[SIAN GOODING, EKATERINA KOCHMAR]

Complex Word Identification as a Sequence Labelling Task

MOTIVATION

Complex word identification (CWI) is concerned with the detection of words in need of simplification. The context of a word has an impact on its perceived complexity:

	Proportion of complex annotations
Elephants have four molars …	17/20
… new molars emerge in the back of the mouth.	0/20

Current approaches to CWI, including state-of-the-art systems, have limitations in that they require **extensive feature engineering** and/or consider word complexity in **isolation from the context**.

CONTRIBUTIONS

We present a CWI system based on ***sequence labelling*** that can:

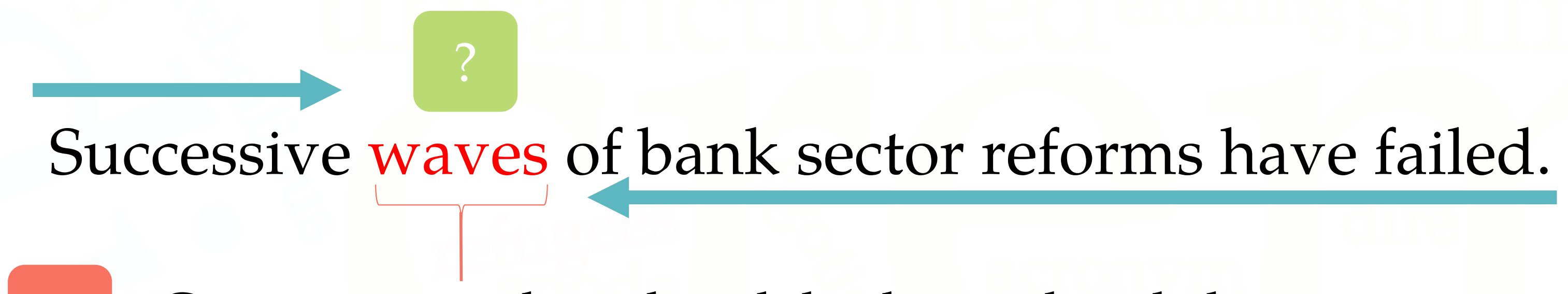
1. Take **word context** into account;
2. Rely on **word embeddings only**, eliminating the need for extensive feature engineering;
3. Be used with **complex word and complex phrases.**

SYSTEM

We use a sequential architecture by Rei (2017) with 300-dimensional GloVe embeddings.

The design is highly suited to the task of CWI as:

