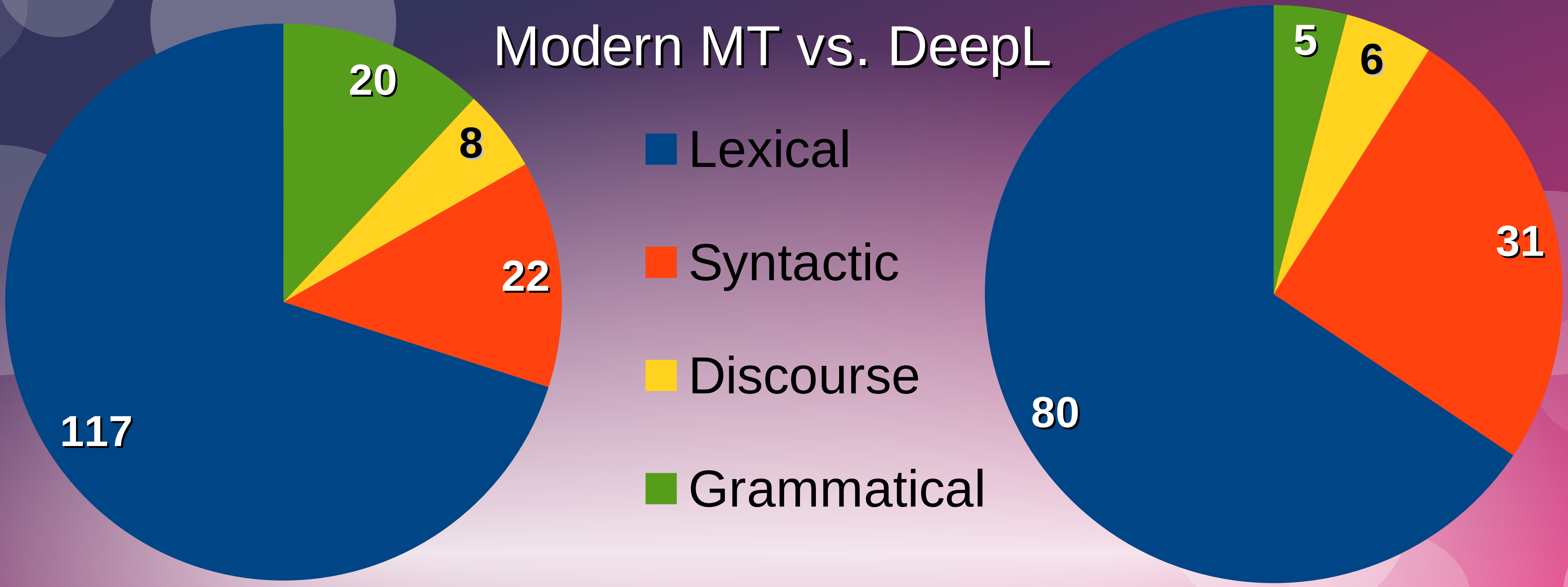
Limitations

- NMT is in the process of constant development and even some of the preliminary results might be already obsolete
- Numerous textual features that are difficult for NMT & impossibility to have all of them in the final program
- Limitations related to one language pair, domain and corpus size

