CluBERT:
A Cluster-Based Approach for Learning Sense Distributions in Multiple Languages

Tommaso Pasini, Federico Scozzafava and Bianca Scarlini
Sapienza NLP Group
{pasini, scozzafava, scarlini}@di.uniroma1.it

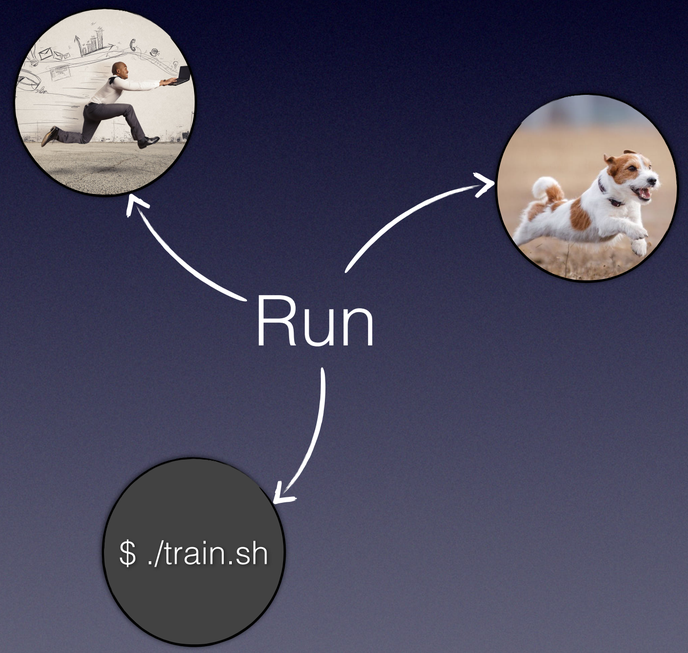
Supported by the ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research and innovation programme

CluBERT:
A Cluster-Based Approach for Learning Sense Distributions in Multiple Languages

Tommaso Pasini, Federico Scozzafava and Bianca Scarlini
Sapienza NLP Group
{pasini, scozzafava, scarlini}@di.uniroma1.it

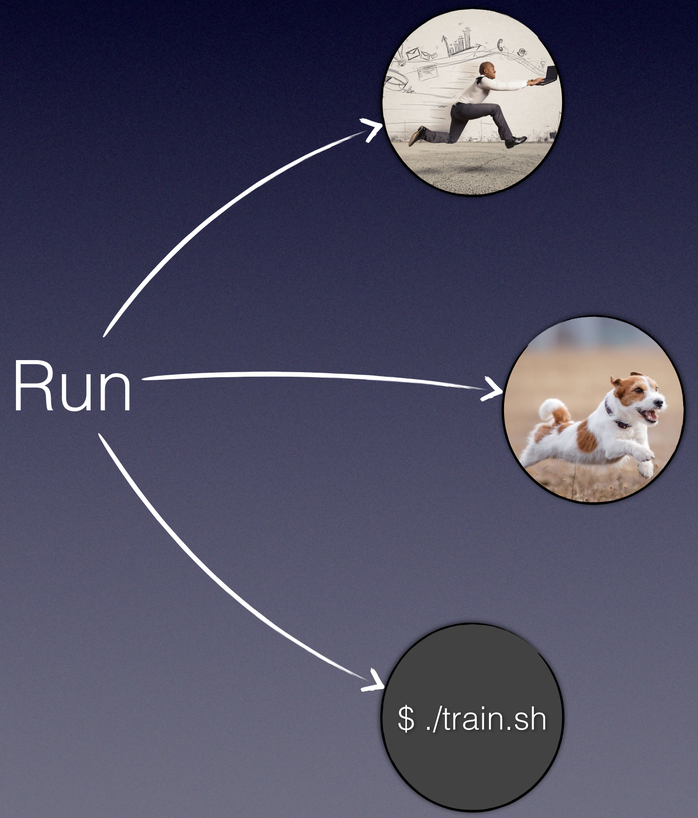
Supported by the ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research and innovation programme

Words are Ambiguous



Words are Ambiguous

A word sense is determined by the context the word occurs in.



- In the UK there is no requirement to obtain a licence to ***run*** a business.
- A person who intends to ***run*** a business can pledge a company mortgage.
- They seemed unable to ***run*** a business at a profit.

- They try to ***run*** to the other side.
- Dogs were ***running*** in the park that day.
- I ***ran*** to catch the train this morning.

- ***Run*** the program for some number of steps and check if it halts.
- The program loader is instructed to ***run*** the program /bin/sh.
- I asked her if she ***ran*** the script to start the training.

Sense Distribution

Sense Distribution

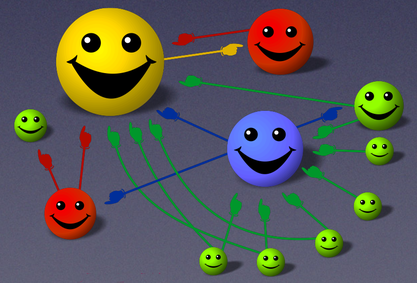
Can we **automatically extract** the **distribution** of word **senses** from a corpus of **raw texts**?

Tools & Resources

BabelNet, a multilingual knowledge base connecting meanings through semantic relations and having each meaning lexicalized in different languages.

BERT, a transformer-based neural model that we use to extract latent and contextualized representations of words.

