Personalized PageRank with Syntagmatic Information for Multilingual Word Sense Disambiguation

Federico Scozzafava, Marco Maru, Fabrizio Brignone, Giovanni Torrisi, and Roberto Navigli

SAPIENZA
UNIVERSITÀ DI ROMA

SAPIENZA
NLP

Babelscape

elexis
european lexicographic infrastructure

ELEXIS project No. 731015

Consolidator Grant
MOUSSE No. 726487

Overview

SyntagRank

A multilingual knowledge-based WSD system
based on the Personalized PageRank algorithm

- Achieves state-of-the-art performances in 5 languages (EN, DE, FR, ES, IT)
- Can be queried via a user-friendly interface at http://syntagnet.org/
- ...or programmatically, via a RESTful endpoint at http://api.syntagnet.org/

What is Word Sense Disambiguation?

Word Sense Disambiguation (WSD) is the task of selecting the proper sense for an ambiguous word in a particular context.

Use the **force**, Luke!

SAPIENZA NLP

How is lexical ambiguity addressed in WSD?

Supervised approaches

- Leverage annotated data
- **Achieve state of the art**
- **Need huge manual effort**

Knowledge-based approaches

- Exploit Lexical Knowledge Bases
- **Scale to different languages**
- **Have lower performances**

Can we improve knowledge-based disambiguation?

Enriching Lexical Knowledge Bases (LKBs) with **syntagmatic relations** proved to be very effective for knowledge-based WSD (Maru et al., 2019)

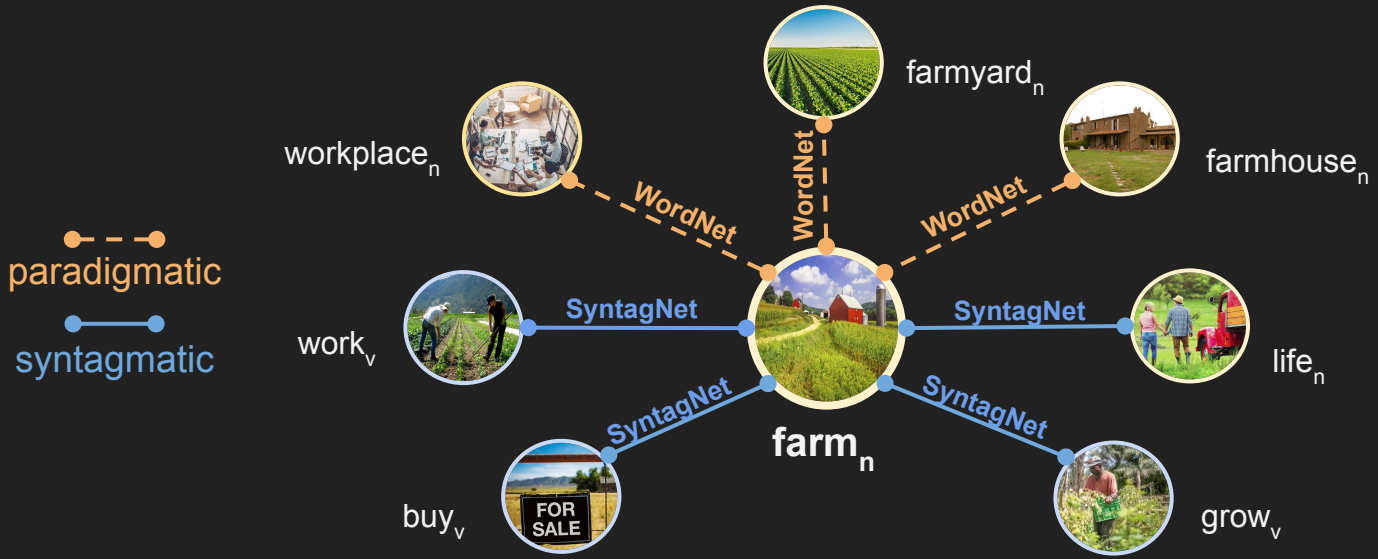
SyntagNet

- ★ A manually-curated database of lexical-semantic combinations
- ★ Captures sense distinctions evoked by syntagmatic relations
- ★ Covers 78,000 lexical and 88,019 semantic combinations
- ★ Links 20,626 WordNet 3.0 synsets with a relation edge

How does SyntagRank's LKB look like?