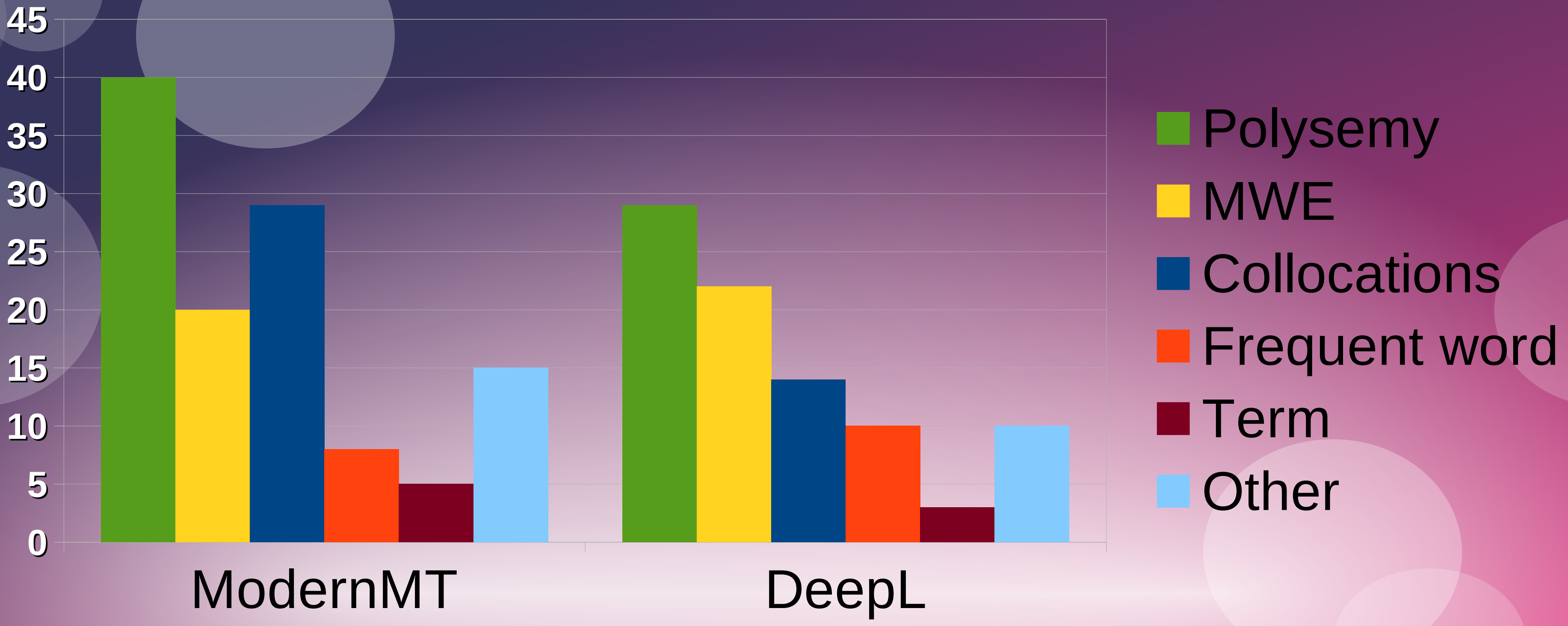
Preliminary results (I)

- 15 % of the texts analysed
- The author does not attempt to generalise these results to any extent and underlines that they apply only within the limits of the size and domain of the texts analysed

