FACULTY OF ARTS AND PHILOSOPHY

LT³, LANGUAGE AND TRANSLATION TECHNOLOGY TEAM

Sofie Labat and Els Lefever

A CLASSIFICATION-BASED APPROACH TO COGNATE DETECTION COMBINING ORTHOGRAPHIC AND SEMANTIC SIMILARITY INFORMATION

Task description

Cognates = words with high formal and semantic cross-lingual similarity

Cognate Detection = distinguish cognates from non-cognates

Use = automatic alignment, bilingual lexicon compilation, CALL

Goals

- create context-independent en-nl gold standard for classifying cognates and non-cognates
- develop supervised binary classifier with these data, based on orthographic and semantic similarity information

True cognates:

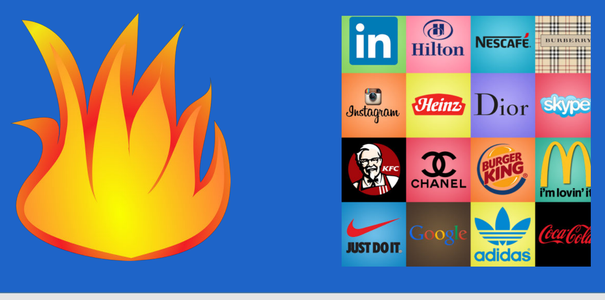
thumb/duim

thunder/donder

False friends:

trap/trap

brand/brand



Data Creation

bilingual list of cognate candidates through statistical word alignment on DPC corpus

500,000 pairs

first selection based on normalised Levenshtein distance (≤ 0.5)

28,503 pairs

manual annotation (cognate, partial cognate, false friend, proper name, error, no standard, etc.) (Labat et al. 2019)

14,619 pairs
9,855 cognates
4,763 non-cognates

Experiments

Machine Learning

- Support Vector Machine
- 5-fold cross-validation
- hyperparameter optimisation

Orthographic similarity Features

- 15 string similarity metrics
- e.g., Dice, Levenshtein, etc.

Semantic similarity Features

- FastText word embeddings
- pre-trained on Wikipedia corpus with skip-gram (Bojanowski et al. 2017)
- pre-trained alignment matrix to map nl>en vector space (Smith et al. 2017)
- cosine similarity

Results

	Cognates			Non-cognates		
Metric	Prec	Rec	F-score	Prec	Rec	F-score
Prefix	82.30	87.99	85.05	62.74	51.66	56.65
Dice	80.40	91.23	85.47	65.84	43.19	52.15
Dice (3gr)	79.94	91.28	85.23	65.05	41.48	50.65
Jaccard	80.71	90.74	85.43	65.34	44.58	52.98
XDice	76.45	95.94	85.09	70.25	24.50	36.32
XXDice	79.89	94.15	86.43	72.56	39.45	51.09
LCSR	83.79	91.99	87.70	72.73	54.55	62.32
NLS	85.23	88.48	86.83	67.42	60.83	63.95
LCSR (2gr)	78.77	91.57	84.69	63.16	36.94	46.60
NLS (2gr)	79.09	90.57	84.44	61.69	38.82	47.64
LCSR (3gr)	79.94	91.28	85.23	65.05	41.48	50.65
NLS (3gr)	80.04	90.97	85.15	64.57	42.05	50.92
Jaro-Winkler	82.24	91.31	86.54	69.11	49.64	57.76
SpSim (opt.1)	85.58	81.05	83.23	57.40	65.01	60.89
SpSim (opt.2)	80.87	87.42	83.99	59.57	47.02	52.33
Sem	83.56	95.53	89.14	82.00	51.99	63.62
Ortho	89.46	91.23	90.33	76.42	72.54	74.42
Ortho + Sem	92.59	94.65	93.61	85.52	80.64	83.00

Average score			
Metric	Prec	Rec	F-score
Prefix	72.52	69.82	70.85
Dice	73.12	67.21	68.81
Dice (3gr)	72.50	66.38	67.94
Jaccard	73.02	67.66	69.20
XDice	73.35	60.22	60.70
XXDice	76.22	66.80	68.76
LCSR	78.26	73.27	75.01
NLS	76.33	74.66	75.39
LCSR (2gr)	70.96	64.25	65.64
NLS (2gr)	70.39	64.69	66.04
LCSR (3gr)	72.49	66.38	67.94
NLS (3gr)	72.30	66.51	68.04
Jaro-Winkler	75.67	70.47	72.15
SpSim (opt.1)	71.49	73.03	72.06
SpSim (opt.2)	70.22	67.22	68.16
Sem	82.78	73.76	76.38
Ortho	82.94	81.88	82.38
Ortho + Sem	89.05	87.64	88.30

Conclusions

- embeddings already obtain good results for cognate classification
- combining orthographic and semantic similarity boosts the performance
- semantic information helps to detect cognates with less orthographic similarity
- semantic information generates fewer false negatives

Contact

els.Lefever@ugent.be

References

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5:135–146.

S. Labat, L. Vandevoorde, and E. Lefever. 2019. Annotation Guidelines for Labeling English-Dutch Cognate Pairs, version 1.0. Technical report, Ghent University, LT3 15-01.

S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax.

GHENT UNIVERSITY