SyntagRank employs an LKB made up of:

1. **WordNet 3.0** (Fellbaum, 1998) + **Princeton WordNet Gloss Corpus** (PWNG)
2. **SyntagNet** (Maru et al., 2019)



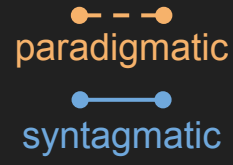
SAPIENZA NLP

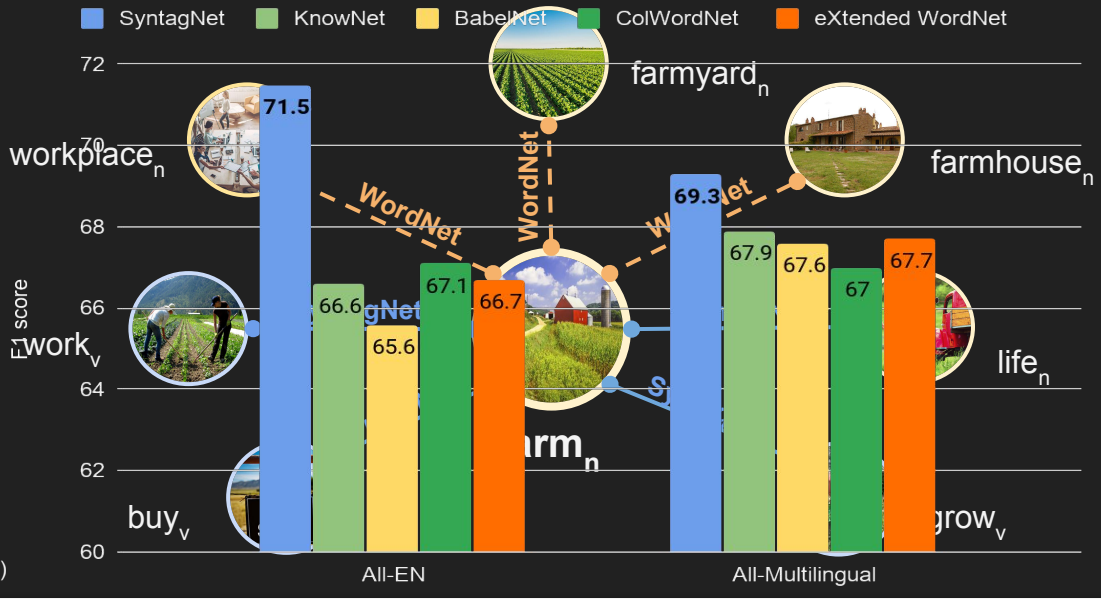
How does SyntagRank's LKB look like?

WordNet provides paradigmatic knowledge, describing each concept in terms of its properties.

SyntagNet provides contextual knowledge capturing the semantic of each concept in terms of actions.

Syntagmatic relations leverage pairs of co-occurring words to comprise contextual semantic information.