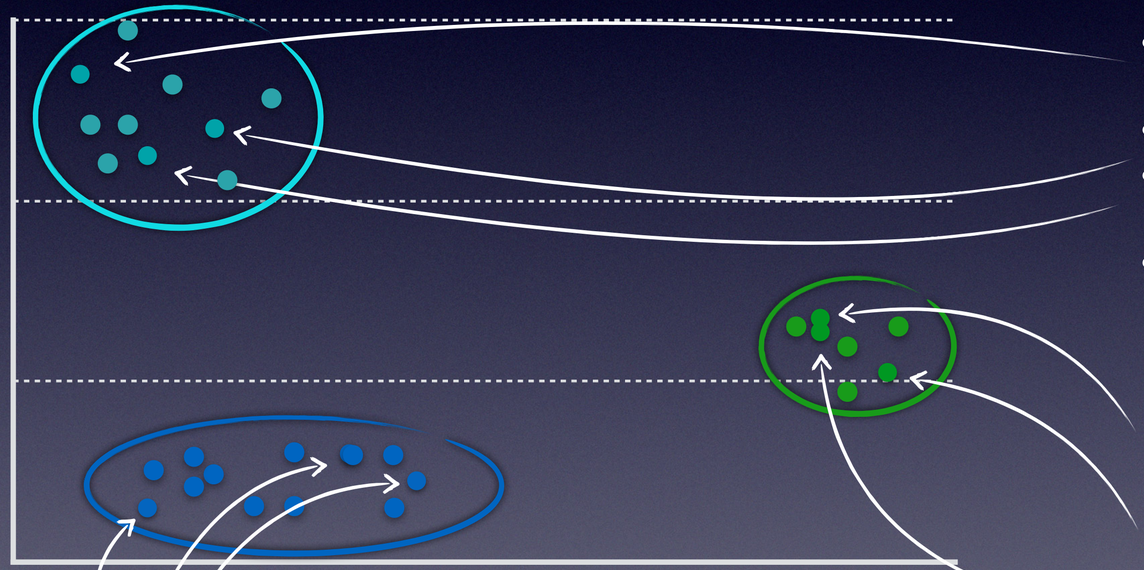
UKB, a knowledge-based approach for WSD. Given a set of words, runs the Personalized PageRank on a knowledge base, e.g., BabelNet, and scores each meaning therein.



Wikipedia, a large corpus of texts in different languages.

CluBERT (1)

Leverages contextualised word embeddings to cluster word occurrences in Wikipedia sentences.



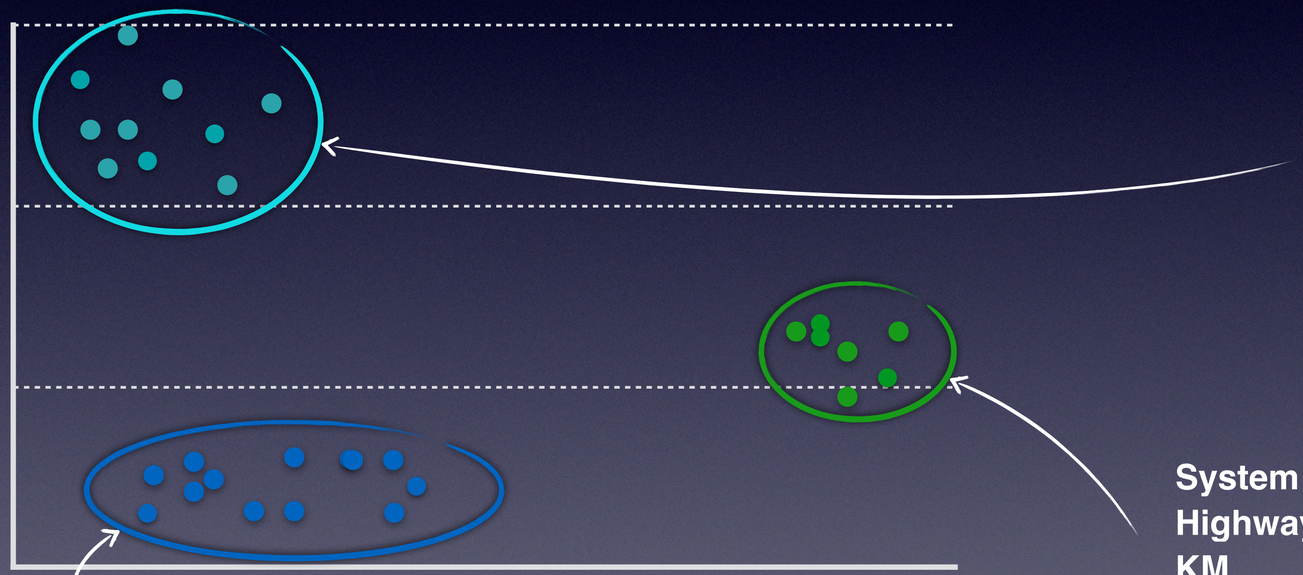
- Since 2002, it has been wholly staffed and ***run*** by local people.
- Health post is ***run*** by a health assistant .
- He completes his education in London and comes back to ***run*** his business.
- …

- The service was ***running*** and accepting requests.
- ***Running*** programs from the bash is possible on any linux OS.
- the operating system switches the processor to ***run*** another program.
- …

- You would ***run*** to the store .
- He was also able to ***run*** for 83 yards in the game.
- It can ***run*** at full throttle for 11 min on one charge .
- …

CluBERT (2)

Extracts the most peculiar word from each cluster.



