Topics & components

- Comparison of low-resource approaches to Turkish-English translation
- Building and comparing four models:
  - Baseline Turkish-English model
  - Transfer learning from a high-resource pair (Zoph et. Al 2016)
  - Transfer learning from related low-resource pairs (Nguyen & Chiang 2017)
  - Iterative back-translation with monolingual Turkish and English data (Hoang et. Al 2018)

Background research

From Professor Richardson’s lecture slides:

- *Transfer Learning for Low-Resource Neural Machine Translation*
- *Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation*
- *Iterative Back-Translation for Neural Machine Translation*
- Papers and videos referenced in the datadan.io blog post on low-resource NMT

Code, data, and corpora

To implement and compare these models, I will need to write four different translation pipelines using OpenNMT-py:

1. The baseline Turkish-English model with no data augmentation
2. The transfer model using a high-resource but unrelated language (possibly Spanish, Portuguese, or Japanese)
3. The transfer model using similar low-resource languages (possibly Azerbaijani or Uzbek)
4. The iterative back-translation model with monolingual Turkish and English data

For all English-Turkish and high-resource English-L2 data, I will use the Church data. For Azerbaijani or Uzbek data, I will use OPUS corpus data and/or the en-az-parallel-corpus. Finally, for the Turkish and English monolingual data, I will use Wikimedia dumps.

Evaluation

- BLEU and METEOR scores on a test set of 2,500 parallel Turkish-English sentences
- Human-evaluated rankings
  - 50 randomly-selected sentences
  - 4 translators (including myself)
  - Custom web app to rank sentences

Expected outcome

I became interested in low-resource approaches after finding that the translations with a baseline model trained on ∼60,000 sentence pairs produced quite bad translations.

Although I haven't had time to carefully review the various approaches, a brief review of the approaches and their results leads me to believe that back-translation will be the most effective approach.

I anticipate that the similar low-resource transfer approach will be the weakest, because the lack of data is simply too great a restriction. It also results in the smallest BLEU score improvements in the article outlining the results.