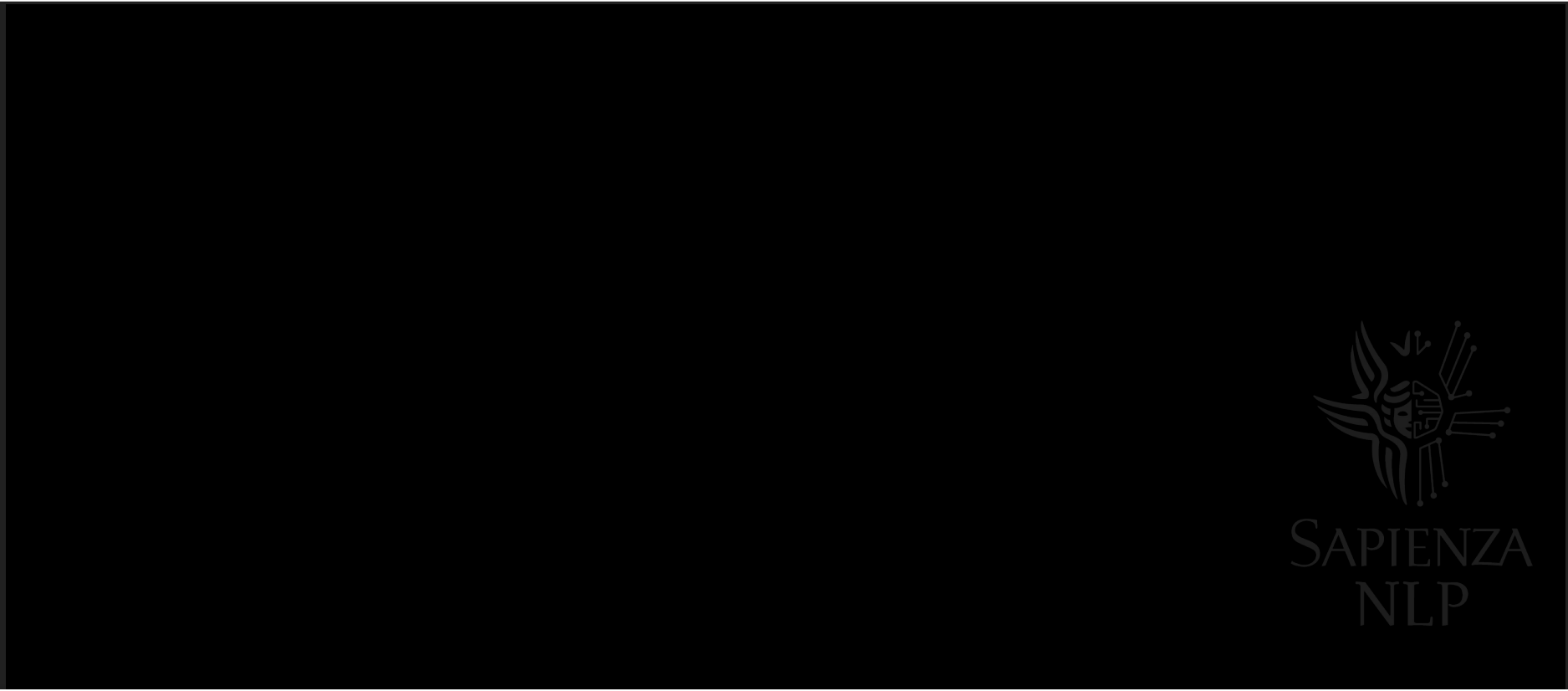




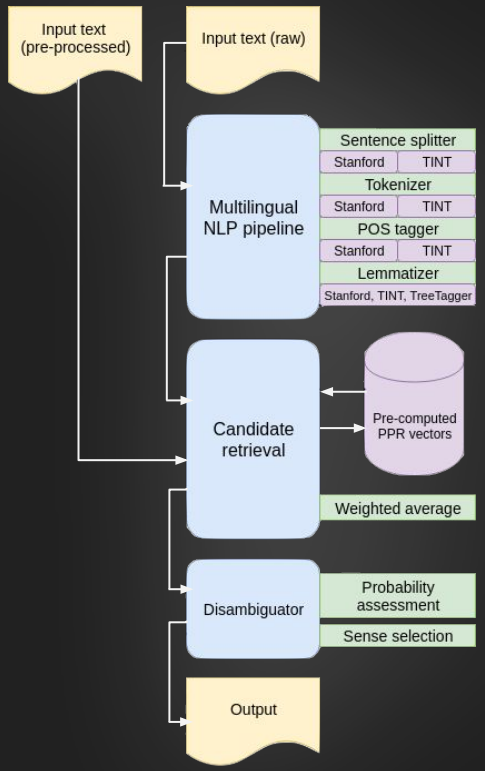
Taken from SyntagNet (Maru et al., 2019)