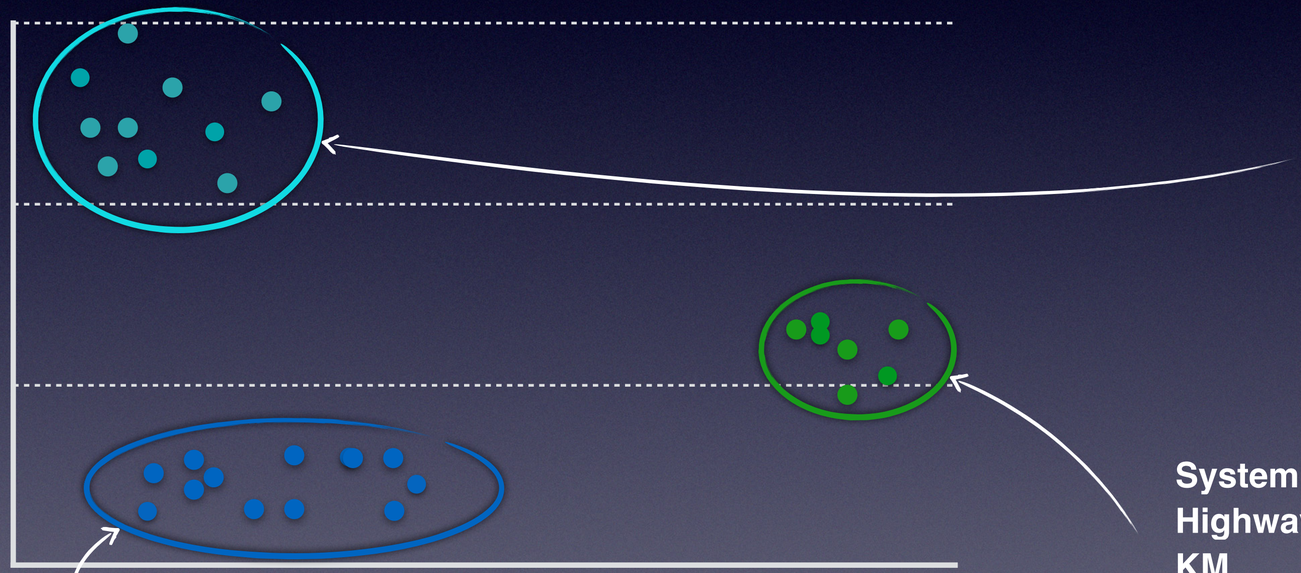
Business	375
Company	201
School	142
Store	41
Shop	18

System	112
Highway	83
KM	50
Mile	21
Street	12

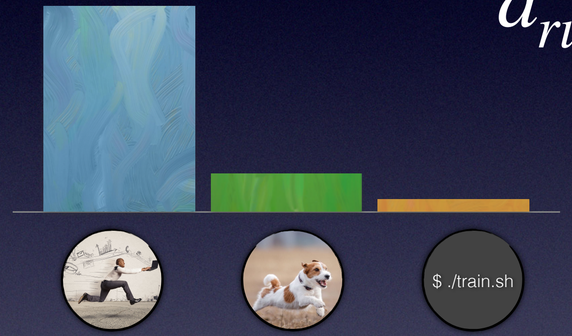
Route	200
Highway	123
KM	74
Mile	39
Street	20

CluBERT (3)

Uses the clusters' BOWs to compute an in-cluster sense distribution for the target word ***run***.

