CIS 526 Homework 5: Extracting Parallel Sentences from Wikipedia

James Yang

April 28, 2015

1 Introduction

Automation of data collection, especially the extraction of parallel sentences in diverse corpora, is a highly sought-after goal in machine translation. Wikipedia is proven to be a very valuable source of data, where certain articles are written in dozens of different languages. Evaluating the quality of translations between each language, however, can prove to be a difficult task since Wikipedia articles in different languages are not direct translations of each other. Articles are often paraphrased, incomplete, or sparsely translated, and as a result of fundamental differences between languages, direct translations do not necessarily exist. In this homework, we were tasked with extracting parallel sentences from multiple Wikipedia articles by evaluating the likelihood that a source sentence is a direct translation of a target sentence based on features that may indicate a good quality translation.

2 Scoring Metric

Given a source sentence $s \in S$, we want to find $f \in F$ such that

$$(s, f) = \underset{s' \in S, f' \in F}{\operatorname{argmax}} \left(\text{Score}(s', f')\right)$$

The goal is to find the proper scoring function $\text{Score}(s, f)$ that provides sentence pairs that maximize the F-Score which compares the precision and recall of the evaluation given a ground truth. The F-Score function is given by

$$\text{F-Score} = 2\frac{P(h,e) \cdot R(h,e)}{P(h,e) + R(h,e)}$$

where P and R are precision and recall defined as:

$$R(h,e) = \frac{|h \cap e|}{|e|} \quad , \quad P(h,e) = \frac{h \cap e|}{|h|}$$

where h is a hypothesis sentence for a given source sentence e.

It is assumed that any of the source sentences can be translated to any of the target sentences, regardless of location of either sentence in their respective articles. Therefore, distance of a target sentence from a source sentence for an article and its respective translation is not penalized.

3 Features

The following features are used to evaluate the translation: sentence length ratio, translation overlap, and proper noun presence.

Sentence length ratio is given by

$$r_{s,f} = \frac{|f|}{|s|}$$

Translation overlap $o_{s,e}$ is given by translating the target hypothesis and counting the number of unique intersections between words in the source sentence and hypothesis sentence

$$o_{s,f} = |s \cap T(f)|$$

where $T(x)$ is the word-level translation of target sentence x to the source language. Stop words such as "and", "but", "or", "with", etc. were highly prevalent throughout documents. Due to general lack of influence these words have on the translation of a sentence, these words were entirely rejected in the translation overlap count.

Finally, a library of proper nouns is tracked, where the presence of a proper noun $w_p \in P$ is weighted.

$$\delta_{w_p}(s, f) = \begin{cases} 1 & , \quad \text{if } w_p \in s \text{ and } w_p \in f \\ 0 & , \quad \text{otherwise} \end{cases}$$

The final scoring function is given by the following:

$$\text{Score}(s, f) = (1 - k_l + |r_{s,f}|) \left(o_{s,f} + k_p \delta_{w_p}(s, f)\right)$$

where k_l and k_p are empirically determined values.

A pairing is considered valid if $\text{Score}(s, f) > k_t$ where k_t is also an empirically determined value.

4 Results

The following is a table of the empirically determined values for the constants used in evaluating the probability of a parallel sentence.

Constants	**Value**
k_l	1.18
k_p	2.5
k_t	0.2

Table 1: Table of empirical constants.

These values gave a final F-Score of 0.6949.