SAPIENZA NLP

The SyntagNet Explorer



System Architecture of SyntagRank



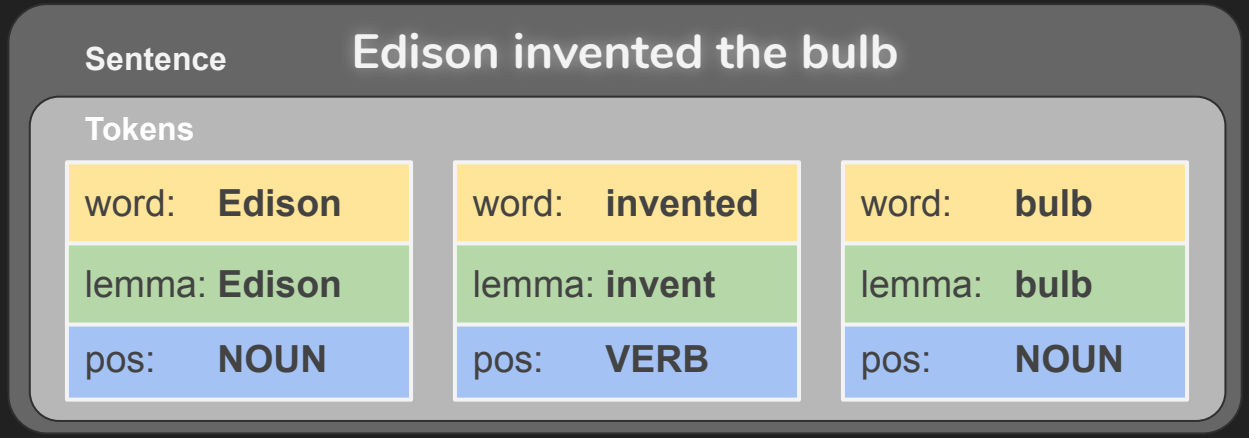
SyntagRank uses an optimized implementation of the Personalized PageRank (PPR) algorithm.

It has three main modules:

1. **Multilingual NLP Pipeline**
2. **Candidate Retrieval**
3. **Disambiguator**

SAPIENZA NLP

Module 1: the Multilingual NLP Pipeline



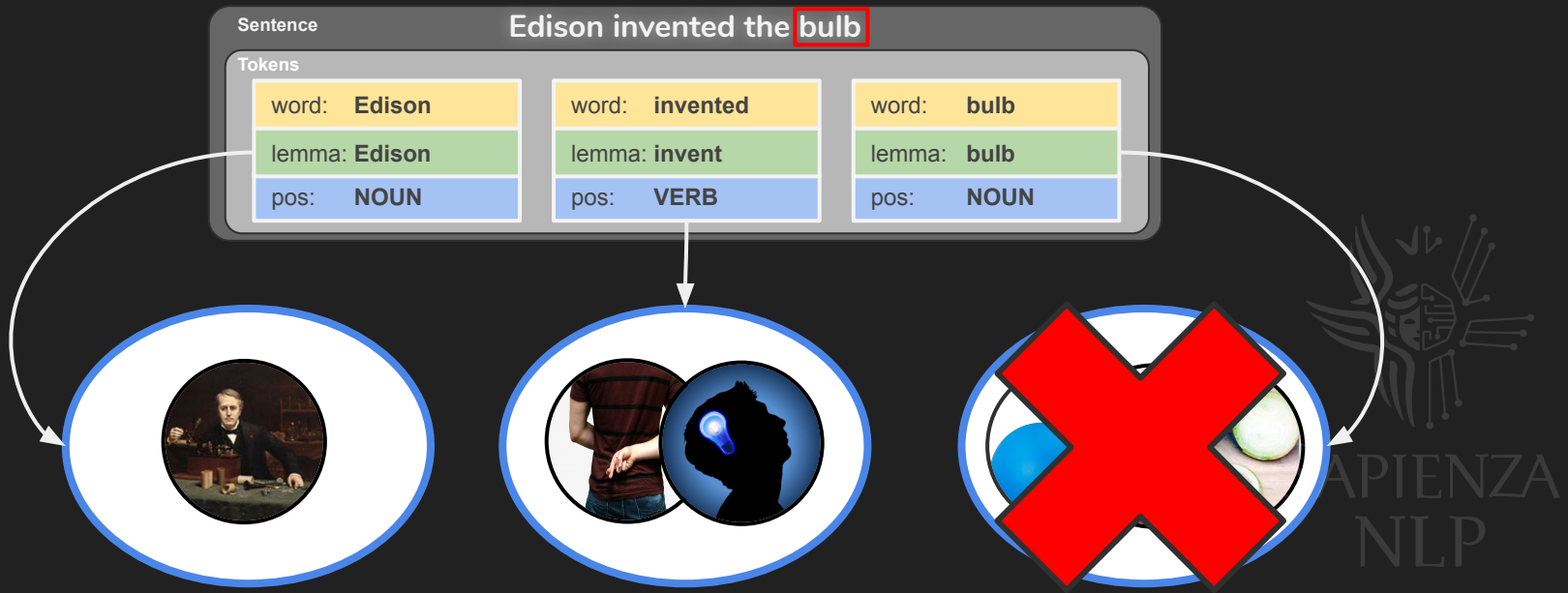
Depending on the input language, **SyntagRank** employs:

- The Stanford CoreNLP suite (Manning et al., 2014)
- The models provided by The Italian NLP Tool (Palmero Aprosio and Moretti, 2016, TINT)

SAPIENZA NLP

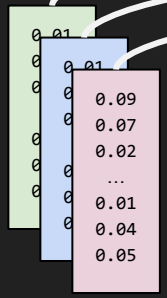
Module 2: the Candidate Retrieval

For each content word in the input sentence, we collect its candidate concepts from the LKB. Then, to disambiguate a target word, we make use of the ***w2w*** heuristic (Agirre et al., 2014).



Module 2: the Candidate Retrieval (PPR)

Due to the Linearity Theorem (Jeh and Widom, 2003), computing the PPR vector starting from a set of nodes is equivalent to the weighted average of the PPR vectors of the single starting points.



Weighted Average

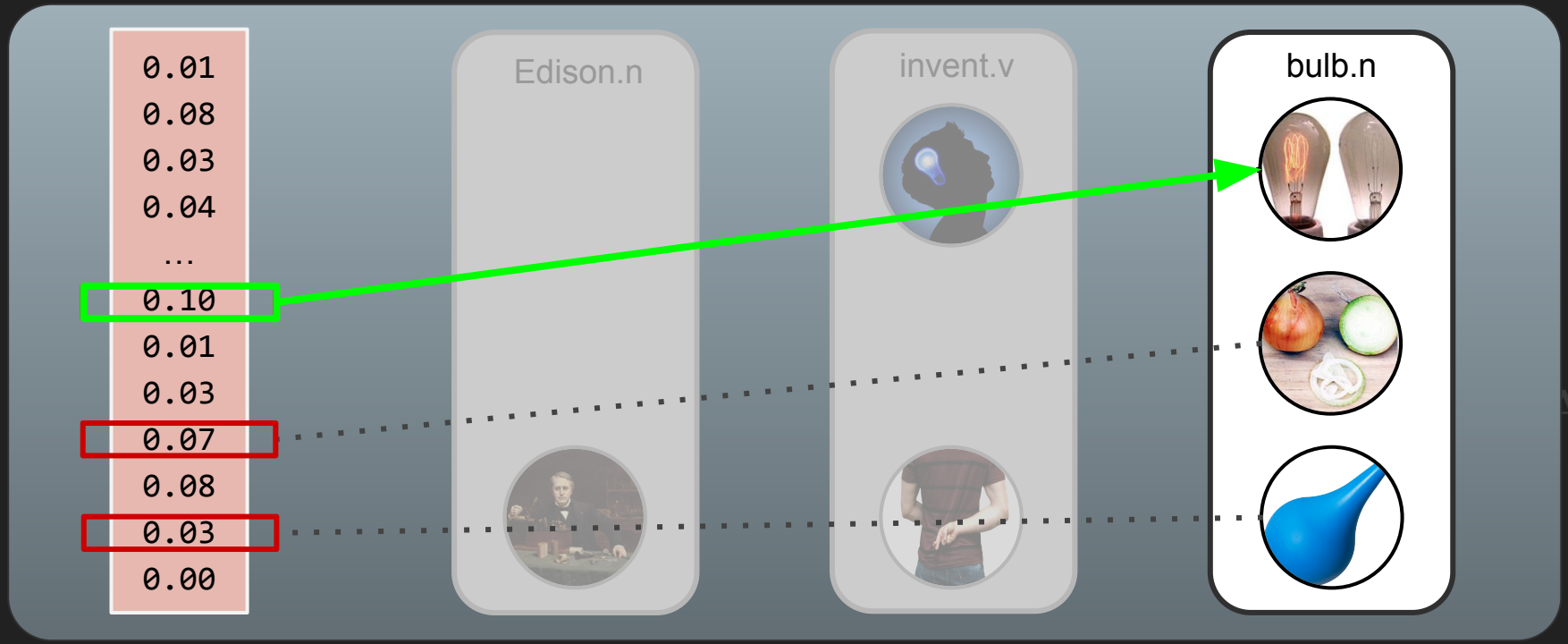
$$\sum \frac{1}{N * |senses_w|} freq_{ws}$$

