Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison

Alessandro Raganato, José Camacho Collados
and Roberto Navigli

Dipartimento
di Informatica

Sapienza
Università di Roma

lcl.uniroma1.it/wsdeval

Word Sense Disambiguation (WSD)

Given the word in context, find the correct sense:

The **mouse** ate the cheese.

A **mouse** consists of an object held in one's hand, with one or more buttons.

International Workshops on Semantic Evaluation

Many evaluation datasets have been constructed for the task:

- Senseval 2 (2001)
- Senseval 3 (2004)
- SemEval 2007
- SemEval 2013
- SemEval 2015

International Workshops on Semantic Evaluation

Many evaluation datasets have been constructed for the task:

- Senseval 2 (2001) WN 1.7
- Senseval 3 (2004) WN 1.7.1
- SemEval 2007 WN 2.1
- SemEval 2013 WN 3.0
- SemEval 2015 WN 3.0

Problem:

- different formats, construction guidelines and sense inventory

Building a Unified Evaluation Framework

Our goal:

- build a unified framework for all-words WSD (training and testing)
- use this evaluation framework to perform a fair quantitative and qualitative empirical comparison

Building a Unified Evaluation Framework

Our goal:

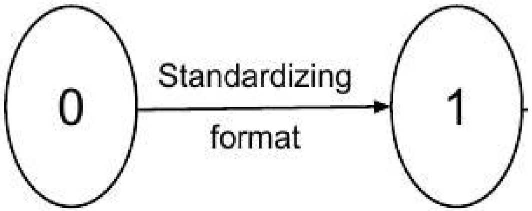
- build a unified framework for all-words WSD (training and testing)
- use this evaluation framework to perform a fair quantitative and qualitative empirical comparison

How:

- standardizing the WSD datasets and training corpora into a unified format
- semi-automatically converting annotations from any dataset to WordNet 3.0
- preprocessing the datasets by consistently using the same pipeline.

Building a Unified Evaluation Framework

Pipeline for standardizing any given WSD dataset:

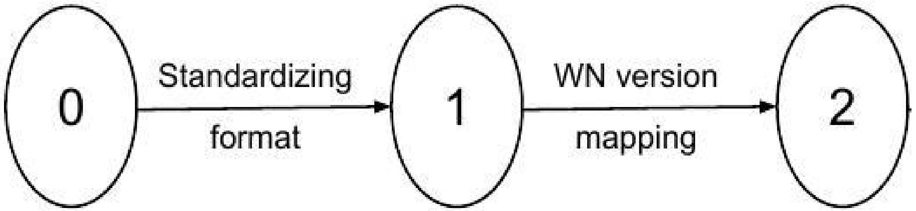


Standardizing format:

- convert all datasets to a unified XML scheme, where preprocessing information (e.g. lemma, PoS tag) of a given corpus can be encoded

Building a Unified Evaluation Framework

Pipeline for standardizing any given WSD dataset:



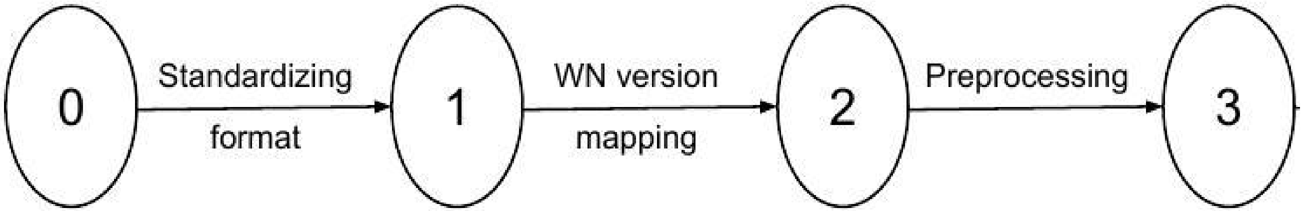
WN version mapping:

- map the sense annotations from its original WordNet version to 3.0
  - carried out semi-automatically (Daude et al., 2003)

Jordi Daude, Lluis Padro, and German Rigau. *Validation and tuning of wordnet mapping techniques*. In Proceedings of RANLP 2003.

