Complex Word Identification with Ensemble-Based Voting

Cambridge ALTA

{ Sian Gooding, Ekaterina Kochmar } Dept. of Computer Science and Technology

University of Cambridge

(1) Motivation

Example: Hundreds of protesters have been detained in both cities.

- hundreds ∼ more than 100
- protesters ∼ demonstrators
- detained ∼ arrested
- cities ∼ ?

Word	Binary	Proportion
hundreds	1	0.05
protesters	1	**0.30**
detained	1	**0.75**
cities	0	0.00

CWI is an important task in its own right:

1. facilitates more targeted text adaptation
2. helps avoid unnecessary & educationally harmful "oversimplification"
3. alleviates data sparsity: definition can be provided if no simpler alternative available

(2) CWI Shared Task 2018

Data:

- 3 data sources: News (News), WikiNews (WiNs) and Wikipedia (Wiki)
- Content words and phrases annotated via MTurk by 10 native and 10 non-native speakers (metadata not used or released)
- 2 settings: `bin` if at least one annotation as CW, `prob` – proportion of annotators

Annotation:

Data	0_{bin}	1_{bin}	0.05_{prob}	1.0_{prob}
News$_{tr}$	60.41	39.59	13.52	0.39
News$_{dev}$	60.54	39.46	13.83	0.28
News$_{ts}$	61.72	38.28	12.70	0.29
WiNs$_{tr}$	58.48	41.52	16.25	0.17
WiNs$_{dev}$	59.43	40.57	14.25	0.11
WiNs$_{ts}$	57.58	42.42	16.71	0.16
Wiki$_{tr}$	55.07	44.93	16.66	0.52
Wiki$_{dev}$	51.15	48.85	19.31	0.14
Wiki$_{ts}$	49.54	50.46	18.62	0.23

(3) Challenges

1. Context-specific annotation: up to 10% words receive different annotations; e.g., *tragedy* from 0.00 to 1.00 – interaction of surrounding context, multiple senses and sequence labelling effect

2. Sequence labelling effect:
Beethoven's *Symphony*$_{0.6}$ No.7, Bruckner's *Symphony*$_{0.1}$ No.6 and Mendelssohn's *Symphony*$_{0.0}$ No.4 comprise a nearly complete list of *symphonies*$_{0.3}$ in this key in the Romantic era.

3. Phrase annotation:

- *future*$_{0.05}$ ∪ *generations*$_{0.25}$ = *future generations*$_{0.15}$
- *traditional*$_{0.2}$ ∪ *connection*$_{0.0}$ ∪ *country*$_{0.05}$ ≠ *traditional connection to that country*$_{0.0}$

4. Proper nouns: 0.0 − 0.45 for *Eurozone*, 0.0 − 0.05 for *Barack*, 0.05 − 0.3 for *Brexit*, and 0.0 − 0.1 for *Copenhagen*, *Estonia*, *Hungary*, *Warsaw*, etc.

(4) CAMB System Overview

- Preliminary experiments confirm that ensemble-based approaches work best
- **Method for `bin` setting**:
  - Wikipedia & News – AdaBoost with 5,000 estimators
  - WikiNews – ensemble voting classifier using AdaBoost and Random Forest
- **Method for `prob` setting**: Linear Regression; round the classifier's predictions to the nearest value on [0.00, ..., 1.00] with the step of 0.05

(5) Experimental Results

Features Overview:

1. **Word N-grams and PoS**: words, character bigrams and PoS tags
2. **Lexical Features**: word length, number of syllables, number of senses, hyper- and hyponyms from the WordNet
3. **Dependency Parse Relations**: number of dependency relations for the target word
4. **Lexicon-Based Features**: presence/absence in the *SubIMDB*, the *Simple Wikipedia* and *Ogden's Basic English* list; CEFR level from the *Cambridge Advanced Learners Dictionary*
5. **Word Frequency** in the Google N-grams
6. **Psycholinguistic Features** from the MCR Psycholinguistic Database: *word familiarity*, *imageability*, *concreteness*, *age of acquisition*, etc.

Feature selection:

- `bin`: all features for News & WiNs; all but MCR features for Wiki
- `prob`: all features for News; all but MCR features for WiNs & Wiki

	Binary (F-Score)	Probabilistic (MAE)
News	**0.8736**	0.0558
WiNs	**0.8400**	**0.0674**
Wiki	**0.8115**	**0.0739**

- `bin`: News trained on News; all training data on WiNs & Wiki
- `prob`: all training data on News & WiNs; Wiki & WiNs training data on Wiki

(6) Analysis

- **Per-Genre Performance:** Unique words

	News	WiNs	Wiki
Total	13,461	7,559	5,439
Unique	3,376	3,334	3,157
%	25.08	44.10	58.44

  - Classifiers in both settings perform best on News: News contains lowest number of complex words & lowest number of unique words ⇒ less challenging
  - Wiki – more challenging for humans (highest CW %) and machines (lowest results) + highest number of unique words

- **Phrase Classification**

Data	Acc	P	R	F-Score
CW pres.	0.6987	0.8049	0.8231	0.8139
N-gram	0.8004	0.8015	0.9977	0.8889
Greedy*	**0.8004**	**0.8004**	**1.000**	**0.8891**

- **Performance across PoS**

Data	Size	Acc	P	R	F
Total	3,701	0.86	0.82	0.79	0.85
Nouns	2,427	0.86	0.80	0.76	0.84
Verbs	718	0.84	0.83	0.81	0.84
Adj's	435	0.88	0.86	0.86	0.87
Adv's	111	0.91	0.89	0.92	0.91

- Nouns represent the largest proportion of all test items, while showing the lowest precision and recall
- Dependency on context: 88.94% of misclassified instances in News, 61.31% in WiNs and 52.78% in Wiki have multiple labels
- Proper nouns are problematic: 12.56% of misclassified instances in News, 22.02% in WiNs and 22.92% in Wiki

(7) Conclusions

Our systems **scored first** on all 3 text genres in the `bin` classification track, and on 2 out of 3 genres in the `prob` track. Further analysis identifies future directions for this research.

1. contextualisation of CWI
2. better phrase complexity prediction
3. personalisation of CWI with level of education, L1 and level of language competence

Contact Information

Email Sian.Gooding@cl.cam.ac.uk
Ekaterina.Kochmar@cl.cam.ac.uk
Web www.cl.cam.ac.uk/∼ek358/