1. BiLSTM provides contextual information from both the left and right context of a target word;

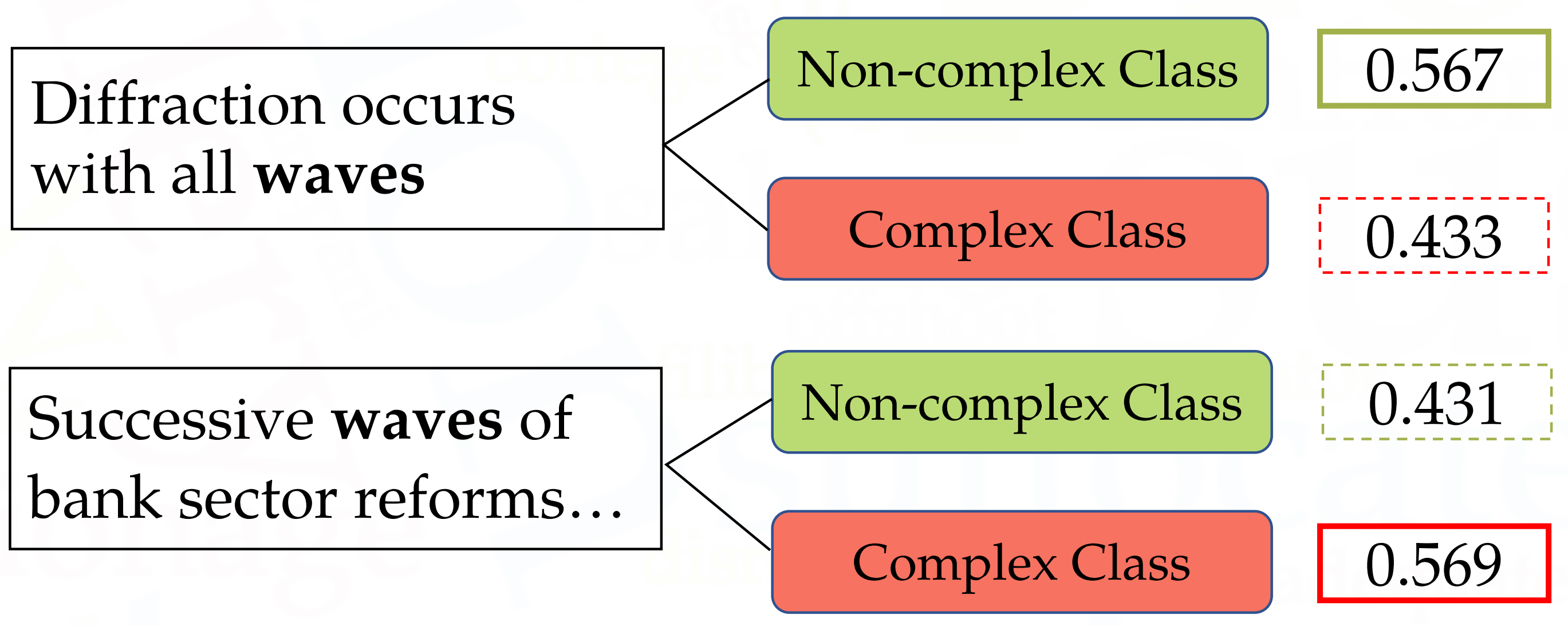


Successive waves of bank sector reforms have failed.

2. Context is combined with both word and character-level representations (Rei et al., 2016);
3. Language modelling objective enables the model to learn better composition functions and to predict the probability of individual words in context.

Prior work has found word frequency and length to be highly informative features, therefore we chose an architecture which uses sub-word information and a language modelling objective.

A word is considered complex if the probability of it belonging to the complex class > 0.50:



Diffraction occurs with all **waves**

Non-complex Class 0.567

Complex Class 0.433

Successive **waves** of bank sector reforms…

Non-complex Class 0.431

Complex Class 0.569

Phrases are considered complex if the average probability of each word belonging to complex class > 0.50 (excluding stop words):

successive waves

$$x = \frac{0.939 + 0.569}{2}$$

if x > 0.50: Complex Class

else: Non-complex Class

DATA

We use the English section of the **CWI datasets from Yimam et al. (2017)**, which contains annotated texts from professionally written *News*, *WikiNews* and *Wikipedia* articles in the following format:

Sentence	Word	Bin	Prob
They drastically…	drastically	1	0.5

As a sequential model expects complete word inputs we adapt the format to:

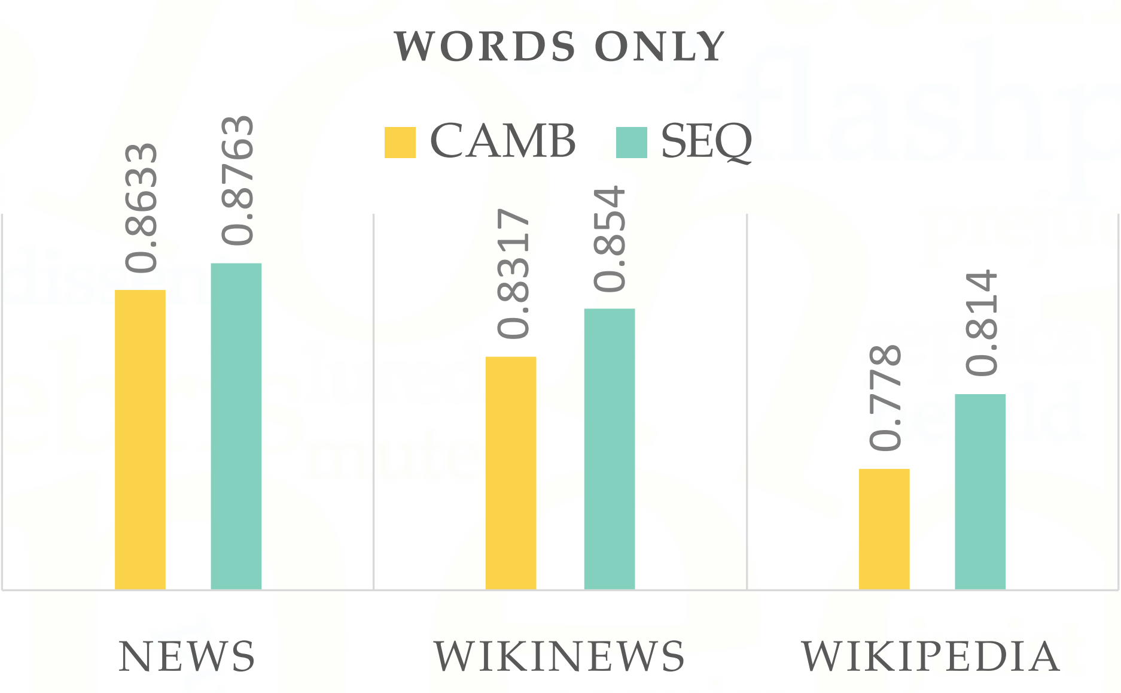
They	N
drastically	C
…	

where *C* is *complex* and *N non-complex.*

RESULTS

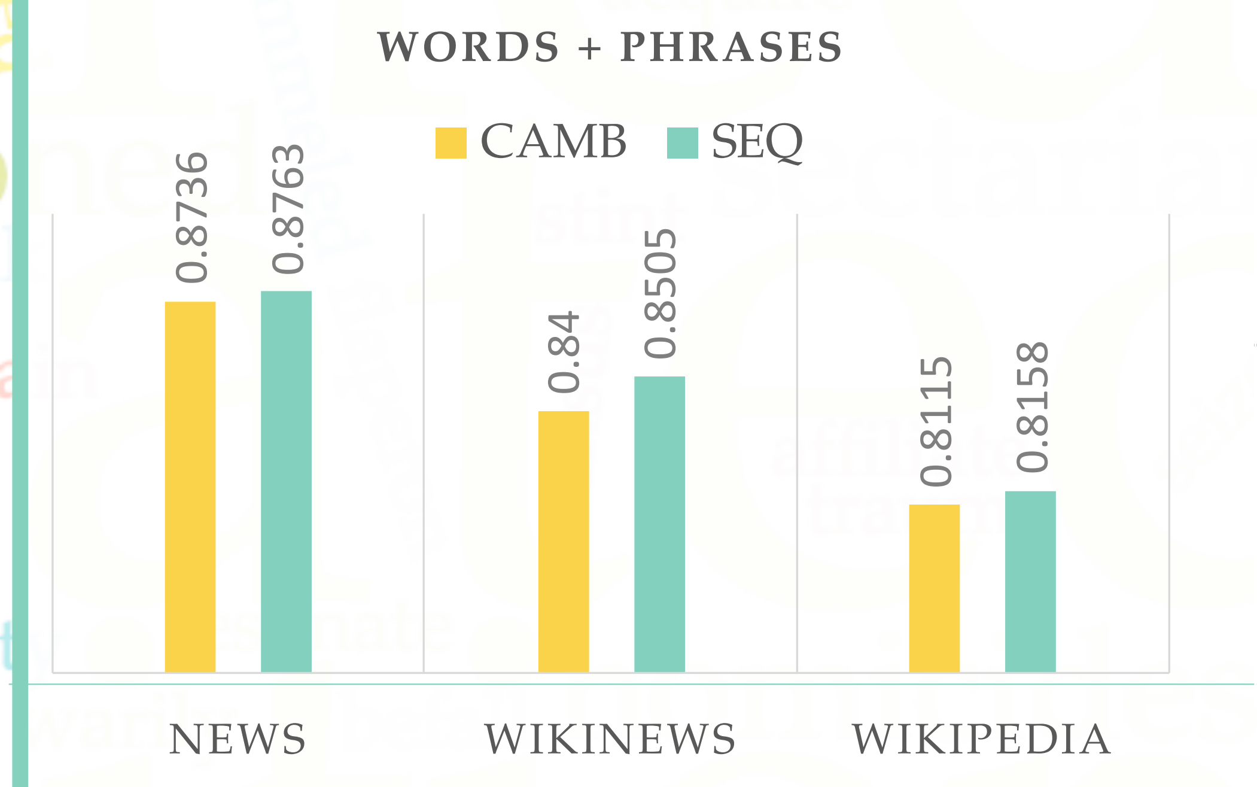
SEQ results compared to current state-of-the-art in complex word identification CAMB (Gooding and Kochmar 2018).

Evaluated with **macro-averaged F1** based on CWI 2018 shared task.



	CAMB ✓	CAMB ✗
SEQ ✓	3002	205
SEQ ✗	145	349

- Phrase classification by CAMB marks all phrases as complex whereas SEQ considers word complexity.



	CAMB ✓	CAMB ✗
SEQ ✓	3443	207
SEQ ✗	145	457

SEQ system achieves significantly better results than CAMB according to continuity corrected McNemar test (Edwards 1948):

Words only
(p=0.0016, $X^{2=}$= 9.95)

Words + Phrases
(p=0.0016, $X^{2=}$= 9.95)

CONCLUSIONS

✓ SEQ model views **CWI as a sequence labelling task**

✓ Achieves **state-of-the-art results** with a one-model-fits-all approach

✓ SEQ takes **word context** into account

Contexts	CAMB	SEQ	LABEL
Successive **waves** of bank sector reforms have failed	0	1	1
Diffraction occurs with all **waves**	0	0	0

*Example showing context impacting complexity of word **waves***

✓ SEQ can be used to classify both words and phrases in a **unified framework**

✓ **Avoids expensive feature engineering**

CONTACT INFORMATION

Sian Gooding, Ekaterina Kochmar [shg36, ek358] @cam.ac.uk

CWI SEQ Models available: github.com/siangooding/cwi

Cambridge ALTA
Institute for Automated Language Teaching and Assessment
UNIVERSITY OF CAMBRIDGE