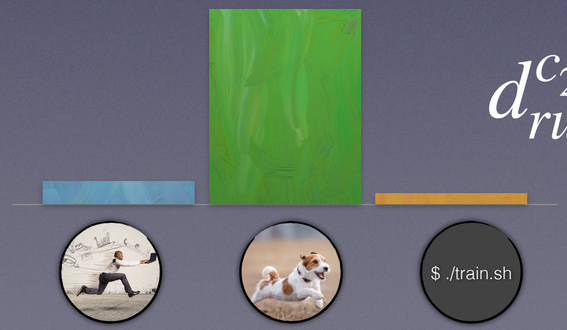


Business	375
Company	201
School	142
Store	41
Shop	18



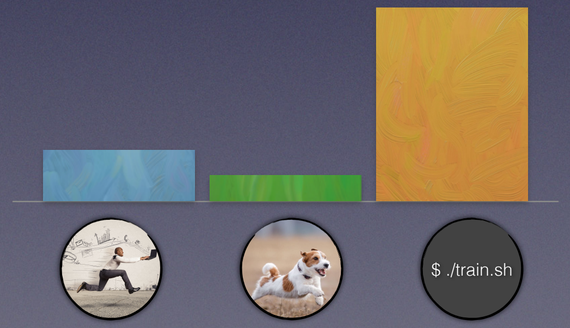
$d^{c_1}_{run}$

Route	200
Highway	123
KM	74
Mile	39
Street	20



$d^{c_2}_{run}$

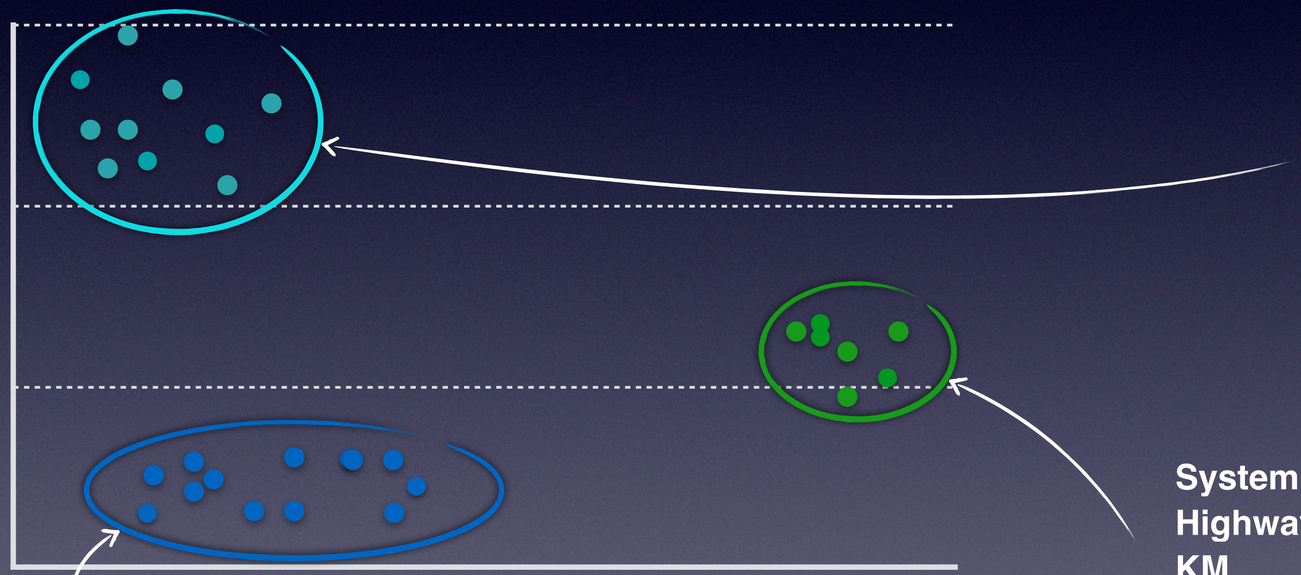
System	112
Highway	83
KM	50
Mile	21
Street	12



$d^{c_3}_{run}$

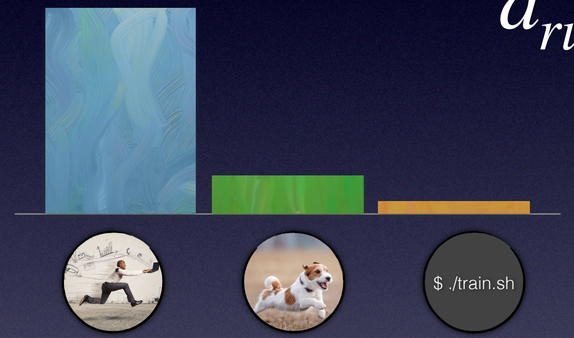
CluBERT (4)

Aggregates the informations and computes the final distribution of senses for the target word ***run***.



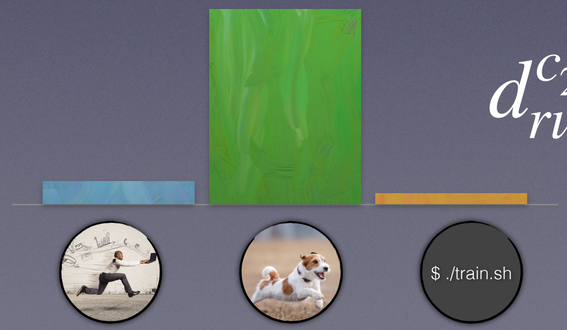
Business	375
Company	201
School	142
Store	41
Shop	18

$d^{c_1}_{run}$



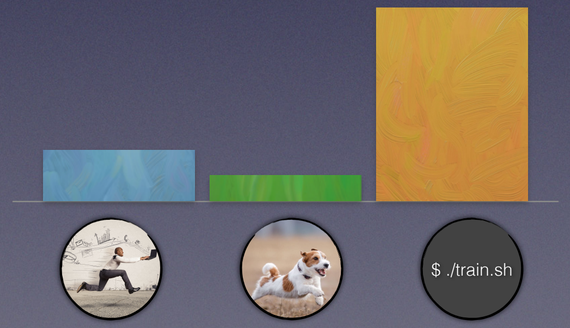
Route	200
Highway	123
KM	74
Mile	39
Street	20

