Comparing Formulaic Language in Human and Machine Translation: Insight from a Parliamentary Corpus

Yves Bestgen

Laboratoire d'analyse statistique des textes
Statistical Analysis of Text Laboratory (SATLab)
UCLouvain

ParlaCLARIN III @ LREC2022
Introduction

• Neural machine translation systems
  • Bridging the gap between human and machine translation
    Wu et al. 2016, Popel et al. 2020

• Little research about the processing of multiword units (MWU)
  Monti et al. 2018, Zaninello et al. 2020

• Unfortunate because of the importance of MWU in language use
  • Including in Foreign Language Learning and Translation
    Baker 2007, Sinclair 1991
A Recent Study

Comparing multiword units in human and machine translation.

- Focused on a specific type of multiword units

**Formulaic sequences (FS)**
*Habitually occurring lexical combinations* (Laufer & Waldman 2011)

- traffic jam, wide range, very good, dramatic increase, depend on,
  by the way, as far as I know

- Showed that neural machine translations contain **fewer strongly-associated** formulaic sequences
  made of **relatively rare words**

  self-fulfilling prophecy, sparsely populated, sunnier climes
The Usage-Based explanation

• Similar difference in foreign language learning
• Both results could be explained by
  • The usage-based model of language learning

A major determining force in the acquisition of formulas is the frequency of occurrence and co-occurrence of linguistic forms in the input

Durrant & Schmitt 2009

• Since frequency of use also affects neural models

Koehn & Knowles 2016, Li et al. 2020
A Competing Explanation

• The previous study was based on quality newspaper articles

Translation of news implies a higher degree of re-writing and re-telling than in any other type of translation

Ponomarenko 2019

• Less literal translation than that expected from a machine system
A Competing Explanation

• The previous study was based on quality newspaper articles

Translation of news implies a higher degree of re-writing and re-telling than in any other type of translation  

Ponomarenko 2019

• Less literal translation than that expected from a machine system

• Parliamentary corpora as an answer

• Translation accuracy is the main objective

The target text is a faithful, accurate and consistent translation of the source text

Sosoni 2011
Study Aim

To determine whether machine translations of parliamentary texts differ from human translations in the use of phraseology.

- Hypothesis
  - Fewer strongly-associated formulaic sequences made of relatively rare words in neural machine translations
Method

• Translation corpora (from French to English)
  • Parliamentary corpus
    • Preprocessed version of the *Europarl* corpus
      by the *EuroparlExtract* toolkit
      Koehn 2005; Ustaszewski 2019
    • 200 randomly selected speeches (± 120,000 words)
Method

• Translation corpora (from French to English)
  • PLECI corpus: Newspaper articles
    • 279 articles (± 500,000 words)
Method

• Translation corpora (from French to English)
  • Parliamentary corpus
  • PLECI corpus: Newspaper articles

• Three well-known neural machine translation (NMT) systems
Method

• Formulaic sequences analysis
  • CollGram
    Bernardini 2007, Durrant & Schmitt 2009, Bestgen & Granger 2014
    • A technique for scoring texts on formulaicity by means of
      - Lexical association indices for identifying COLLocation
      - Applied to the word biGRAMs

This talk is focused on one indice (more in the paper)
CollGram

• How does CollGram work?
  • Word bigrams are first extracted from the text to be evaluated
  • A native reference corpus is then necessary

• For this study, I used the BNC (www.natcorp.ox.ac.uk)
CollGram

- How does CollGram work?
  - Word bigrams are first extracted from the text to be evaluated
  - A native reference corpus is then necessary
    - To calculate for each bigram in the text
      - Its Mutual Information (MI)

MI Identifies mainly low-frequency, but strongly-associated, FSs

- asylum seekers, democratically elected, blatant violation, vast majority,
  - publicly denounce, money laundering, subsidiarity principle,
  - utmost importance, hardly surprising, left unsaid, takes precedence

(Examples from the Parliamentary corpus)
...hardly surprising...
...hardly surprising...
CollGram

Text

...hardly surprising...