Contextual PPR Vector

0.01
0.08
0.03
...
0.01
0.02
0.10

Pre-computed PPR vectors

Module 3: the Disambiguator



WSD Evaluation - Setting

SyntagRank was tested on the five English WSD evaluation datasets of Raganato et al, 2017:

- Senseval-2 (Edmonds and Cotton, 2001)
- Senseval-3 (Snyder and Palmer, 2004)
- SemEval-2007 (Pradhan et al., 2007)
- SemEval-2013 (Navigli et al., 2013)
- SemEval-2015 (Moro and Navigli, 2015)

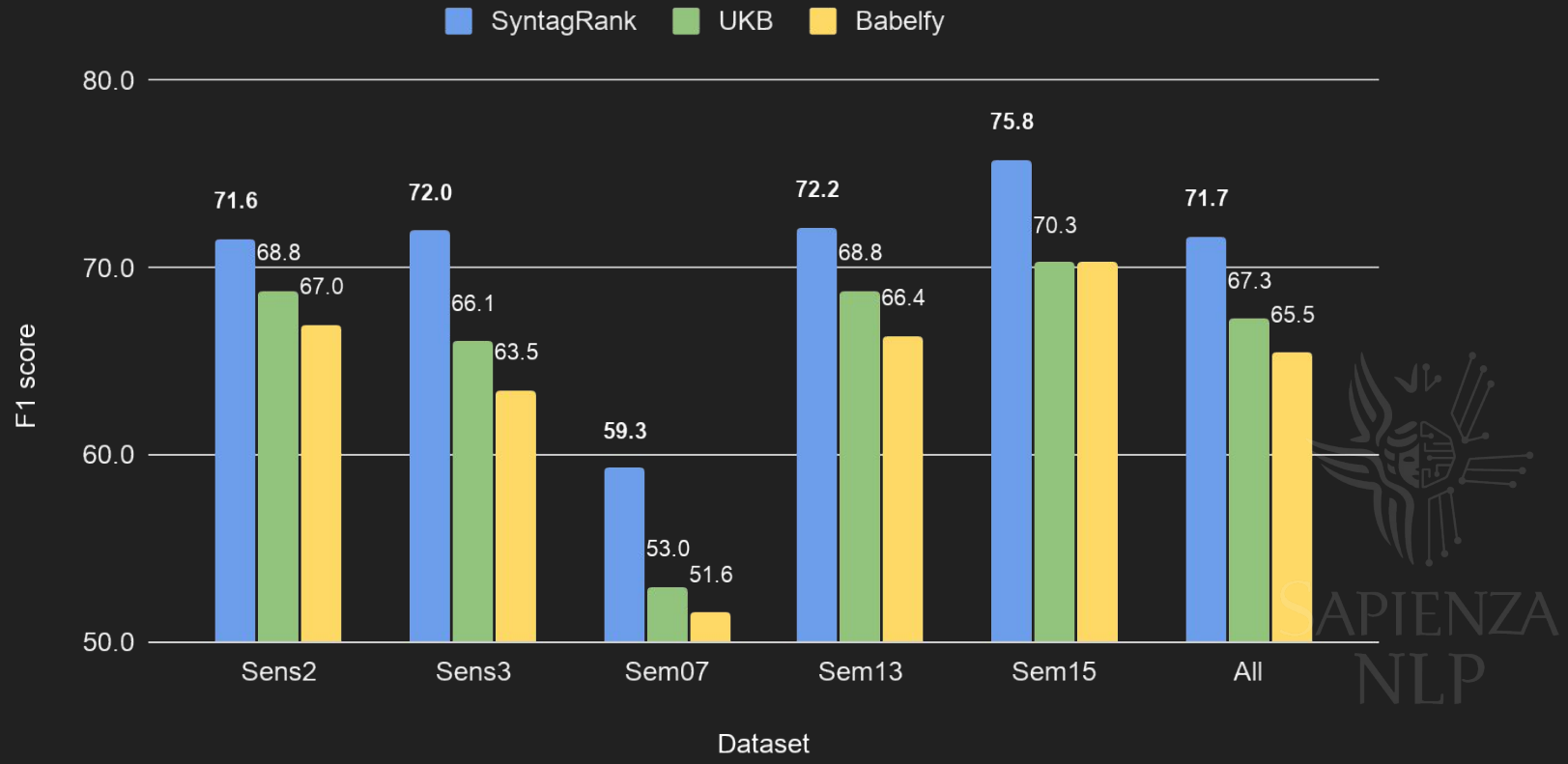
Competitor systems:

- Babelfy (Moro et al., 2014)
- UKB (Agirre et al., 2014)

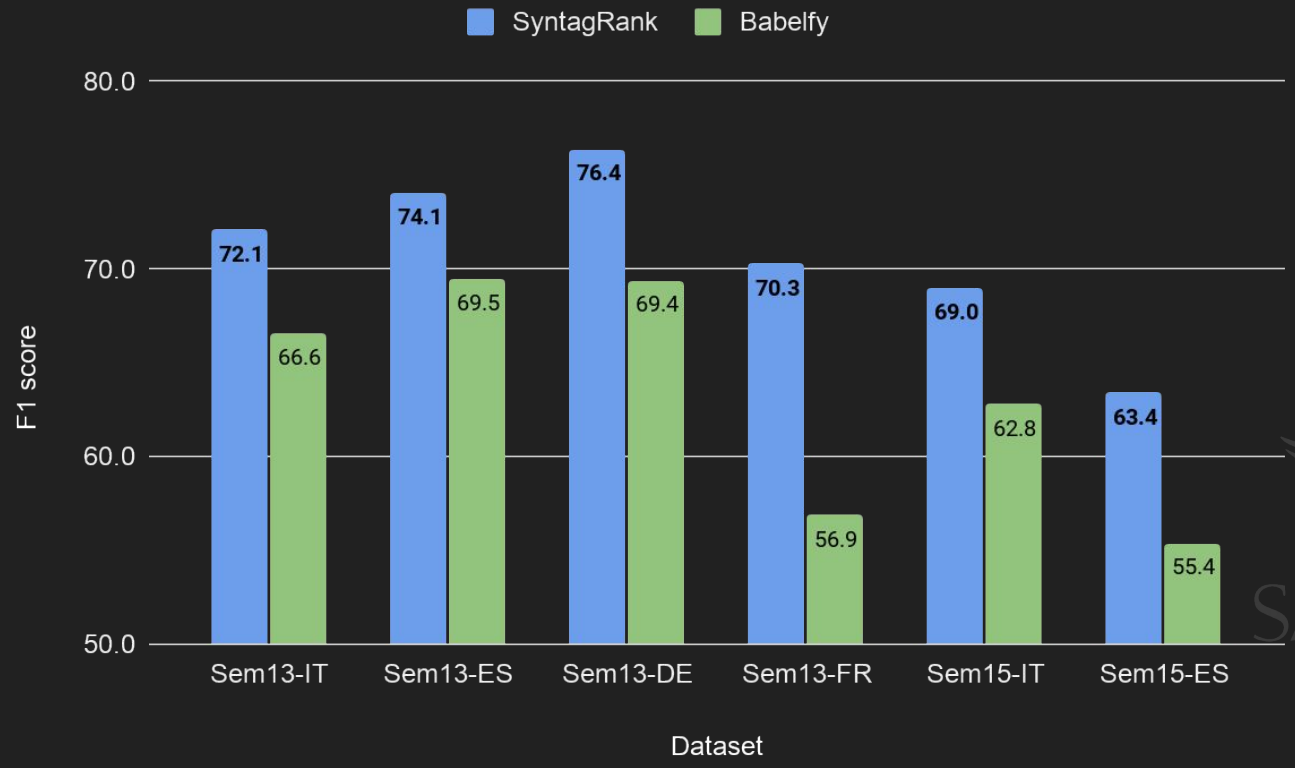
The multilingual evaluation leverages the German, Spanish, French and Italian annotations available in the amended version of the SemEval-2013 and SemEval-2015 evaluation datasets.

Evaluation data available at https://github.com/SapienzaNLP/mwsd-datasets.