$d^{c_2}_{run}$



System	112
Highway	83
KM	50
Mile	21
Street	12

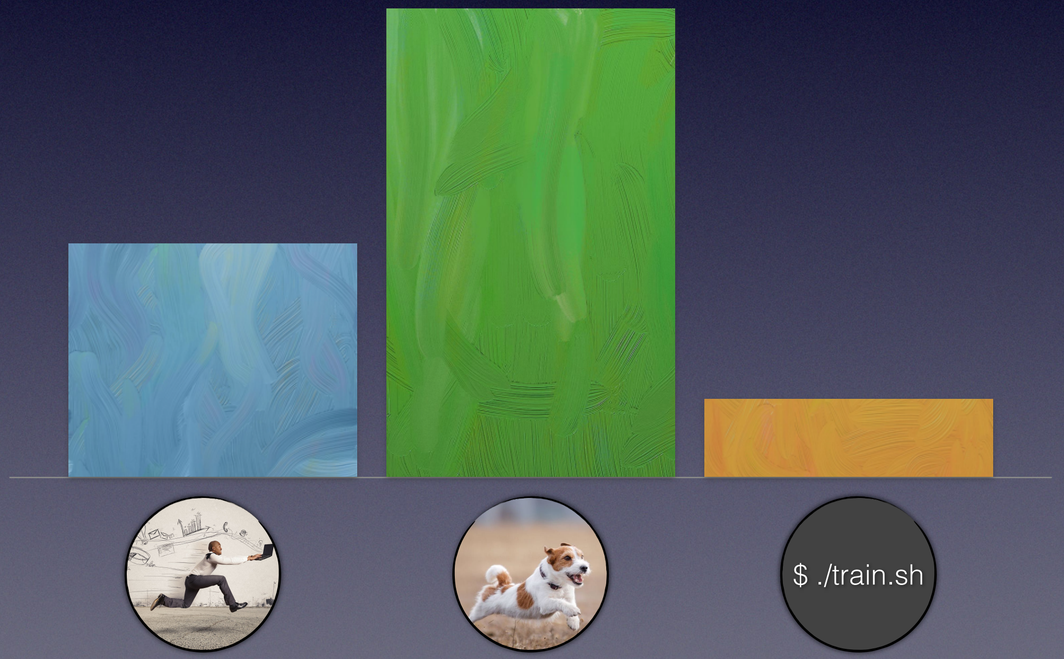
$d^{c_3}_{run}$



CluBERT (4)

Aggregates the informations and computes the final distribution of senses for the target word ***run*** (d_{run})**.**

$$d_{run} = \frac{\sum_{c_i \in \mathcal{U}_{run}} |c_i| \, d_{run}^{c_i}}{\sum_{c_i \in \mathcal{U}_{run}} |c_i|} =$$



How can we measure if our distributions are good?

Divergence Experiment

We measure the difference between our automatically-induced distributions and a manually-curated distribution

Divergence Experiment
CluBERT Setting

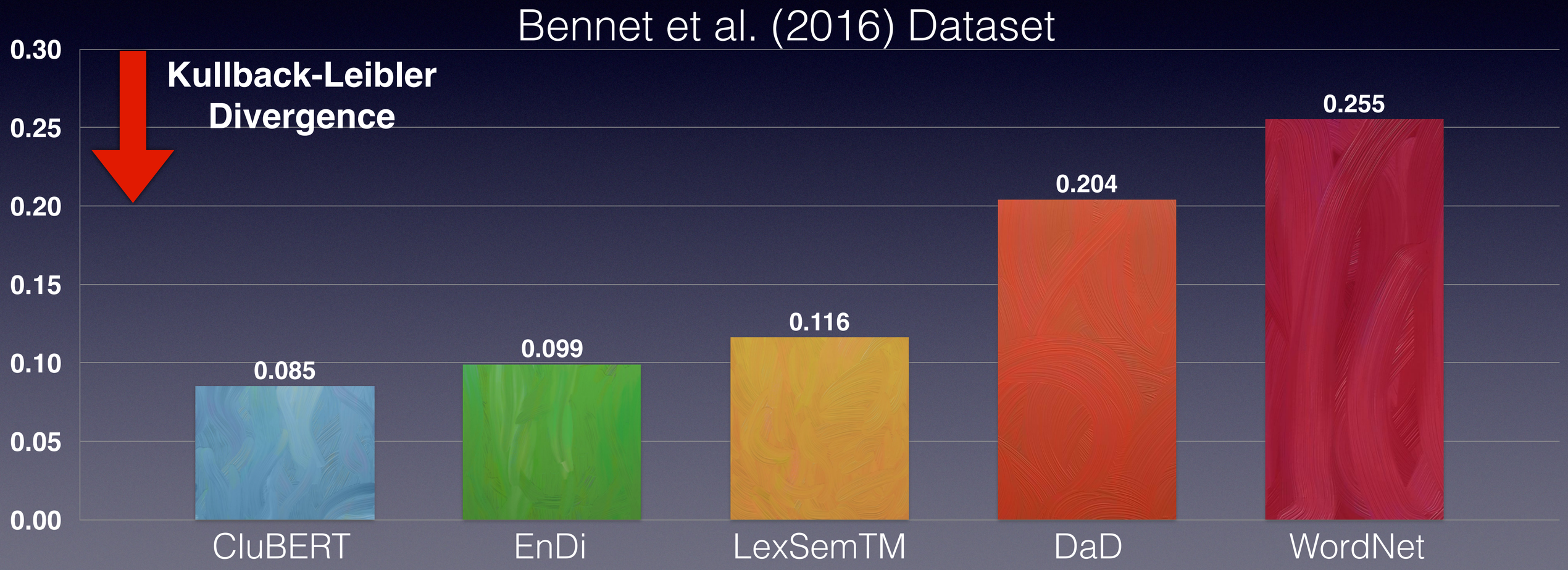
- We used BERT multilingual to represent word's occurrences.
- Cluster BOW size: 5
- *k*-means: we set *k* to the number of senses in WordNet of the target word.

WSD Experiment
Dataset & Competitors

- Dataset
  - **Bennet et al. (2016) dataset,** provides manually-annotated distributions for 50 distinct English lemmas.
- Competitors:
  - **EnDi & DaD**, two knowledge-based approaches.
  - **LexSemTM,** a topic-modelling approach.
  - **WordNet,** distribution of senses according to their frequency in SemCor.