Building a Unified Evaluation Framework

Pipeline for standardizing any given WSD dataset:

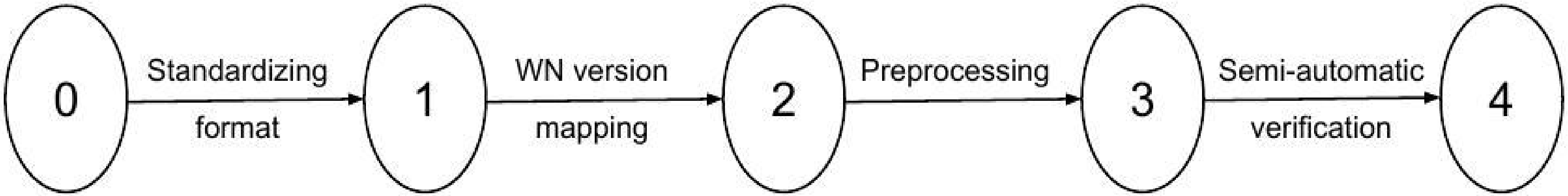


Preprocessing:

- use the Stanford coreNLP toolkit for part of speech tagging and lemmatization

Building a Unified Evaluation Framework

Pipeline for standardizing any given WSD dataset:



Semi-automatic verification:

- develop a script to check that the final dataset conforms to the guidelines
- ensure that the sense annotations match the lemma and the PoS tag provided by Stanford CoreNLP

Data - evaluation framework

- Training data:
  - **SemCor**, a manually sense-annotated corpus
  - **OMSTI** (One Million Sense-Tagged Instances), a large annotated corpus, automatically constructed by using an alignment based WSD approach

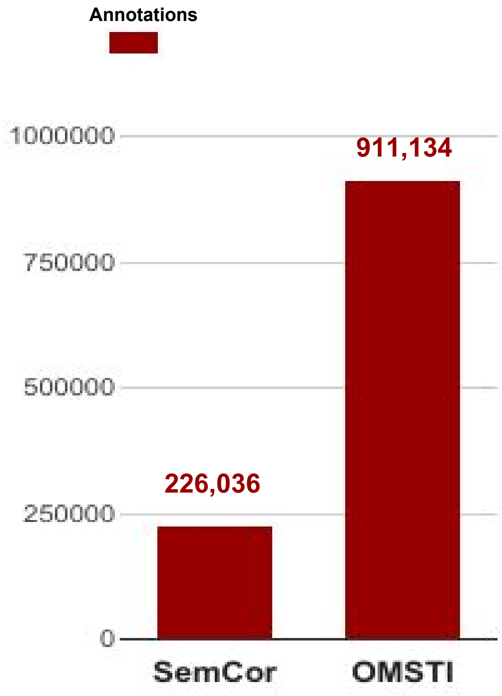
Data - evaluation framework

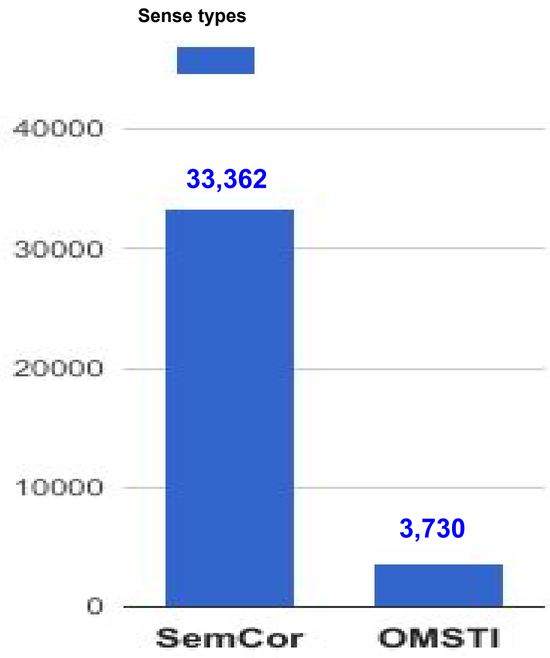
- Training data:
  - **SemCor**, a manually sense-annotated corpus
  - **OMSTI** (One Million Sense-Tagged Instances), a large annotated corpus, automatically constructed by using an alignment based WSD approach

- Testing data:
  - **Senseval 2**, covers nouns, verbs, adverbs and adjectives
  - **Senseval 3**, covers nouns, verbs, adverbs and adjectives
  - **SemEval 2007**, covers nouns and verbs
  - **SemEval 2013**, covers nouns only
  - **SemEval 2015**, covers nouns, verbs, adverbs and adjectives