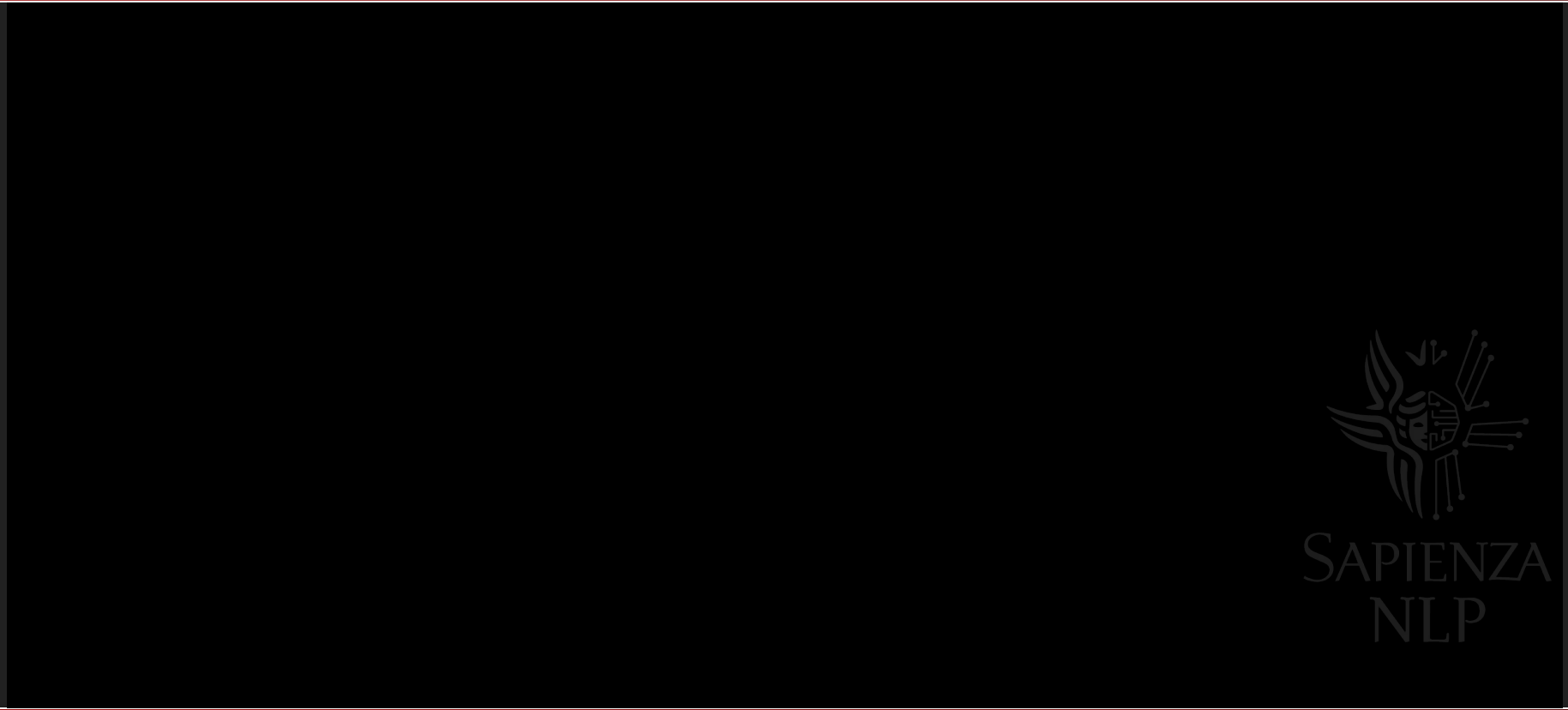
English WSD Evaluation



Multilingual WSD Evaluation



The SyntagRank Web Interface



Usage of the RESTful service

It is possible to query **SyntagRank** programmatically through a RESTful service.

The APIs come with two different methods:

- `disambiguate`: processes a raw text provided as input
- `disambiguate_tokens`: accepts a pre-processed text as input to be disambiguated

Both methods require the input language to be specified.

Full APIs documentation available at: http://syntagnet.org/api-documentation

Thank you for your attention!

SCAN ME!

Try SyntagRank at http://syntagnet.org/

Come visit us at http://nlp.uniroma1.it/

SAPIENZA UNIVERSITÀ DI ROMA

SAPIENZA NLP

Babelscape

elexis european lexicographic infrastructure

ELEXIS project No. 731015

erc

Consolidator Grant MOUSSE No. 726487