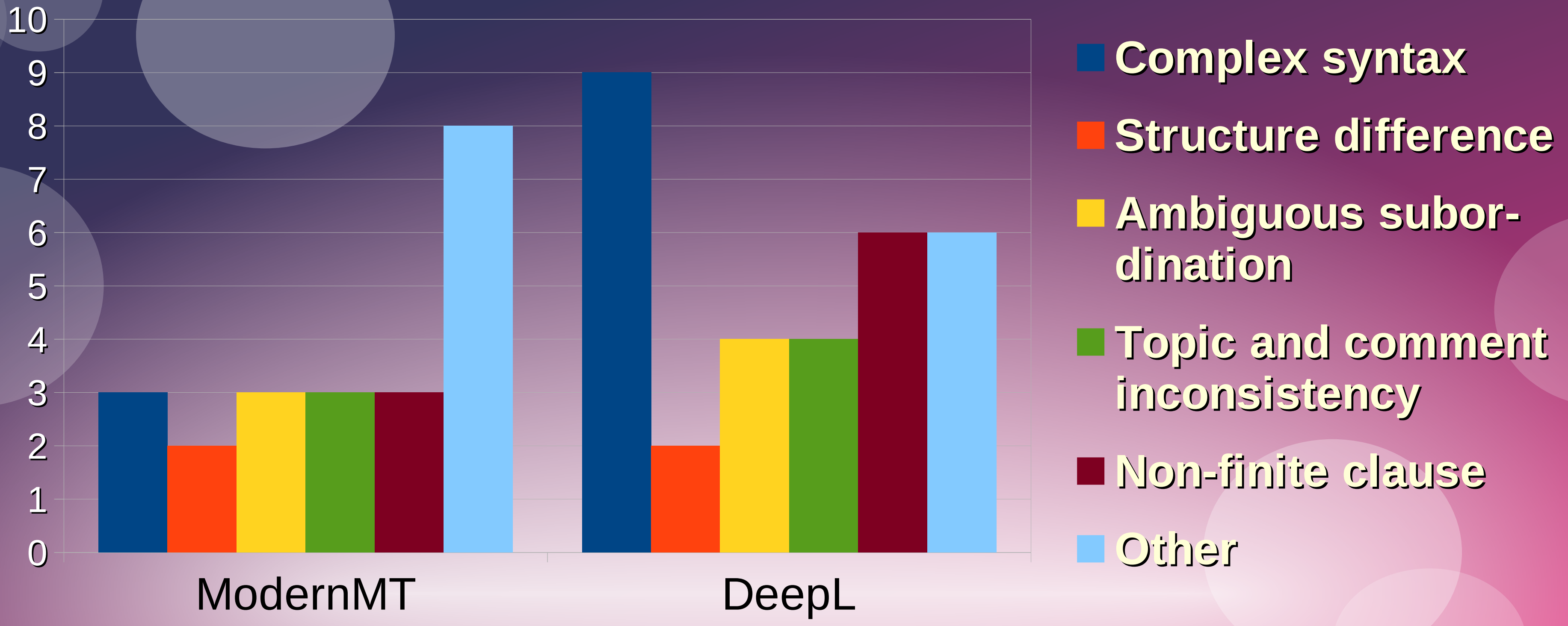
Preliminary results (II): sources of complexity



Preliminary results (III): lexical sources of complexity



Preliminary results (IV): syntactic sources of complexity



Aknowledgements

First and foremost, thank you to my supervisor Prof Dr R. Mitkov for all his support

As well as to other representatives of the EM TTI & University of Wolverhampton teaching & research team who have introduced me to NLP and always find time to answer my questions

Thank you for your attention!

References (I)

- Barreiro, A., Monti, J., Orliac, B. and Batista, F. (2013) When Multiwords Go Bad in Machine Translation. *Monti, J., Mitkov, R., Corpas Pastor, G., Seretan, V., (eds). Workshop on Multiword Units in Machine Translation and Translation Technology* [online]. Nice, 2 September 2013, pp. 26-33. [Accessed 10 November 2020]. Available at: < http://www.mt-archive.info/10/MTS-2013-W4-Barreiro.pdf>
- Bernth, A. (1999) A Confidence Index for Machine Translation. *TMI 99: 8th International Conference on Theoretical and Methodological Issues in Machine Translation* [online]. Chester, UK, August 1999, pp. 120-127. [Accessed 10 November 2020] Available at: <https://www.semanticscholar.org/paper/A-Confidence-Index-for-Machine-Translation-Bernth/d943c6fcd39798e0b5a9794e53152d43db273181>
- Bernth, A. and Gdaniec, C. (2001) MTranslatability. *Machine Translation* [online]. 16(3), pp. 175-218. [Accessed 10 November 2020]. Available at: <http://www.jstor.com/stable/40006963>
- Birch, A., Osborne, M. and Koehn, P. (2008) Predicting Success in Machine Translation. *Conference on Empirical Methods in Natural Language Processing* [online]. SIGDAT, Honolulu, 25-27 October 2008, pp. 745-754. [Accessed 10 November 2020]. Available at: <https://www.aclweb.org/anthology/D08-1078.pdf >
- Daems, J., Vandepitte, S., Hartsuiker, R.J., and Maken, L. (2017) Identifying the Machine Translation Error Types with the Greatest Effect on Post-Editing Effort. *Frontiers in Psychology* [online]. 8, 1282. [Accessed 10 November 2020]. Available at: <https://doi.org/10.3389/fpsyg.2017.01282>

