Strong Baselines for Complex Word Identification across Multiple Languages

Pierre Finnimore, Elisabeth Fritzsch, Daniel King, Alison Sneyd, Aneeq Ur Rehman,
Fernando Alva-Manchego, Andreas Vlachos

The University Of Sheffield.

UNIVERSITY OF CAMBRIDGE

DATASETS (SECOND CWI SHARED TASK)

Dataset / Language	Train	Dev	Test
English (EN) - News	14,002	1,764	2,095
English (EN) - WikiNews	7,746	870	1,287
English (EN) - Wikipedia	5,551	694	870
Spanish (ES)	13,750	1,622	2,232
German (DE)	6,151	795	959
French (FR)	N/A	N/A	2,251

MONOLINGUAL MODELS

Ours: Logistic Regression (LR) + 25 features

Target-based	Named Entity type, part-of-speech, num. hypernym, num. tokens, language-normalised num. characters in each word, unigram probabilities
Subword-based	prefixes, suffixes, num. syllables, num. complex punctuation marks (e.g. hyphens)
Sentence-based	sentence length, n-grams from the whole sentence

Dataset	Ours	Best CWI ST	SotA
EN-News	86.0	**87.4**	**87.4**
EN-WikiNews	81.6	**84.0**	**84.0**
EN-Wikipedia	76.1	**81.2**	**81.2**
ES	**77.6**	77.0	77.0
DE	74.8	74.5	**75.5**

same simple model

different complex models

Few language-independent features and Logistic Regression achieve comparable results to SotA complex models in monolingual and cross-lingual Complex Word Identification

Sentence	Target Word or Multi-word Expression	Complex?
Both China and the Philippines flexed their muscles on Wednesday	flexed	Yes
	flexed their muscles	Yes
	muscles	No

Cross-lingual

Features:

- number of syllables in the target
- number of tokens in the target
- number of complex punctuation marks (e.g. hyphens)
- sentence length
- unigram probabilities

Dataset	Ours	SotA
FR	75.8	76.0

LR + 5 Features

MTL + NNs + Random Forests Ensemble

https://github.com/sheffieldnlp/cwi

@feralvam

CROSS-LINGUAL MODELS

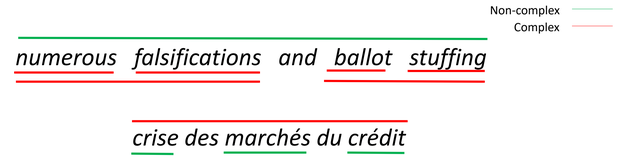
Ours: LR + 5 language-independent features

EN	ES	DE	Test	Macro F1
		X	EN-WikiNews	62.3
	X		EN-Wikipedia	63.1
	X		EN-News	67.2
		X	ES	72.6
X	X		DE	73.4
		X	FR	75.8

Adding EN to the training languages decreases performance for both ES and FR, but not for DE.

DATASETS ANALYSIS

Every sub-word in 25% of multi-word expression instances has the opposite label



Every target multi-word expression in Spanish, German and French is labelled as complex