Cross-lingual complex word identification with multi-task learning

Joachim Bingel and Johannes Bjerva

UNIVERSITY OF COPENHAGEN

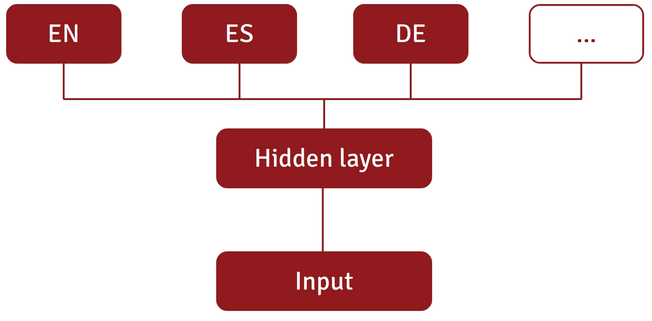
Languages as tasks

Lets us share data between languages

Generalizes better to new languages (without training data!)

Training by alternating between languages

For languages without training data, predict most similar language and use output layer for that language



Model and Features

Ensemble of 10 multitask networks and 10 random forests

Features include length, frequency, character perplexity, semantic specificity (measured by WordNet synsets, hypernyms and hyponyms), inflectional complexity, POS, target-sentence similarity

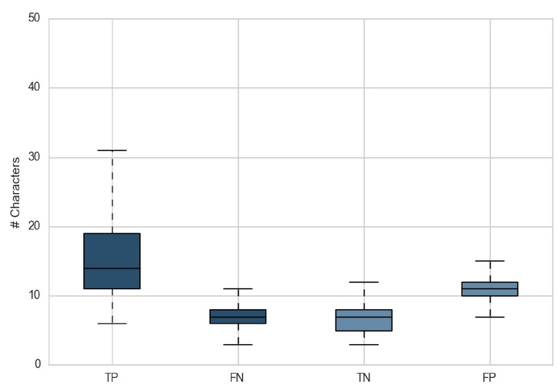
Results

Language	MAE	Rank	Δ (system)	F_1	Rank	Δ (system)
French	0.066	1	0.012 (TMU)	0.7595	1	0.013 (TMU)
German	0.075	2	-0.013 (TMU)	0.6621	5	-0.083 (TMU)
Spanish	0.079	3	-0.007 (TMU)	0.7458	5	-0.024 (TMU)

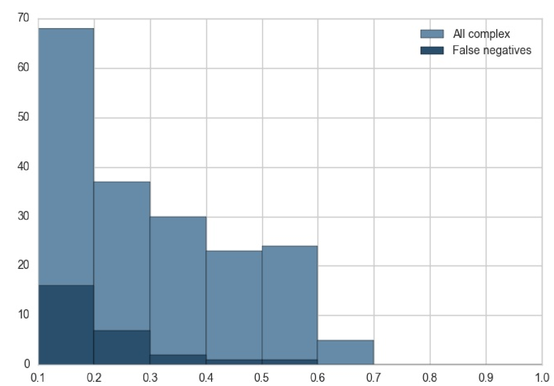
Analysis

Word length good predictor, but false positives tend to be long and false negatives short

False negatives are mostly words rated as complex by very few annotators



Target length in characters per error type



Distribution of false negatives per complexity degree in the German test set

Contact

bingel@di.ku.dk