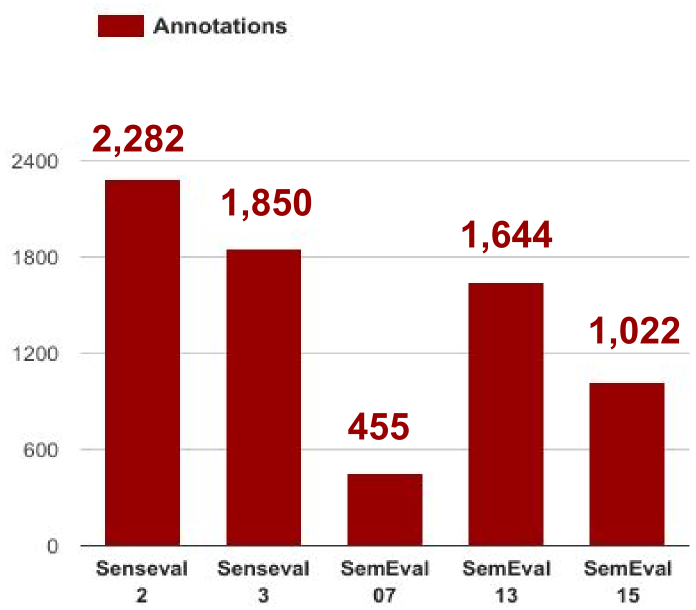
  - **ALL**, the concatenation of all five testing data