Divergence Experiment

Results - **The Lower The Better**



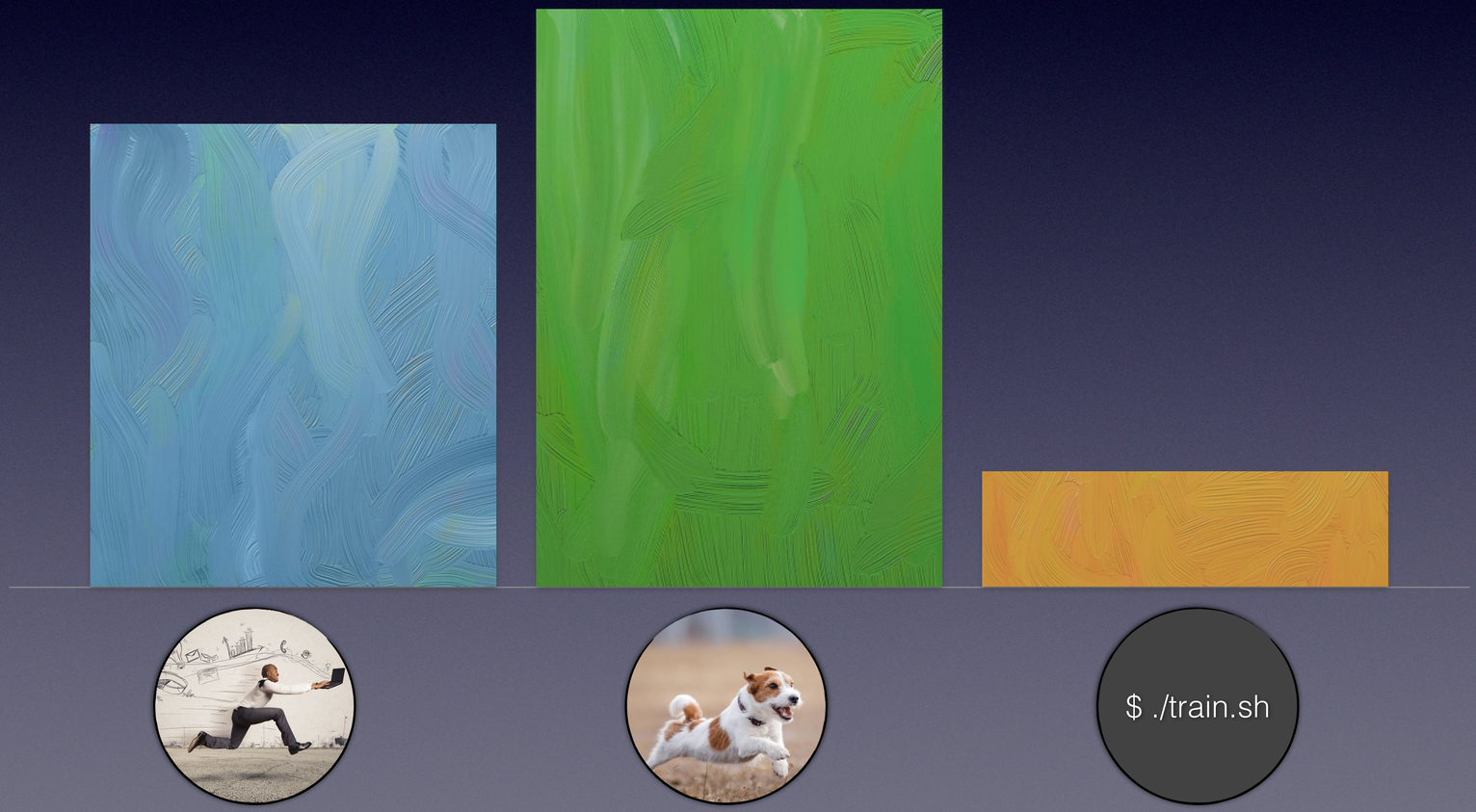
WSD Experiment

We evaluate CluBERT's Most Frequent Sense in the Word Sense Disambiguation task.

- You would ***run*** to the store.
- He was also able to ***run*** for 83 yards in the game.
- It can ***run*** at full throttle for 11 min on one charge.
- I went out while the program was ***running***.
- I ***ran*** to catch the train today.
- You better ***run*** if you want to arrive in time.
- Kids were ***running*** in the garden the whole afternoon.
- She decided to ***run*** her own business.

The Most Frequent Sense

Is the word sense in which the word is used in most of its occurrences.



- ✓ You would ***run*** to the store .
- ✓ He was also able to ***run*** for 83 yards in the game.
- ✓ It can ***run*** at full throttle for 11 min on one charge.
- – I went out while the program was ***running***.
- ✓ I ***ran*** to catch the train today.
- ✓ You better ***run*** if you want to arrive in time.
- ✓ Kids were ***running*** in the garden the whole afternoon.
- – She decided to ***run*** her own business.

