Entity Tagging & Linking

Afonso Mendes, David Nogueira, Pedro Balage, André Martins

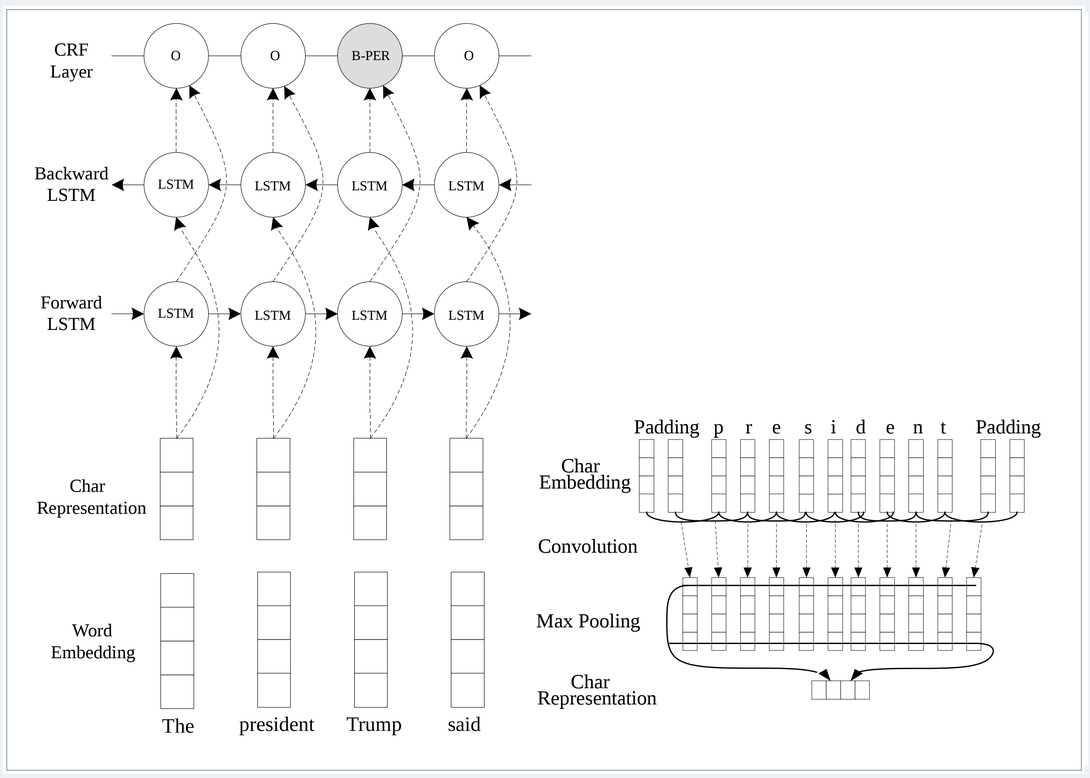
amm@priberam.pt, david.nogueira@priberam.pt
pedro.balage@priberam.pt, afm@priberam.pt

Goals

- To develop statistical models for entity recognition and linking.
- Explore multilingual approaches for the core languages of the project (English, German, Spanish) and for Portuguese.

Named Entity Recognition (NER)

- **Strong industrial implementation** that uses a sequence labelling model with gazetteers.
- New approaches using **deep neural networks** with bidirectional LSTMs and character- and word-level embeddings can achieve higher scores.



Architecture for the Neural Named-Entity Recognition

Text Analysis Conference (TAC)

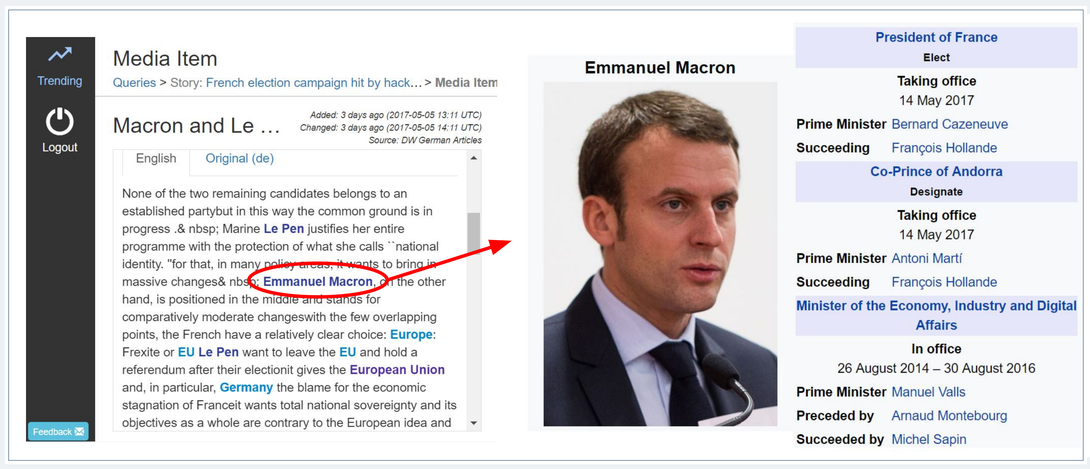
- The goal of TAC Knowledge Base Population (KBP) is to develop and evaluate technologies for populating knowledge bases (KBs) from unstructured text.
- TAC is sponsored by the U.S. National Institute of Standards and Technology (NIST) since 2008.
- **SUMMA participated at TAC 2017** in the track of Entity Discovery and Linking (EDL).
- From the 22 teams that participated, SUMMA **ranked as first for English EDL** and third for Spanish EDL.