References (II)

- Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *The 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* [online]. Minneapolis, Minnesota, June 2019, pp. 4171-4186. [Accessed 10 November 2020]. Available at: <https://www.aclweb.org/anthology/N19-1423.pdf>

- Hung, B.T., Ngueng, L.M. and Shimazu, A. (2012) Sentence Splitting for Vietnamese-English Machine Translation. *IEEE Fourth International Conference on Knowledge and Systems Engineering* [online]. Danang, 17-19 August 2012, pp.156-160. [Accessed 10 November 2020]. Available at: < https://ieeexplore.ieee.org/document/6299413>

- Kittredge, R. I. (2016). 'Sublanguages and controlled languages' in *R. Mitkov (ed.) The Oxford Handbook of Computational Linguistics (2 ed.)* [online] Oxford University Press, 2015, pp. 1-26. [Accessed 11 November 2020]. Available at: <DOI: 10.1093/oxfordhb/9780199573691.013.015>

- Mitkov, R. and Schmidt, P. (1998) On the Complexity of Pronominal Anaphora Resolution in Machine Translation. *C. Martin-Vide (ed.) Mathematical and Computational Analysis of Natural Language: Selected papers from the 2nd International Conference on Mathematical Linguistics*. Tarragona, 1996. Amsterdam: John Benjamins, pp. 207-222

-

References (III)

- Ngueng, Q.P., Vo, A.D., Shin, J.C. and Ock, C.Y. (2018) Effect of Word Sense Disambiguation on Neural Machine Translation: A Case Study in Korean. *IEEE Access* [online]. 6, pp.38512-38523. [Accessed 10 November 2020]. Available at: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8399736>
- O’Brien (2005) Methodologies for Measuring the Correlations Between Post-Editing Effort and Machine Translatability. *Machine Translation* [online]. 19(1), pp. 37-58. [Accessed 10 November 2020]. Available at: <https://link.springer.com/article/10.1007/s10590-005-2467-1>
- O’Brien, S. (2006) Pauses as Indicators of Cognitive Effort in Post-Editing Machine Translation Output. *Across Languages and Cultures* 7(1), pp. 1-21. [Accessed 10 November 2020]. Available at: <https://www.researchgate.net/publication/250006640_Pauses_as_Indicators_of_Cognitive_Effort_in_Post-editing_Machine_Translation_Output>
- Popović, M. and Arčan, M. (2015) Identifying main obstacles for statistical machine translation of morphologically rich South Slavic languages. *EAMT 2015: 18th Annual Conference of the European Association for Machine Translation* [online]. Antalya, 11-13 May 2015, pp.97-104. [Accessed 10 November 2020]. Available at: <https://www.aclweb.org/anthology/W15-4913.pdf>
- Tiedemann, J. (2012) Parallel Data, Tools and Interfaces in OPUS. *8th International Conference on Language Resources and Evaluation*, Istanbul, 23-25 May 2012. Istanbul: European Language Resources Association (ELRA), pp. 2214-2218

References (IV)

- Underwood, N.L. and Jongejan, B. (2001) Translatability Checker: A Tool to Help Decide Whether to Use MT. *Maegaard, B. (ed.) Proceedings of MT Summit VIII* [online]. Santiago de Compostela, Spain, 18-22 September 2001, pp. 363-368. [Accessed 10 November 2020]. Available at: <http://www.mt-archive.info/MTS-2001-Underwood.pdf>