CIS 526: Project Proposal
Extracting parallel sentences from Wikipedia

Anshul Sharma

Extracting Parallel Sentences: Challenge Problem 5

State of the art Statistical Machine Translation (SMT) systems greatly depend upon training corpora consisting of pairs of sentences in the languages being translated between. Each pair consists of a sentence in a particular language A, and its equivalent translation in language B.

However, corpora containing such pairs remain few and far between and manual sentence correlation is not an efficient option. Therefore, **this challenge deals with extracting pairs of sentences given a corpus of Wikipedia articles in two different languages**. The corpus will consist of pairs of articles on the same subject in different languages, and the task will be to automatically find pairs of mutually translatable sentences.

Getting Started

The default directory would contain a script (*align.sh*) which will use training data from *data/de/train* to compute alignments on the EuroParl corpus.

./align.sh

This generates an *alignment.txt* file which is used to extract parallel sentences from the data as follows:

python parallel

This generates another file called *output.txt* which contains the parallel sen-

tences chosen for the training data. The contents of this file are then evaluated by using:

python grade < output.txt

The grade function consists of an implementation of the *BLEU* metric which will score the candidate sentences according to their similarity with the reference sentences.
You should submit a file containing parallel english sentences chosen for all the sentences in *data/de/test*. This can be achieved by using:

python parallel -d ../data/de/test

The Challenge

Given a noisy parallel/comparable corpus of articles from Wikipedia, and a pair of target languages, the challenge is to extract pairs of parallel sentences in the given languages such that each extracted pair is the best mutual translation/representation of the sentences according to the respective languages. This involves maximizing the average BLEU score computed using the selected parallel sentence for a given sentence and the reference sentence.

The problem can be solved in multiple ways, including techniques from:

[1] Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. *Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment.*

[2] Magdalena Plamada, Martin Volk. 2013. *Mining for Domain-specic Parallel Text from Wikipedia.*

[3] Sissay Fissaha Adafre, Maartin de Rijke. 2006. *Finding Similar Sentences across Multiple Languages in Wikipedia.*

[4] Pascale Fung, Percy Cheung. 2004. *Multi-level Bootstrapping for Extracting Parallel Sentences from a Quasi-Comparable Corpus.*

A number of features can also be used to improve the selection of parallel sentences:

- *The number of phrasal similarities.*
Given a mapping between different phrases in the target languages, it can be used to compute the number of phrases in common between the two sentences.

- *Using a SMT model*
Given a Machine Translation model, which can translate *language A* into *language B*, the model can be used to translate all *language A* sentences in the extracted pairs to *language B*. Then, a similarity metric such as BLEU can be used to measure the similarity between the pair of sentences, both represented in *language B*.