Effect of translationese on machine translation quality

Stephanie Chen
schen751@stanford.edu

CS224N Winter 2019

Background + Problem

- **Translated texts are structurally different from untranslated texts:** result of balancing fidelity to source language, fluency in target language
- "Translationese" features include higher use of common words, part-of-speech structures from the source language, pronoun frequency
- Research[1] shows that **language models (LM) trained on translated text outperform LMs trained on original target-language text** in statistical machine translation (MT)
- But most research, corpora, and benchmarks ignore translation direction
- In this project we aim to show the **wide impact of translationese across model architectures and languages**

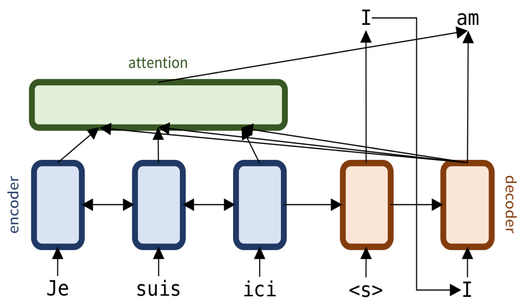
Approach

- **Supervised phrase-based statistical MT (PBSMT)**
  - Translation model *p(s|t),* language model *p(t)*
  - Train translation model without regard for direction
  - Train one language model on source-to-target translated corpus (T-L), one language model on original target-language corpus (O-L)
- **Supervised seq2seq neural MT (NMT)**
  - Single-layer biLSTM encoder-decoder with attention
  - Train one model end-to-end with T-L corpus, one end-to-end with O-L corpus
- **Unsupervised PBSMT**
  - Bootstrap PBSMT model using monolingual (non-parallel) corpora, iteratively backtranslate to learn[2]
  - Train one model with T-L target corpus, one with O-L target corpus

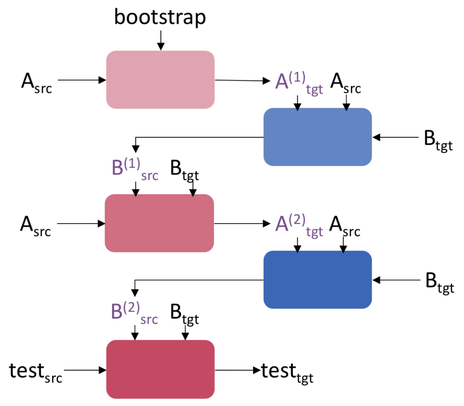
[1] Lembersky, Ordan, & Wintner. Language Models for Machine Translation: Original vs. Translated Texts. 2011.
[2] Lample, Ott, et al. Phrase-Based & Neural Unsupervised Machine Translation. 2018.

Data

- Europarl directed corpora from French, German, Italian, Dutch, and Romanian to English
- **High-resource languages:** fr, de
  - 140k sentences parallel, 200k T-L, 370k O-L, 115k monolingual
- **Medium-resource languages:** it, nl
  - 100k parallel, 100k T-L, 370k O-L, 50k monolingual
- **Low-resource language:** ro
  - 90k parallel, 12k T-L, 80k O-L, 7k monolingual

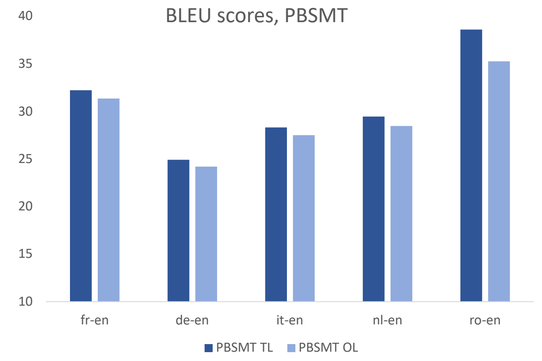


Top: NMT encoder-decoder attention architecture. **Bottom:** Iterative backtranslation for unsupervised translation using monolingual corpora.

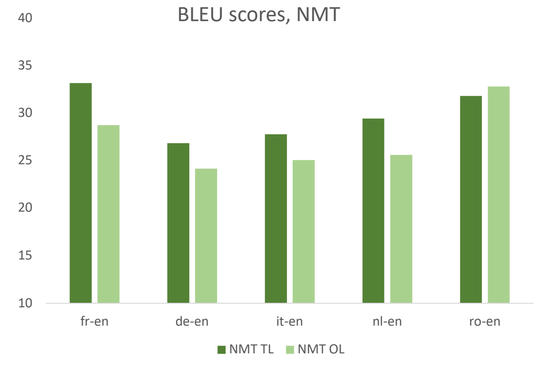


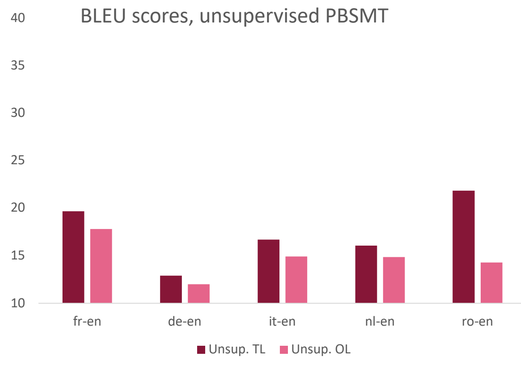
Results + Conclusions

- **T-L models outperform O-L models in all cases***
  *with the exception of Romanian NMT, possibly due to corpus size issues
- Results **consistent even at different degrees of influence of the target corpus** (isolated to LM in PBSMT, end-to-end in NMT, non-parallel to source in unsupervised)
- Results **consistent across corpus size**, with biggest improvements in low-resource PBSMT (both sup. and unsup.)
- Practical implications
  - Augmenting corpora for low-resource translation with translationese in related languages
  - Translation direction of training & test corpora matters!



Graphs: BLEU scores for the three systems. **Example:** Sample output from PBSMT and NMT systems; note differences in translation of register and tense.





Example

French: Donc, j'ai dit un peu l'inverse de ce que vous venez de dire.

Reference: So I, in fact, said practically the opposite of what you have just said.

PBSMT T-L: Therefore, I said a little the opposite of what you have just said.

PBSMT O-L: Therefore, I have said a bit the opposite of what you have just said.

Neural T-L: I therefore said somewhat the opposite of what you have just said.

Neural O-L: So I said a little bit of what you have just said.