Unsupervised Machine Translation

Ivana Kvapilíková

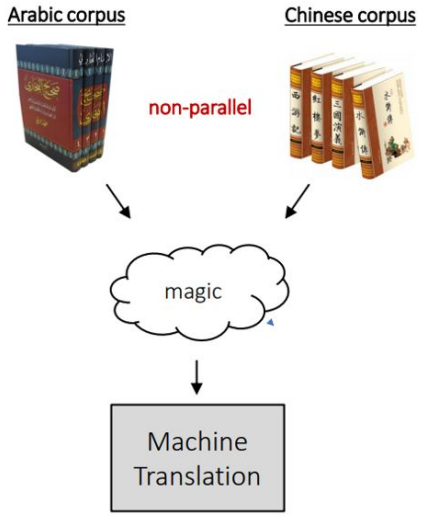
September 17, 2019

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Introduction

- 2nd year PhD student
- **Dissertation topic:** Machine Translation from Monolingual Data
- **Supervisor:** Ondřej Bojar
- **Projects:** GAUK, NEUREM
- **Research visit** to the University of the Basque Country (Oct–Nov 2019)

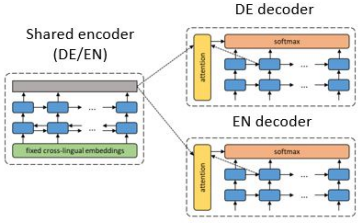
Unsupervised Machine Translation



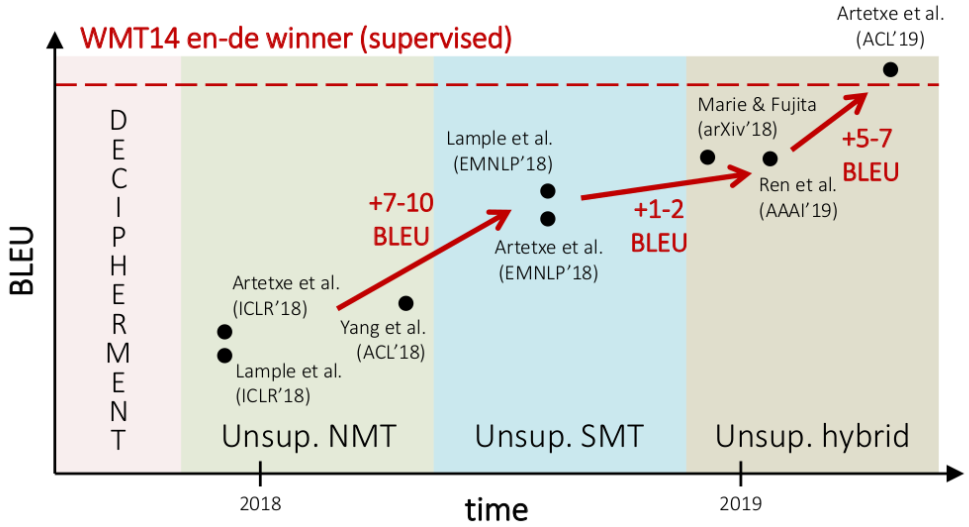
Statistical (phrase-based) model

DE	EN	p
den Vorschlag	the proposal	0.6227
den Vorschlag	's proposal	0.106
den Vorschlag	a proposal	0.0341
den Vorschlag	the idea	0.025

Neural model

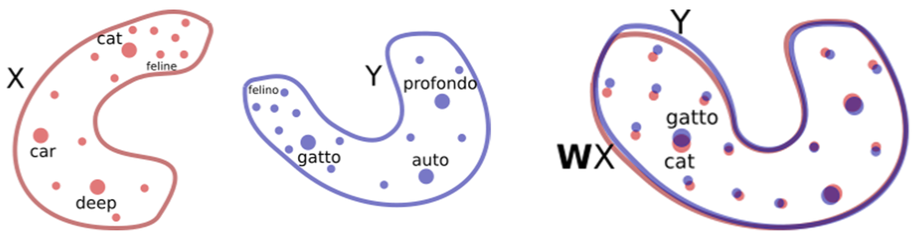


Phrase-based vs. Neural Models



Unsupervised Cross-lingual Embeddings

- Individual training on monolingual corpora
  - word/phrase embeddings (Word2Vec) + Mapping



- Joint training on concatenated/merged corpora
  - subword embeddings (fastText)
  - cross-lingual language model pre-training

Research Plan

- Use a **pre-trained cross-lingual language model (XLM)** and fine-tune for unsupervised NMT by denoising and back-translation (Lample et al., 2019)
- **Analyze cross-lingual embeddings** from the pre-trained model before and after fine-tuning
- Can **cross-lingual mapping** be applied to contextualized embeddings (XLM embeddings)?
  Let's try out...

Thank you for your attention!

References

- Mikel Artetxe, Gorka Labaka, Eneko Agirre. *An Effective Approach to Unsupervised Machine Translation*. ACL 2019.
- Mikel Artetxe. *Unsupervised Machine Translation*. MTM 2019.
- Guillaume Lample, Alexis Conneau *Cross-lingual Language Model Pretraining*. 2019
- Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, Marc'Aurelio Ranzato. *Phrase-Based & Neural Unsupervised Machine Translation*. EMNLP 2018.