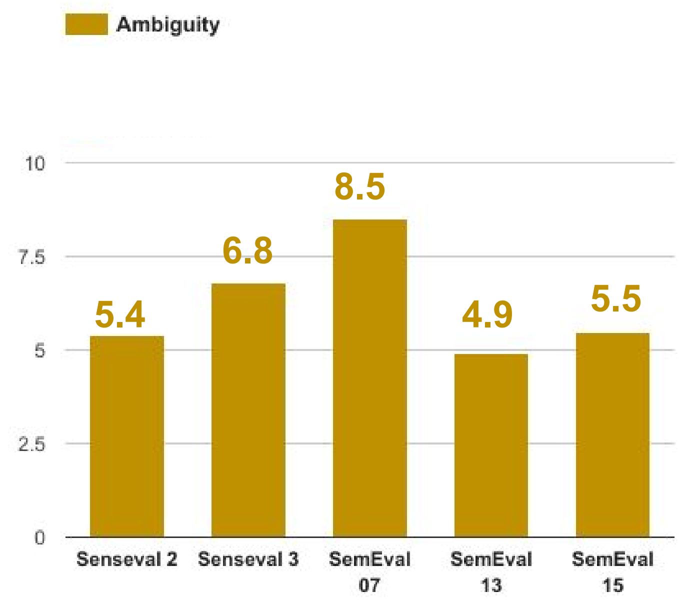
Statistics - training data





Statistics - testing data



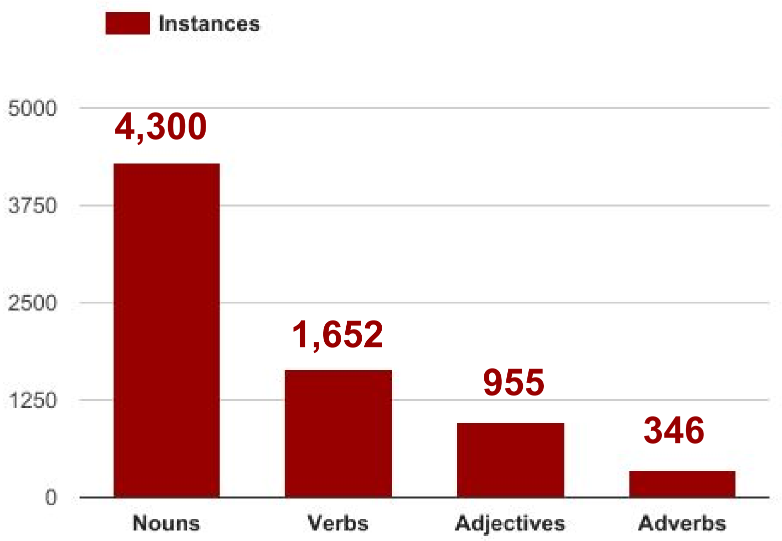


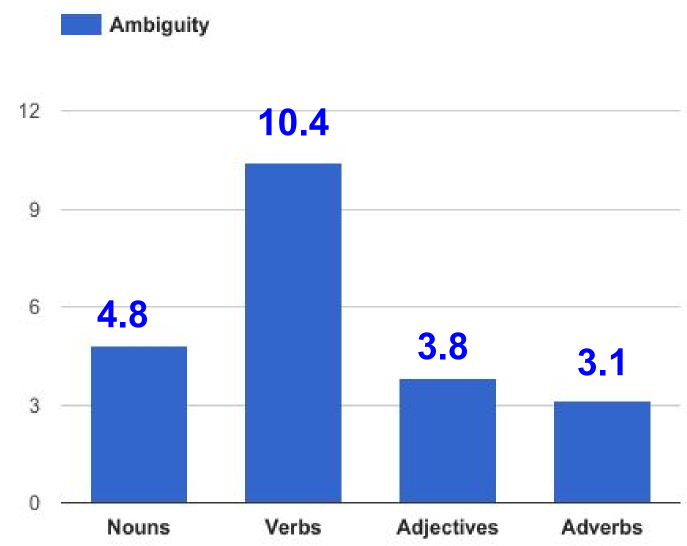
Statistics - testing data (ALL)

- **ALL**, the concatenation of all the five evaluation datasets
  - Total test instances: 7.253

Statistics - testing data (ALL)

- **ALL**, the concatenation of all the five evaluation datasets
  - Total test instances: 7.253





Evaluation

Evaluation: Comparison systems

- **Knowledge-based**

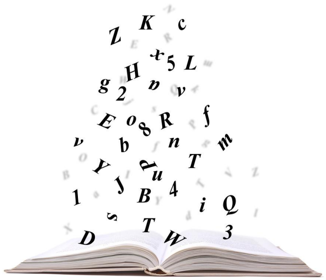
- **Supervised**

Evaluation: Comparison systems

- **Knowledge-based**
  - Lesk_extended (Banerjee and Pedersen, 2003)
  - Lesk+emb (Basile et al., 2014)
  - UKB (Agirre et al., 2014)
  - Babelfy (Moro et al., 2014)

Evaluation: Comparison systems (knowledge-based)

Lesk (Lesk, 1986)



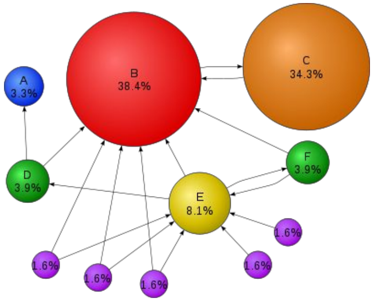
Based on the **overlap between the definitions of a given sense and the context of the target word**. Two configurations:

- ***Lesk_extended*** (Banerjee and Pedersen, 2003): it includes related senses and tf-idf for word weighting.

- **Lesk+emb** (Basile et al., 2014): enhanced version of Lesk in which similarity between definitions and the target context is computed via word embeddings.

Evaluation: Comparison systems (knowledge-based)

UKB (Agirre et al., 2014)

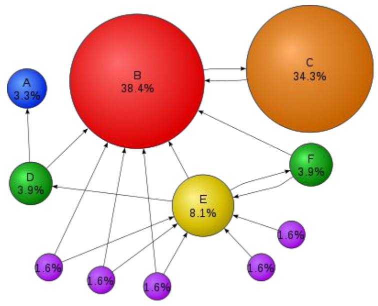


Graph-based system which exploits **random walks over a semantic network**, using Personalized PageRank.

It uses the standard WordNet graph plus disambiguated glosses as connections.

Evaluation: Comparison systems (knowledge-based)

UKB (Agirre et al., 2014)



Graph-based system which exploits **random walks over a semantic network**, using Personalized PageRank.

It uses the standard WordNet graph plus disambiguated glosses as connections.

NEW - UKB*: enhanced configuration using sense distributions from SemCor and running Personalized PageRank for each word.

