Project Proposal

ON

Machine Translation of Noisy Text (MTNT)

Stevens Institute of Technology

Project Guide:

Professor Abdul Rafae Khan

Submitted by:

Janmejay Mohanty

jmohanty@stevens.edu

Course Name:

CS 541-A Artificial Intelligence

Abstract

In most modern Machine Translation (MT) systems, noisy or non-standard input text can result in devastating mistranslations, and there has been a growing research interest in developing noise-resistant MT systems. However, there are currently no publicly available parallel corpora of naturally occurring noisy inputs and translations, therefore past research has had to rely on synthetically constructed datasets for evaluation. The author presents a benchmark dataset for Machine Translation of Noisy Text (MTNT) in this study, which includes noisy Reddit comments and professionally supplied translations. On the huge number of sentences per language pair, we commissioned translations of English comments into French and Japanese, as well as French and Japanese comments into English. We look at the many types of noise in this dataset both qualitatively and quantitatively.

Git-hub Link:

https://github.com/Janmejay1998/mtnt

Data

For each different language, we are selecting a set of groups ("subreddits") that we know contains many comments in that language such as English: Since a huge majority of the discussions on Reddit took place in the medium of English, so we don't restrict our collection to any community in particular.

English: As huge majority of the discussions on Reddit are conducted in English, we are not restricting our collection to any specific community.

French: The total of the French side of the parallel training data which is provided for the English-French WMT 2015 translation procedure. This amount is to be approximately 40.86 million sentences.

Japanese: We had collected three small/medium sized MT datasets: KFTT (Neubig, 2011), JESC (Pryzant et al.) and TED talks (Cettolo et al., 2012), amounting to be approximately ≈ 4.19 million sentences.

The data files for MTNT can be found in this link:
https://github.com/Janmejay1998/mtnt/tree/master/resources

Most of the data we are dealing with are from Reddit based data.