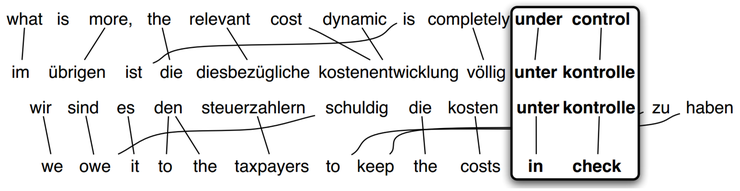
Paraphrase Extraction with Neural Machine Translation

CMPT 726 - Machine Learning, Fall 2016

Contribution

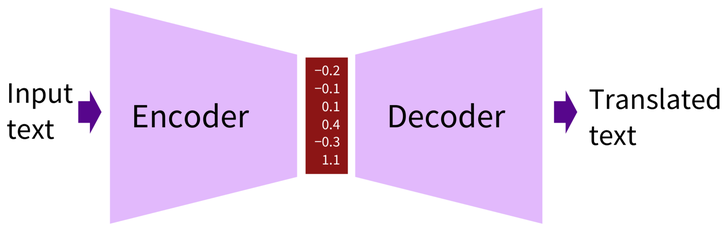
- Paraphrase extraction was performed in [2] using a pivoting method on bilingual parallel corpus
- We propose to use neural machine translation (NMT) with soft attention model to artificially create parallel corpus, and extract paraphrases from it using pivoting method
- Paraphrases can in turn can help improve the performance of Neural Machine Translation systems by addressing the rare word problem in machine translation.



Bilingual pivoting across parallel corpora used for paraphrase extraction. Figure from [3]

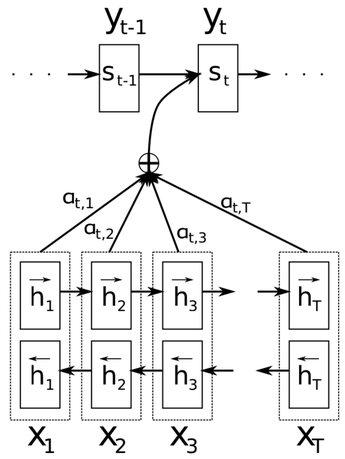
NMT Encoder-Decoder with Attention

- Typical NMT systems have recurrent neural networks (LSTMs or GRUs in various configurations) as encoder and decoder
- A source sentence is encoded into a fixed length vector which is then decoded in steps by the decoder



Soft Alignments

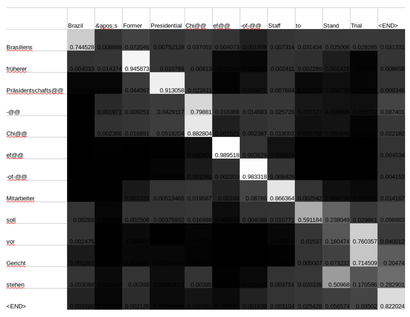
- Soft alignment allows better encoding of the source sentence
- Decoder uses a weighted sum of encoded word vectors, weighing functions are learnt during training



Soft alignment model learns to give high weight-age for relevant input hidden vectors for decoding each word. Figure from [1]

Algorithm

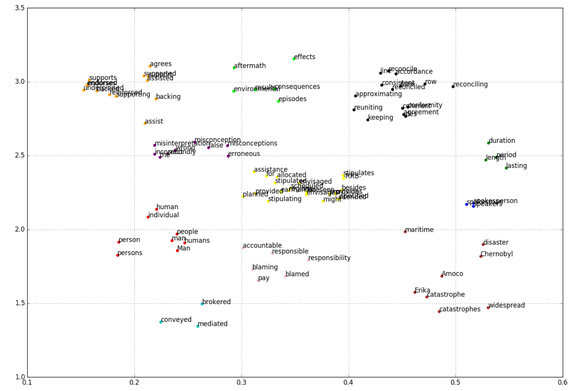
- Translate source corpus to a foreign language
- Collect repeatedly occurring words/phrases in the foreign language
- Map back to the source side through alignment matrix to extract source words



Example of an alignment matrix

Experiments and results

- We use an implementation of [4] for translation and first 100k sentences from English-German corpus of Europarl version 7 for experiments.



t-SNE plot of paraphrastic words based on similarity

Evaluation

- Word2vec is used to get vector representation of words to calculate cosine simailarity and Word Mover's Distance.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv*, 2014.

[2] Colin Bannard and Chris Callison-Burch. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[3] Chris Callison-Burch, Philipp Koehn, and Miles Osborne. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24, New York City, USA, June 2006. Association for Computational Linguistics.

[4] Rico Sennrich, Barry Haddow, and Alexandra Birch. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of First Conference on Machine Translation (WMT2016)*, 2016.

This work was done under the supervision of Dr. Anoop Sarkar