LIMSI Submission for WMT'14 QE Task

Guillaume Wisniewski[1,2] Nicolas Pécheux[1,2] Alexandre Allauzen[1,2] François Yvon[1]
(1) LIMSI–CNRS, Orsay France
(2) Université Paris Sud, Orsay France

Highlights

- Participation to the word-level quality estimation task for English to Spanish translations (binary condition)
- Use of 16 'dense' features ⊕ binary classifier trained with a specific method to optimize the f_1 score

Features

- 16 'dense' features (no lexicalized information)
- Three main classes of features

Association Features derived either from **IBM 1 scores** (max, arithmetic mean, geometirc mean, ...) or from **pseudo-references** (e.g. target word in the pseudo-reference)

Fluency Features

3 different language models

- a 'traditional' 4-gram LM
- a continuous-space 10-gram LM
- a 4-gram LM based on POS

3 different kinds of features

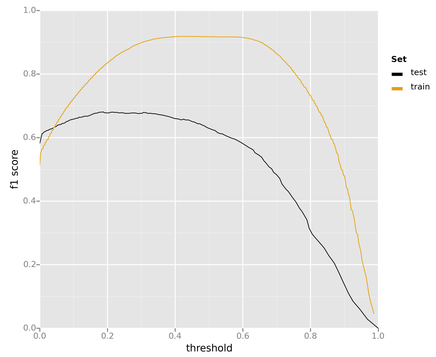
- probability of the word given the history
- ratio with the 'best' probability that can be achieved by replacing this word
- features to describe back-off behavior

Prior probability

- Most useful features: language models + prior probability

Learning Strategy

- The QE task can naturally be framed as a binary classification problem
- Logistic Regression and Random Forest as 'base' classifier
- Optimization of the f_1 score on the train set:
  1. train a classifier
  2. enumerate all possible trade-offs between recall and precision by varying the threshold of the decision function ($O(n \cdot \log n)$)
  3. find the trade-off with the optimal f_1 score ($O(n)$)



Experiments

- Experiments on a internal test set made of 200 sentences
- Overall performance is not good enough to consider the use of such a system in a real-word application
- Results on the official test set is much worse

Prediction performance for the two learning strategies considered

Classifier	thres.	$\text{recall}_{\text{BAD}}$	$\text{precision}_{\text{BAD}}$	f_1 score
Random forest	0.43	0.64	0.69	0.67
Logistic regression	0.27	0.51	0.72	0.59

Failure Analysis

- 2 kinds of information
  - Compute the score for each POS
  - 1st baseline: choose the label randomly
  - 2nd baseline: always predict BAD
- We are better at predicting the 'quality' of plain words

	Random Forest	Random Classifier	Always BAD
VERB	0.73	0.45 +0.28	0.58 +0.15
ADJ	0.70	0.42 +0.28	0.53 +0.17
NOUN	0.69	0.41 +0.28	0.52 +0.17
ADV	0.69	0.42 +0.27	0.54 +0.15
PRON	0.72	0.46 +0.26	0.60 +0.12
overall	0.67	0.41 +0.26	0.52 +0.15
DET	0.62	0.40 +0.22	0.49 +0.13
PUNCT	0.56	0.35 +0.21	0.43 +0.13
ADP	0.61	0.42 +0.19	0.52 +0.09
CONJ	0.57	0.38 +0.19	0.47 +0.10

What we Have Learned

- Predicting confidence at the word level is hard
- Need for more information about preprocessing and annotation convention
- Difficult to interpret results

Acknowledgments

This work was partly supported by ANR project Transread (ANR-12-CORD-0015). Warm thanks to Quoc Khanh Do for his help for training a SOUL model for Spanish.