

Entity Tagging & Linking

Mariana Almeida, Afonso Mendes, André Martins, Samuel Broscheit, Filipe Aleixo



m1a@priberam.pt, amm@priberam.pt, afm@priberam.pt
samuel.broscheit@googlemail.com, filipe.aleixo@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)

Media Item

Queries > Stories of Query: Venezuela in News > Story: Venezuela protests show no signs of ... > Media Item: Venezuela

Venezuela protests show no signs of letting up... Added: 4 days ago (2017-05-04 19:17 UTC) Changed: 3 days ago (2017-05-05 14:26 UTC) Source: DW English Articles

Together with her supporters and family, **Lilian Tintori**, the wife of jailed Venezuelan opposition leader **Leopoldo Lopez**, stood in front of the hilltop **Ramo Verde** jail on **Thursday** and demanded to see her husband. **Luis Almagro**, **Organization of American States (OAS)**, wrote via **Twitter**: "The **Venezuela** government has refused to confirm the health of political prisoner **Leopoldo Lopez**. Family and lawyers have not seen him in more than a month." "I demand to visit **Leopoldo Lopez** based on the commitments that **Venezuela** has with the **Inter-American System of Human Rights**," Almagro wrote. **Lopez**, a former mayor, was sentenced to nearly 14 years in jail in **2015** following the last major anti-government

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

Entity Linking

The screenshot shows a media item titled "Macron and Le ..." with a photo of Emmanuel Macron. A red arrow points to the name "Emmanuel Macron" in the text, which is highlighted in red. The sidebar on the right contains biographical information for Emmanuel Macron, including his position as President of France, his taking office date (14 May 2017), and his predecessor (François Hollande).

- State-of-art industrial implementation, with document-level and copora-level coherence steps
- Good results for English and Spanish.
- The model has a small number of parameters, being easily portable between different languages.
- New approaches using deep neural networks are being tested.

What are the Challenges?

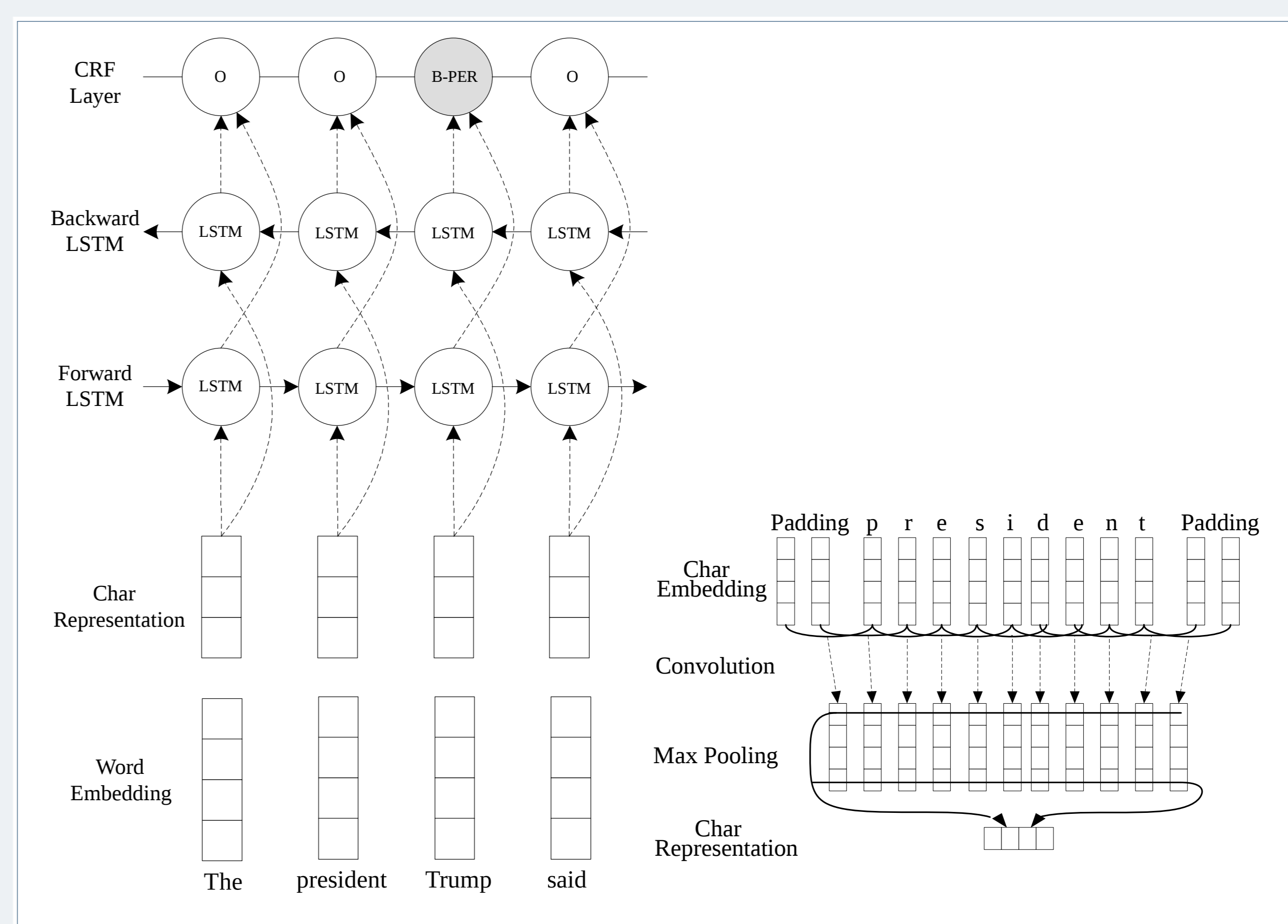
- Named entity recognition performance is directly connected with the amount of data used for training.
- Ensemble methods for named entity recognition.
- Learn a suitable cross-lingual representations for entities.
- Explore multilingual approaches.

Text Analysis Conference (TAC)

- Participation in 2016 Entity Discovery and Linking (EDL) track.
- Planning 2017 participation in the same task.
- NER was the main bottleneck of our TAC 2016's submission.

NER System	Model	Training Data	NER	NERC
Summa (TAC-2016)	Linear	TAC+OntoNotes	83.1	76.1
Summa (2017)	Linear	TAC	81.67	77.61
Summa (2017)	Linear	TAC+OntoNotes	85.68	81.23
SummaNN (2017)	Neural	TAC+OntoNotes	87.87	84.49
USTC (TAC-2016)	Neural	TAC+iFLYTEK	90.55	87.79

SUMMA NER system performance with different models and datasets.



Architecture for SummaNN System.

- Our Entity linking system has state-of-art performance.

Entity Linking System	NERL	KBIDs	CEAFm
Best Team (TAC 2016)	81.1	81.1	83.2
Summa (TAC 2016)	81.8	81.0	83.3
Summa Improved	83.4	81.3	86.1

Entity Linking system performance using entities provided by the USTC system.

Demo

The screenshot shows the priberam EDL interface. A search for "Lisbon" is performed. The results show a list of mentions: "Lisbon => Lisbon", "Portuguese => Portugal", "Lisboa => Lisbon", "Portugal => Portugal", and "European Union => European Union". A text snippet is shown with "Lisbon" highlighted in red. The interface also includes a "Submit" button and a "Mentions list" section.