Complex Word Identification Based on Frequency in a Learner Corpus

Tomoyuki Kajiwara[†‡] and Mamoru Komachi[†] †: Tokyo Metropolitan University ‡: Osaka University

Complex Word Identification (CWI) Shared Task 2018

Systems predict whether words in a given context are complex or non-complex for a non-native speaker.

	Target	Label	Probability
According to Goodyear, a neighbor heard gun **shots**.	shots	0	0.00
According to Goodyear, a neighbor heard gun shots.	According to	1	0.05
A bad part of the **investigation** is that we may not get the why.	investigation	1	0.95

Table 1: Example instances of the English dataset.

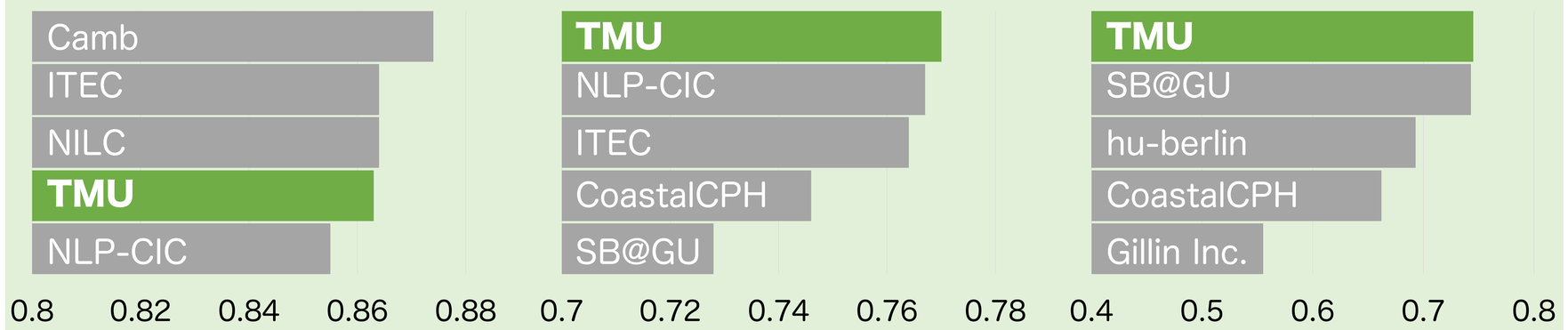
		Train	Dev	Test
English	(News)	14,002	1,764	2,095
	(WikiNews)	7,746	870	1,287
	(Wikipedia)	5,551	694	870
Spanish		13,750	1,622	2,233
German	(Wikipedia)	6,151	795	959
French		0	0	2,251

Table 2: # instances for each dataset.

TMU Systems

- **Classifiers**
  - Random Forest Classifier
  - Random Forest Regressor
- **Features**
  - Number of characters
  - Number of words
  - Frequency in the Wikipedia corpus
  - Frequency in the WikiNews corpus
  - **Frequency in the Lang-8 corpus**

Performance on the Binary Classification Task (macro-averaged F1)



Lang-8 learner corpus

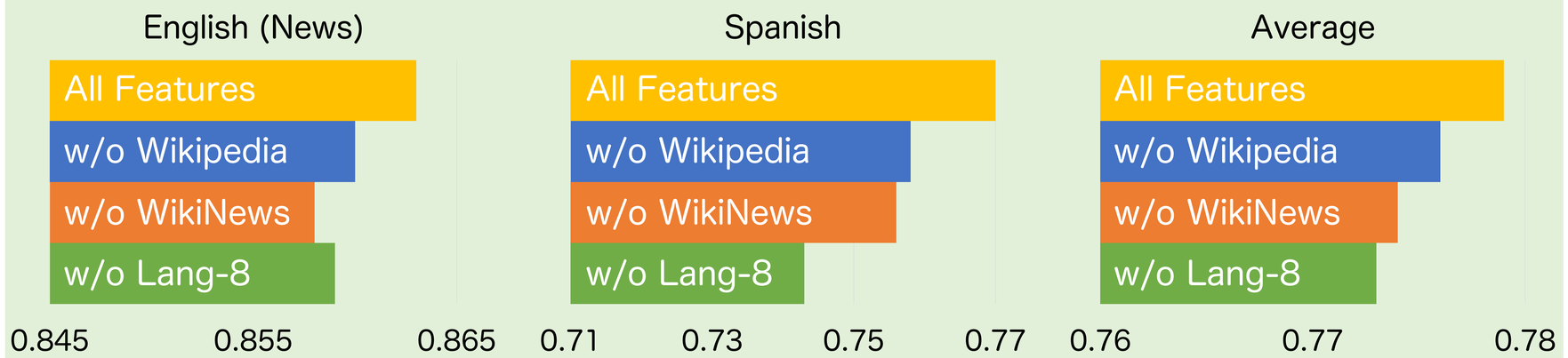
Lang-8 is a learner corpus that can be used on a large-scale in many languages.

	Wikipedia	WikiNews	Lang-8
English	95M	330K	3M
Spanish	20M	110K	190K
German	44M	150K	160K
French	26M	140K	180K

Table 3: # sentences for each corpus.

Language learners tend to use simple words as compared to native speakers. Therefore, we expect the word frequency in the learner corpus to be useful feature for CWI task.

Ablation Analysis



Conclusion

- It was not clear what kind of corpus was useful for estimating word difficulty.
- We discussed the usefulness of a learner corpus for the first time.
- As anticipated, word frequency in the learner corpus is a useful feature for CWI.
- **TMU systems won in 5 out of the 12 tracks.**