Evaluating Adequacy of MT Output without Reference

Yashar Mehdad, Marcello Federico, Daniele Pighin, Hanna Bechara, Angeliki Lazaridou, Nikos Engonopoulos, Antonio Valerio Miceli Barone

MT Marathon 2011 – Trento, Italy

Discussed Issues

- The difficulties of the task.
- The problem with the dataset (WMT data: only few hundred pairs available, L. Specia dataset: scores are not focused on adequacy).
- Processing tools in different languages are not very robust for noisy data (translation output).
- The difficulty of drawing the border line between quality and adequacy
- Feature sets
- Learning algorithms

Proposed Solution

- Using L. Specia's dataset.
- Extracting the global features that can not be implemented in MT decoders.
- Extracting features in different levels: surface, lexical, syntactic (including shallow), semantic (possibly)
- Classifiers: we start using SVM, we can try using different algorithms.
- We will focus on binary classification

Accomplished Tasks

- Data: Daniele and Yashar
- Hanna: surface based features
  - Length, punctuations, numbers, oov, …
- Yashar: Shallow syntactic and dependency features
  - POS : Adj, Adv, Card, Conj, Dt, Pro, Prep, Verb, F
  - Dep: Adjn, Cprep, Dobj, Root, Subj

- Eleftherios Avramidis (DFKI) proposed to help us and he already sent us the data and some features he used for his recent work about CE.

Tasks to be completed

- Angeliki: Multilingual topic modeling
- Nikos: WSD
- Antonio: Statistical Parsing
- SRL
- Deeper syntax
- Lexical features

Preliminary Results - I

- Dataset: 16K pairs (source: L. Specia)
- ~7.5k: good quality, ~8.5k: bad quality
- ~50 features
- Binary classification

Alg.	**Accuracy**
Logistics	65%
Perceptron	64%
SVM	66%

Preliminary Results - II

- Dataset: 16K pairs (source: L. Specia)
- ~4k , 5k, 6k, 1.5k: 1,2,3&4 score for quality
- ~50 features
- Multiclass classification

Alg.	**F1**
Logistics	46%
Perceptron	43%
SVM	43%

Conclusion & Future Work

- The preliminary experiments shows the difficulty of the task.
- A framework to continue working in this direction.
- Need more investigation in this direction.
- Lack of dataset.

Future:

- Feature tuning and selection.
- Adding more relevant features.
- Different learning strategies.
- Using more data.