Entity Linking System	NERLC	CEAFmC	Rank
SUMMA	79.4	83.1	**1st**
Competitor B	79.0	82.5	2nd
Competitor C	78.2	82.4	3rd

Results for English EDL 2017 with named-mentions.
NERLC evaluates the detection, type classification and linking of a mention against a knowledge-base. *CEAFmC* evaluates the detection and type classification of mention clustering.

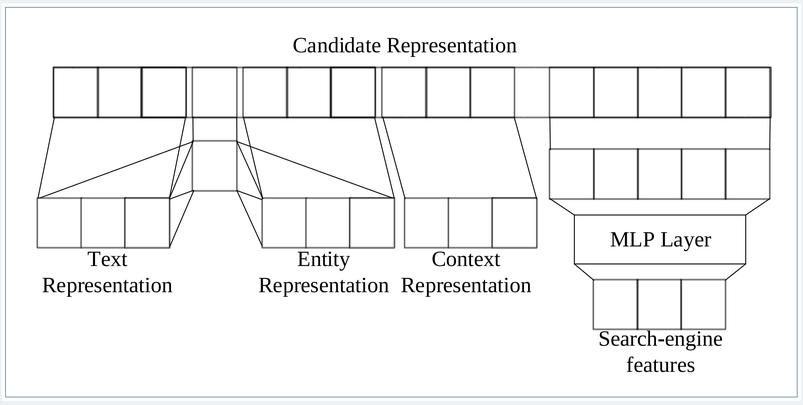
Entity Linking

- **State-of-the-art industrial language-agnostic implementation**, with document-level and corpora-level coherence steps.
- The model has a small number of parameters, being easily deployed in other languages.
- New model using **distributed representations** boosted the results even further.



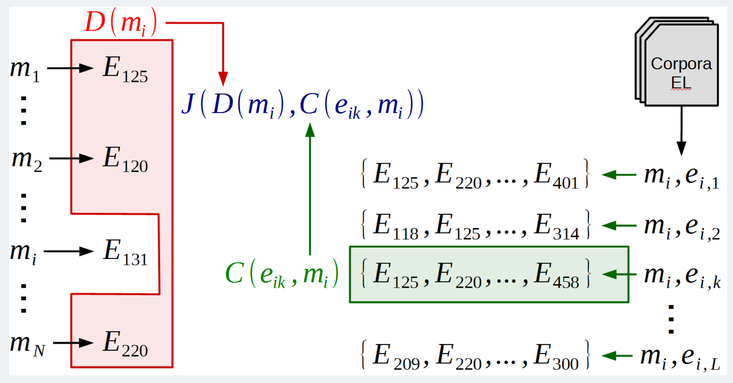
Algorithm for Entity Linking:

1. High-precision **sub-string match mention coreference**
2. **Candidate generation**
3. Features generation: information retrieval engine (KNN algorithm) + prior statistics + mentions candidates co-occurrences
4. **Distributed representation neural network disambiguation**
   A Multi-layer Perceptron (MLP) classifier receives as input the learned representations obtained from the text, each candidate entity representation, context representation and the information retrieval features.



Candidate representation in the neural network model

5. **Nominal mentions disambiguation**
6. **Global NIL clustering** and **cross-document coherence**



Cross document coherence model based on Jaccard similarity

What are the Challenges?

- Named entity recognition performance is directly connected with the amount of data used for training.
- Ensemble methods for named entity recognition.
- Learn suitable cross-lingual entity representations.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 688139 (SUMMA).