Evaluate with Confidence Estimation: Machine ranking of translation outputs using grammatical features

Eleftherios Avramidis, Maja Popovic, David Vilar, Aljoscha Burchardt - name.surname@dfki.de
German Research Center for Artificial Intelligence (DFKI) - Language Technology (LT), Berlin, Germany

Automatic Ranking

Confidence estimation: Evaluate MT with no reference

Our goal: given one source sentence and its translations produced by different MT systems, order (aka rank) the translations from best to worst

Machine Ranking idea:
- Apply machine learning to immitate human evaluation task
- Train a classifier to perform **ranking** on several MT outputs on a sentence level
- Include statistical features of grammatical analysis

Pairwise Decomposition

Ranking is a **result of comparing** MT outputs with each other

We build **one classifier** over all training data, operating on a **binary level:**
- given feature sets from 2 sentence outputs at a time
- decide whether the first output is better than the second

The *n* systems' outputs for each sentence are broken down to *n* × (*n* − 1) pairs, of all possible comparisons between two system outputs, in both directions

Class value is determined as $c(r_i, r_j) = \begin{cases} 1 & r_i < r_j \\ -1 & r_i > r_j \end{cases}$

for the pairwise comparison of systems *i,j* with system outputs *t* and respective ranks *r*

Features

Sentence length: source, target, ratio
Target language model: smoothed 5-gram probability, unigram, bi-gram, tri-gram, count of unknown words
PCFG parsing: parse log likelihood, count of n-best trees, confidence of best parse, avg confidence of all trees, source/target ratios
Shallow grammatical matches: Counts and source/target ratios of NPs, VPs, PPs, NNs and punctuation marks.

Classifiers

Naïve Bayes: relative frequencies of the training pairwise examples. Probability for continues features estimated with LOESS (locally weighted linear regression)

$$p(C, F_1, \ldots, F_n) = p(C) \prod_{n}^{i=1} p(F_i|C)$$

K-nearest neighbour: classification based on the closest training example in the feature space

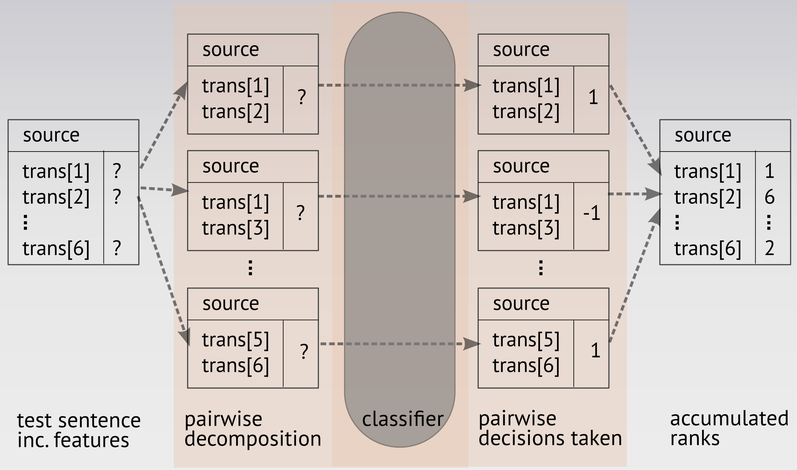

Figure 1: The process of Machine Ranking, performed through pairwise comparisons for 6 system outputs

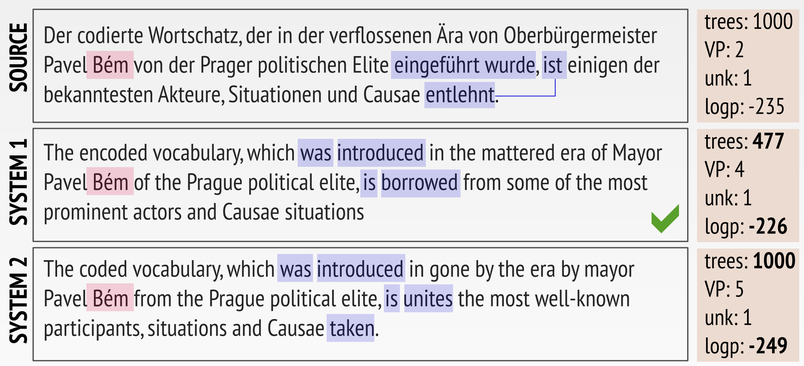

Figure 2: Example of pairwise comparison of two annotated system outputs. Less parsing ambiguity (number of parse trees) and higher parse loglikelihood are if favour of the 1st output

Experiment

Basic experiment (development) German/English
- **training set:** 700 sentences of WMT08, -10.
- **test set**: 184 sent. of WMT10 with human agreement

Extended experiment (comparison)
- **training set:** 1100 sent. of WMT08,-09
- **test set:** the entire WMT10

Results

features	naive Bayes		knn	
	rho	tau	rho	tau
- basic experiment				
ngram	0.19	0.05	0.13	0.01
unk, len	0.67	0.20	0.73	0.24
unk, len, bigram	0.61	0.21	0.74	0.21
unk, len, ngram	0.63	0.19	0.59	0.21
unk, len, trigram	0.67	0.20	0.76	0.21
unk, len, $\log_{parse}$	0.75	0.21	0.74	0.25
unk, len, n_{parse}, VP	0.67	0.24	0.61	0.20
unk, len, n_{parse}, VP, $conf_{bestparse}$	0.78	0.25	0.75	0.24
unk, len, n_{parse}, NP, $conf_{bestparse}$	0.78	0.23	0.74	0.23
unk, len, n_{parse}, VP, $conf_{avg}$	0.75	0.21	0.78	0.23
unk, len, n_{parse}, VP, $conf_{bestparse}$	0.78	0.25	0.75	0.24
unk, len, n_{parse}, VP, $\log_{parse}$	**0.81**	**0.26**	0.75	0.23
- extended experiment				
unk, len, n_{parse}, VP, $\log_{parse}$	**0.60**	**0.23**	0.28	0.02

Kendall's tau: segment level correlation - Spearmann's rho: system level

Discussion

- Correlation comparable wit several reference-aware metrics of WMT10 (e.g. of ρ = 0.47 and τ = 0.12)
- "Tricking" metrics is partially avoided by using source/target ratios. Addition of IBM Model 1 would be useful
- Experimentation with bigger feature sets, more elegant classifiers, various language pairs, could be in further work

This work has been supported by the TaraXÜ Project, financed by TSB Technologiestiftung Berlin – Zukunftsfonds Berlin, co-financed by the European Union – European fund for regional development.