Evaluation: Comparison systems (knowledge-based)

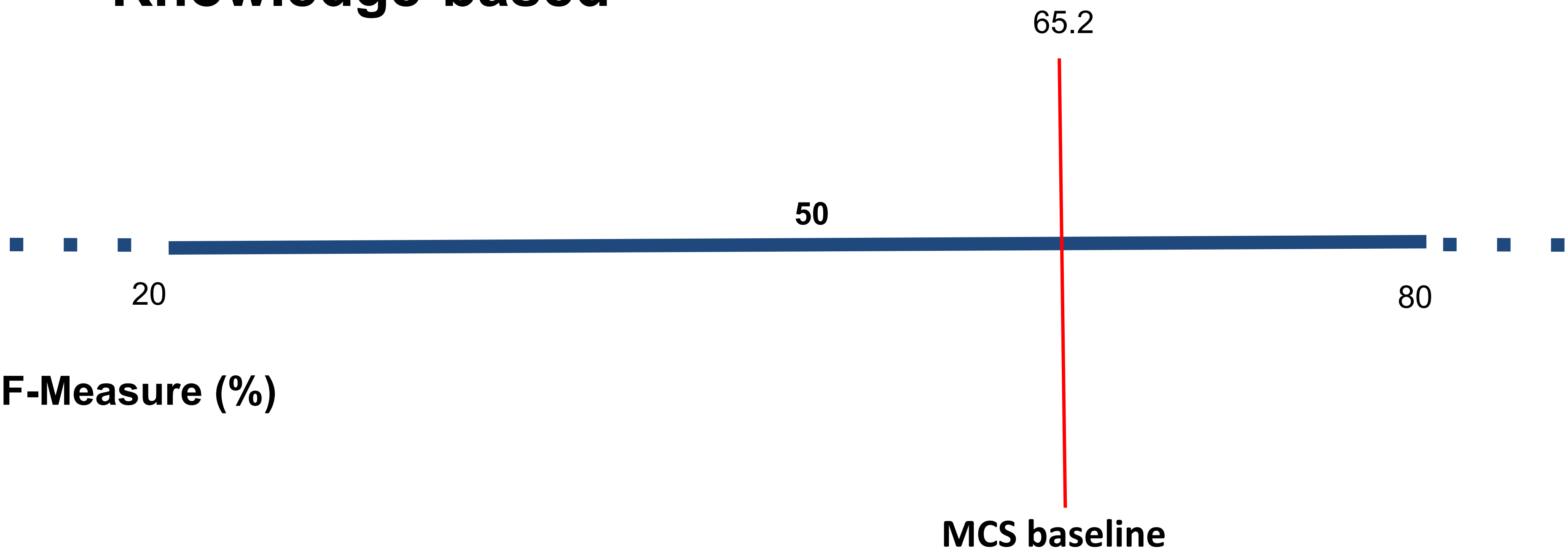
Babelfy (Moro et al., 2014)

Graph-based system that uses **random walks with restart** over a semantic network, creating high-coherence semantic interpretations of the input text.

BabelNet as semantic network. BabelNet provides a large set of connections coming from Wikipedia and other resources.

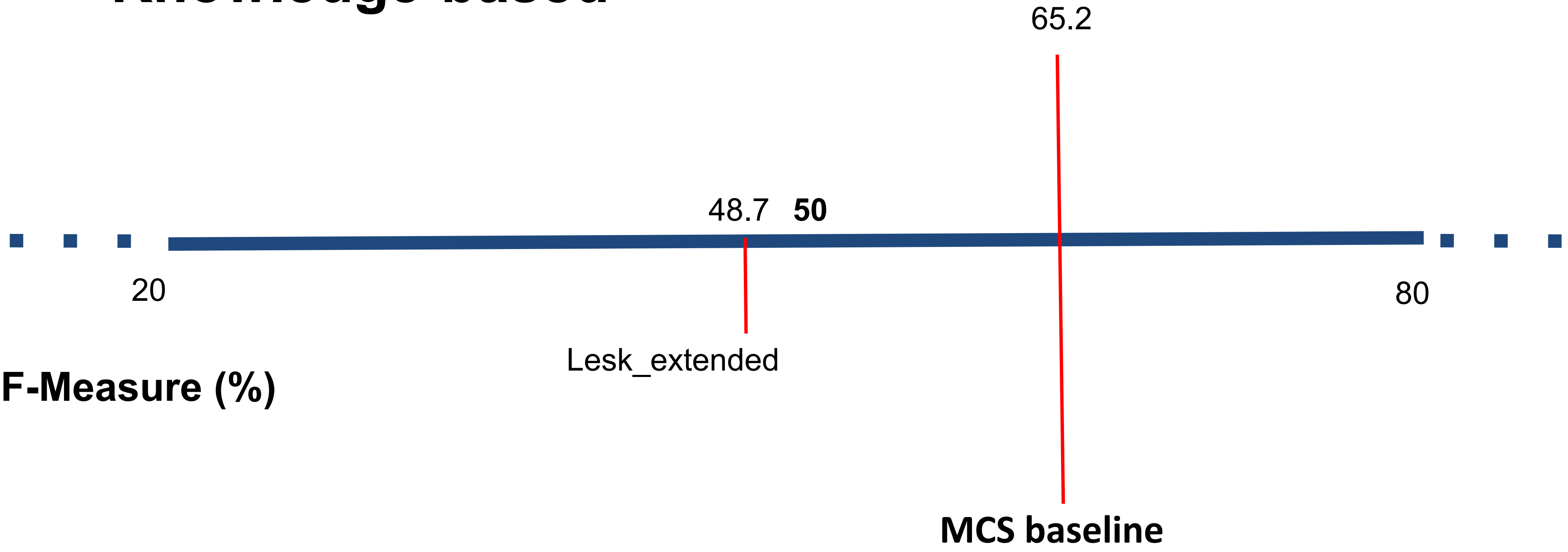
Evaluation: Results on the concatenation of all datasets

