KU LEUVEN

# Deep Learning Architecture for Complex Word Identification

**Dirk De Hertog** [1] **Anaïs Tack** [2,1]

[1] ITEC, imec, KU Leuven Kulak, Kortrijk, Belgium

[2] CENTAL, Université catholique de Louvain, Louvain-la-Neuve, Belgium

dirk.dehertog@kuleuven.be

## CWI SHARED TASK 2018

comparison of techniques for complex word identification

(cf. Shardlow, 2013; Paetzold & Specia, 2016; Yimam et al., 2018)

**4 tracks**

- ✔ **English** monolingual CWI
- ✔ **Spanish** monolingual CWI
- ✘ **German** monolingual CWI
- ✘ **French** multilingual CWI

**2 tasks**

- ✔ **binary** classification (simple/complex)
- ✔ **probabilistic** classification (% of complex annotations)

**gold-standard complex word annotations**

| Both | China | and the | Philippines | flexed | their | muscles | on | Wednesday |
|---|---|---|---|---|---|---|---|---|
| | simple | | simple | complex | | simple | | simple |
| | 0.0 | | 0.0 | 0.4 | | 0.0 | | 0.0 |
| | | | | complex | | | | |
| | | | | 0.25 | | | | |

## DEEP LEARNING ARCHITECTURE

### FEATURES

**engineered features**

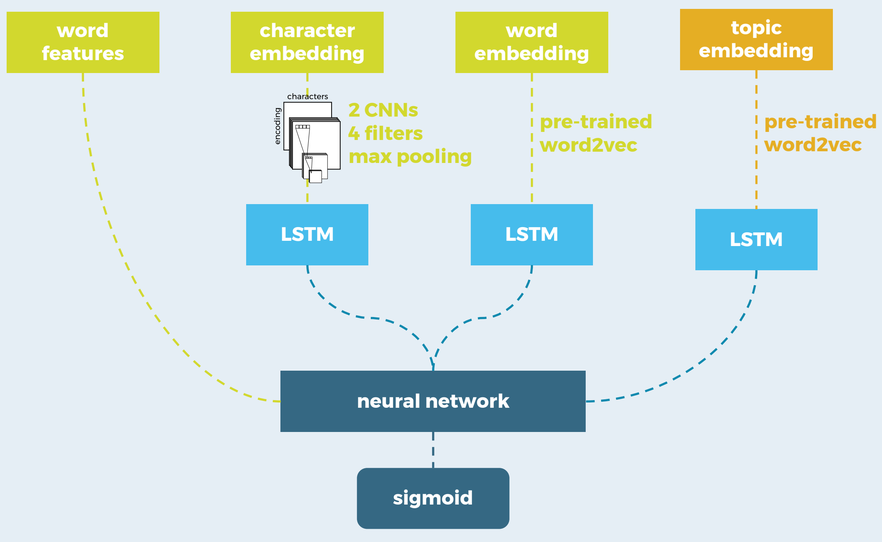
- word length
- word frequency: COW corpus (Schäfer, 2015)
- psycholinguistic norms: MRC database (Wilson, 1988)

**embeddings**

- character embeddings: 16-dimensional encodings
- word embeddings: COW corpus, word2vec (Mikolov, 2013)

### ARCHITECTURE

- 3 input layers at **word level**
- 1 input layer at **sentence level**
- **fully-connected network** w/ 3 layers, moderate dropout




## MONOLINGUAL TRACK — English

- high performance on **probabilistic task**
- more efficient on the **news genre**
- substantial impact of **word** and **character embeddings**

| | binary | | probabilistic | |
|---|---|---|---|---|
| | $F_1$ macro | ⬍ | MAE | ⬍ |
| **News** | .874 | 1 | .051 | 1 |
| | .864 | 2 | .054 | 2 |
| **Wikipedia** | .812 | 1 | .074 | 1 |
| | .782 | 5 | .081 | 2 |
| **WikiNews** | .840 | 1 | .067 | 1 |
| | .811 | 6 | .071 | 3 |

*Results on the English monolingual CWI track, compared to the top-performing system*

## MONOLINGUAL TRACK — Spanish

- top-tier performance on **binary** & **probabilistic tasks**
- good performance despite smaller dataset
- no impact of absence of psycholinguistic norms

| | binary | | probabilistic | |
|---|---|---|---|---|
| | $F_1$ macro | ⬍ | MAE | ⬍ |
| **Wikipedia** | .770 | 1 | .072 | 1 |
| | .764 | 2 | .073 | 2 |

*Results on the Spanish monolingual CWI track, compared to the top-performing system*

## KEY TAKEAWAYS

**Features of complexity**

- 📌 main effects for **character** and **word embeddings**
- 📌 no effects for **contextual** and **topical information**
- 📌 superfluous effects for **engineered features**
  - word length & corpus counts (overlap embeddings)
  - psycholinguistic norms

**English monolingual CWI**

- 📌 importance of sizeable training data

**Spanish monolingual CWI**

- 📌 embeddings account for lack of engineered features

imec