WSD Experiment

Competitors

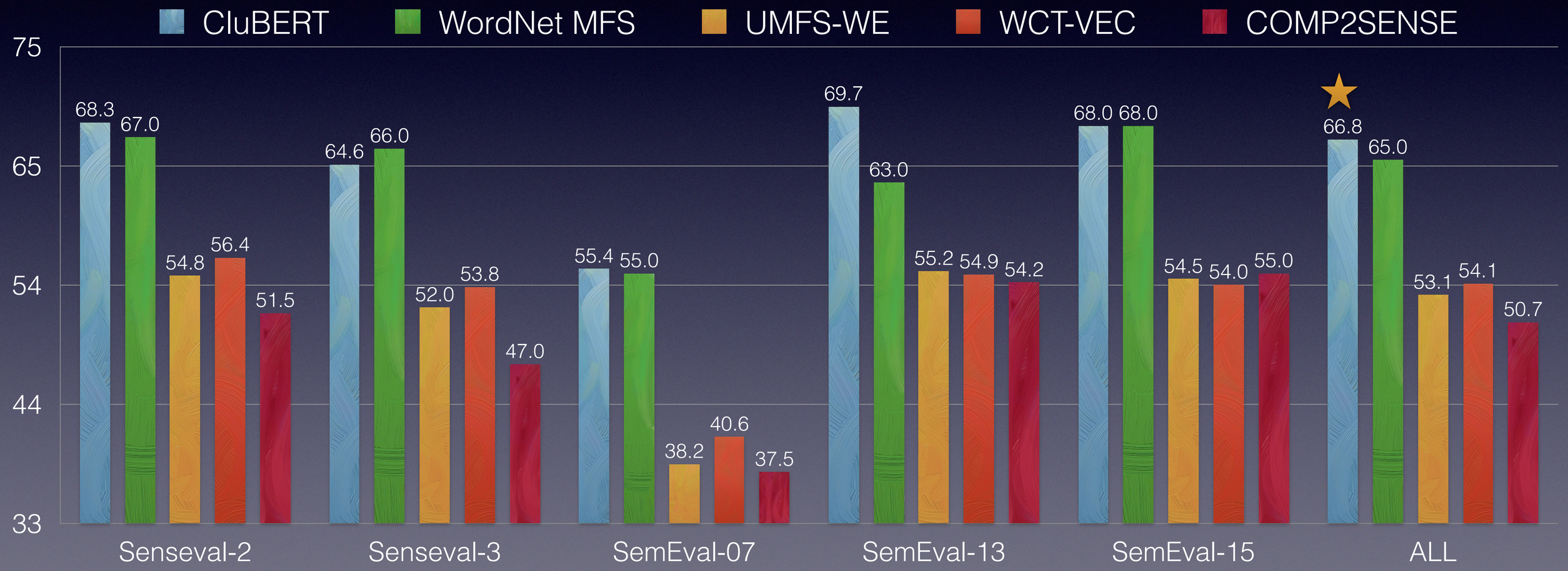
- **WordNet**, is the most widely used knowledge base. Word senses are sorted according to their frequency in SemCor.
- **COMP2SENSE**, a knowledge-based approach relying on the distance between a word and a sense within a semantic network.
- **UMFS-WE & WCT-WE**, two approaches based on the distance between words and sense embeddings.

WSD Experiment
Datasets

- We computed the **F1** on the standard benchmark for the all-words **English Word Sense Disambiguation** task:
  - Senseval-2
  - Senseval-3
  - SemEval-2007 task 17
  - SemEval-2013 task 12
  - SemEval-2015 task 13
  - ALL (the concatenation of all the above)

WSD Experiment

Results



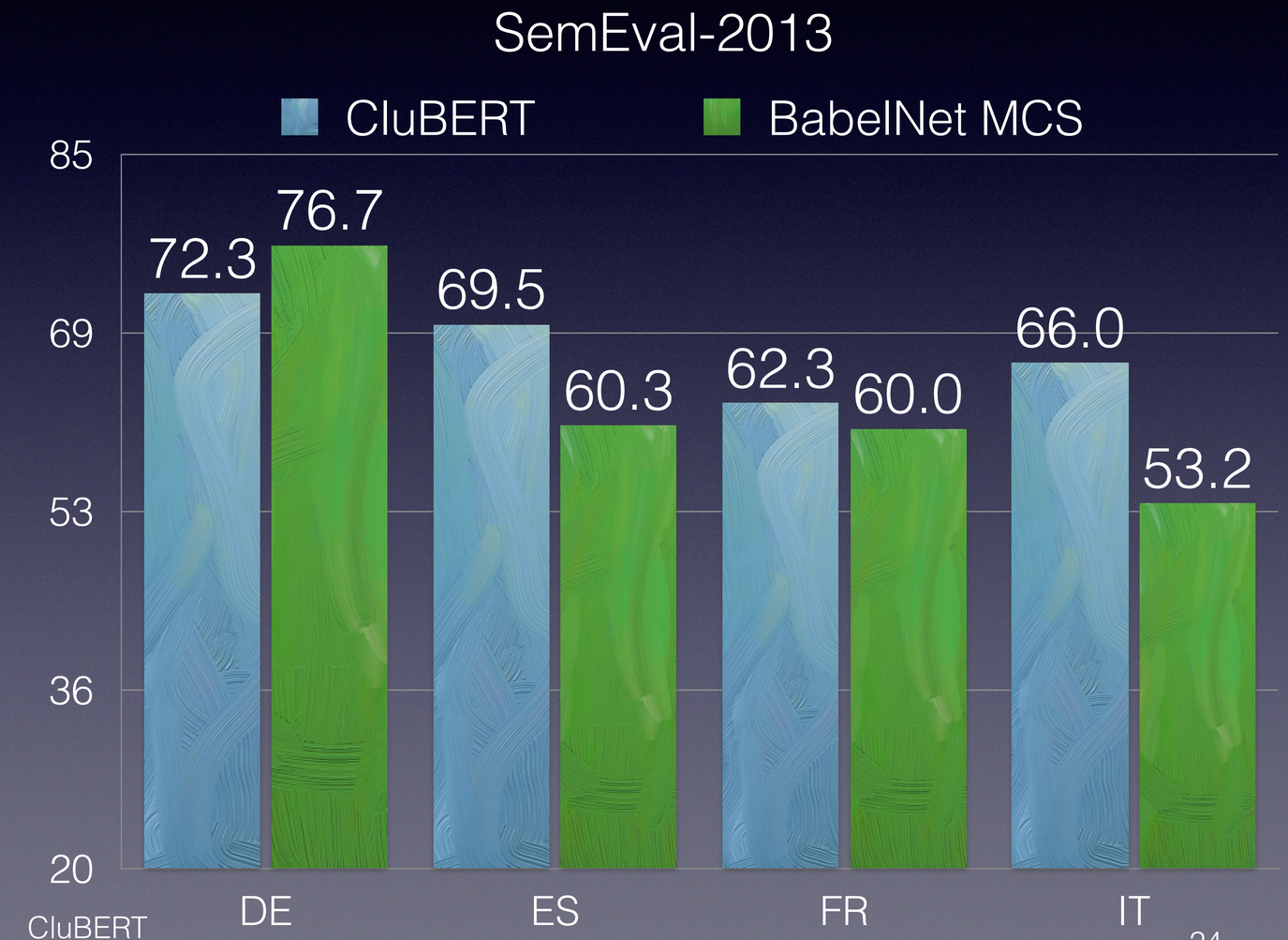
CluBERT
T. Pasini, F. Scozzafava and B. Scarlini