BNC reference corpus MI list

...hardly shown -0.5
...hardly surprising 10.4
...hardly what -0.3
...hardly_surprising_10.4...
CollGram

• How does CollGram work?
  • Word bigrams are first extracted from the text to be evaluated
  • A native reference corpus is then necessary
    - To calculate for each bigram in the text
      - Its Mutual Information (MI)
  • Categorization of each bigram according to its collocation intensity
    - High MI if MI ≥ 5
      Durrant & Schmitt 2009, Bestgen 2018
  • Output
    - Percentage of highly collocational bigrams (for the MI) in the text
Results

• Mean percentages of highly collocational bigrams for MI
Results

• Mean percentages of highly collocational bigrams for MI

Statistically significant differences between Human and NMT Medium effect sizes (Cohen's d)
Results

• Mean percentages of highly collocational bigrams for MI

*Google translations contain fewer highly collocational bigrams for MI*
Results

• Mean percentages of highly collocational bigrams for MI

Similar trends in the two corpora
But the effect sizes are larger in the news corpus
Results

• Average effect size of the difference between human and neural machine translations

| Effect Size | Parliamentary | News |
|-------------|---------------|------|
| 0.25        |               |      |
| 0.5         |               |      |
| 0.75        |               |      |
| 1           |               |      |

Important effect (According to Cohen's criteria)

Medium effect (According to Cohen's criteria)
• This study replicates the news corpus study

• However, the differences are smaller in the parliamentary corpus, which seems better suited to compare human and machine translation
  
  • The less literal nature of the translations in news favors the identification of differences between human and machine translation
Further Works

• (Many) more language pairs
  • Much easier with parliamentary corpora

• Using a genre-specific reference corpus
  • Much easier with parliamentary corpora
Thank you for your attention
Main references

Baker, M. (2007). Patterns of idiomaticity in translated vs. non-translated text. *Belgian Journal of Linguistics* 21, 11–21.

Bernardini, S. (2007). Collocations in translated language. combining parallel, comparable and reference corpora. In Proceedings of the *Corpus Linguistics Conference*. pp. 1–16.

Bestgen, Y., Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing* 26, 28–41.

Durrant, P., Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching* 47, 157–177.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In Conference Proceedings: the tenth Machine Translation Summit, pages 79–86, AAMT.

Koehn, P., Knowles, R. (2017). Six challenges for neural machine translation.

Laufer, B., Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners’ English. *Language Learning* 61, 647–672.

Monti et al. (2018). Multiword units in machine translation and translation technology. In Mitkov et al. (eds.) *Multiword Units in Machine Translation and Translation Technology*, pp. 2–37. John Benjamins.

Ponomarenko, L. (2019). Translating identities in multilingual news. Ph.D. thesis, Universitat Autonoma de Barcelona.

Popel et al. (2020). Transforming machine translation: A deep learning system reaches news translation quality comparable to human professionals. *Nature Communications* 11, 1–15.

Sosoni, V. (2011). Training translators to work for the EU institutions: luxury or necessity? *The Journal of Specialised Translation*, 16:77–108.

Ustaszewski, M. (2019). Optimising the Europarl corpus for translation studies with the Europarl extract toolkit. *Perspectives*, 27:107–123.
d = 0.30
$d = 0.50$
$d = 0.84$
The lexical Association Indices

• Based on the frequencies in a reference corpus

| First word | Second word | (other) | Total |
|------------|-------------|---------|-------|
| larger     | a           | b       | a + b |
| (other)    | c           | d       | c + d |
| Total      | a + c       | b + d   | a + b + c + d |

• "a" is the Observed frequency of the bigram *larger than*

• The Expected frequency of the bigram *larger than* is

\[(a + c) \times (a + b) / (a + b + c + d)\]
The Lexical Association Indices

• Mutual information and t-score

\[ MI = \log_2 \left( \frac{O}{E} \right) \quad t = \frac{O - E}{\sqrt{O}} \]

Smaller the E, larger the MI *
E is small when the frequency of both words is small

Larger the O, larger t should be*
To be large, O needs that the frequency of both words is large

* Everything else being equal
Study Aim

To determine whether machine translations of parliamentary texts differ from human translations in the use of phraseology

• Hypothesis
  • Fewer strongly-associated formulaic sequences made of relatively rare words in neural machine translations

• Outcome
  • A positive conclusion will suggest a way to make machine translations more similar to human translations