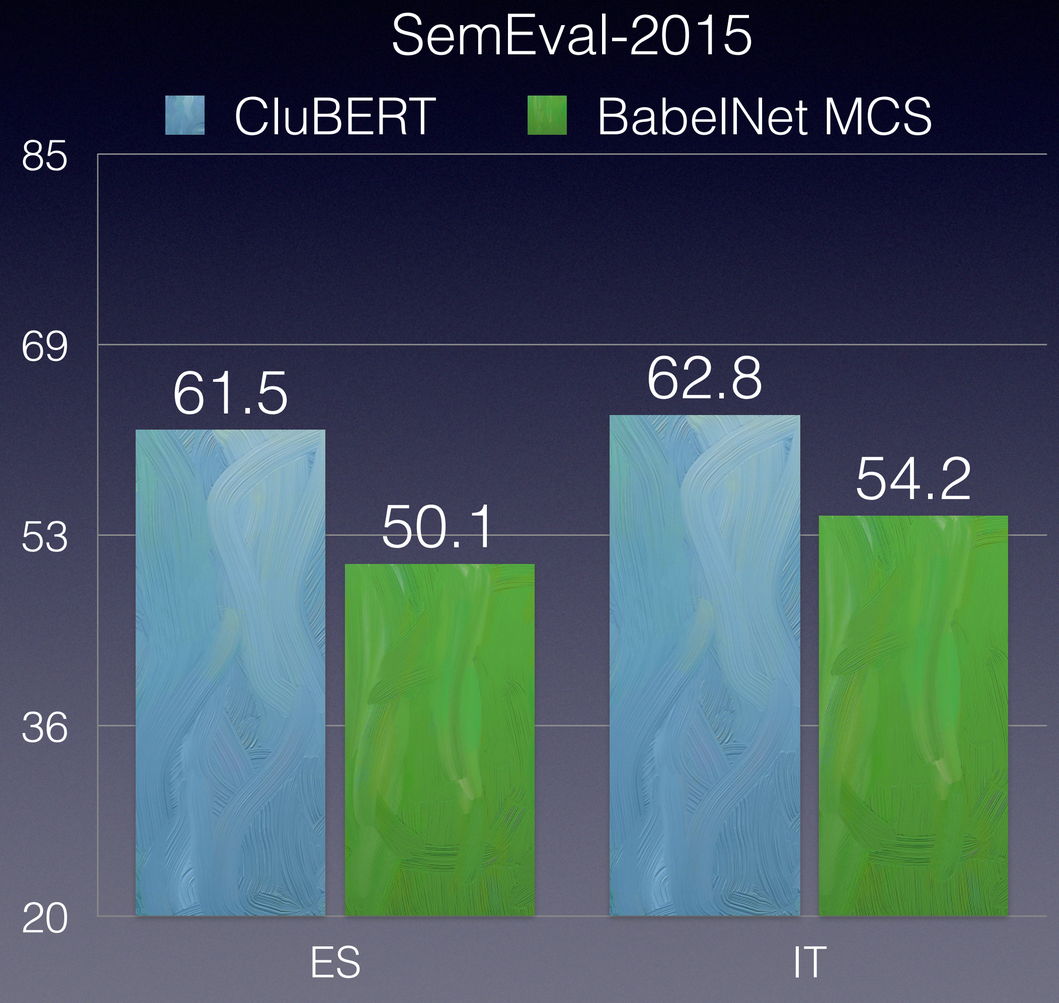
Multilingual WSD Experiment
Setting

- Datasets:
  - French, German, Italian and Spanish datasets of SemEval-2013 task 12.
  - Italian and Spanish datasets of SemEval-2015 task 13.
- Competitors:
  - **BabelNet MCS**, the most common sense according to BabelNet.

Multilingual WSD Experiment

Results





Conclusion

CluBERT is a multilingual approach to automatically induce sense distributions from a corpus of raw sentences.

CluBERT MFS attains better results than manually-curated approaches and beats all its competitors across languages.

We plan refine our approach to build cluster-level BOW and to leverage the clusters to create sense embeddings.

Data Available at: https://github.com/SapienzaNLP/clubert

Supported by the ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research and innovation programme.