Knowledge-based



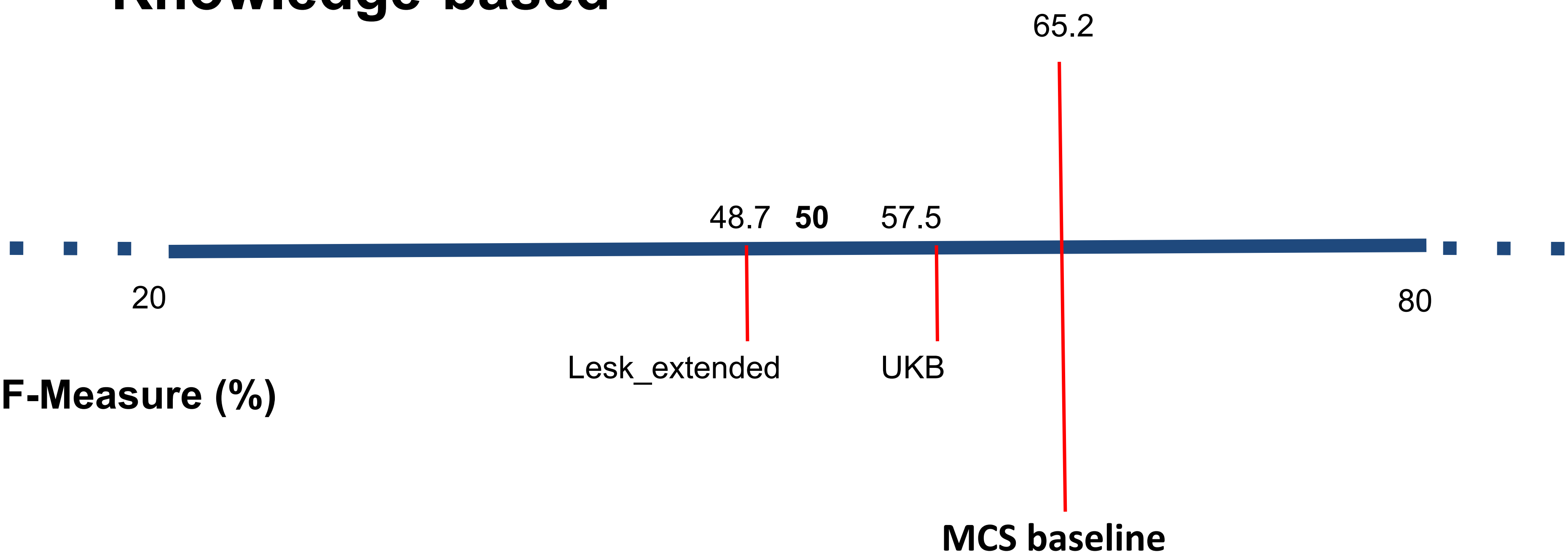
Evaluation: Results on the concatenation of all datasets

Knowledge-based



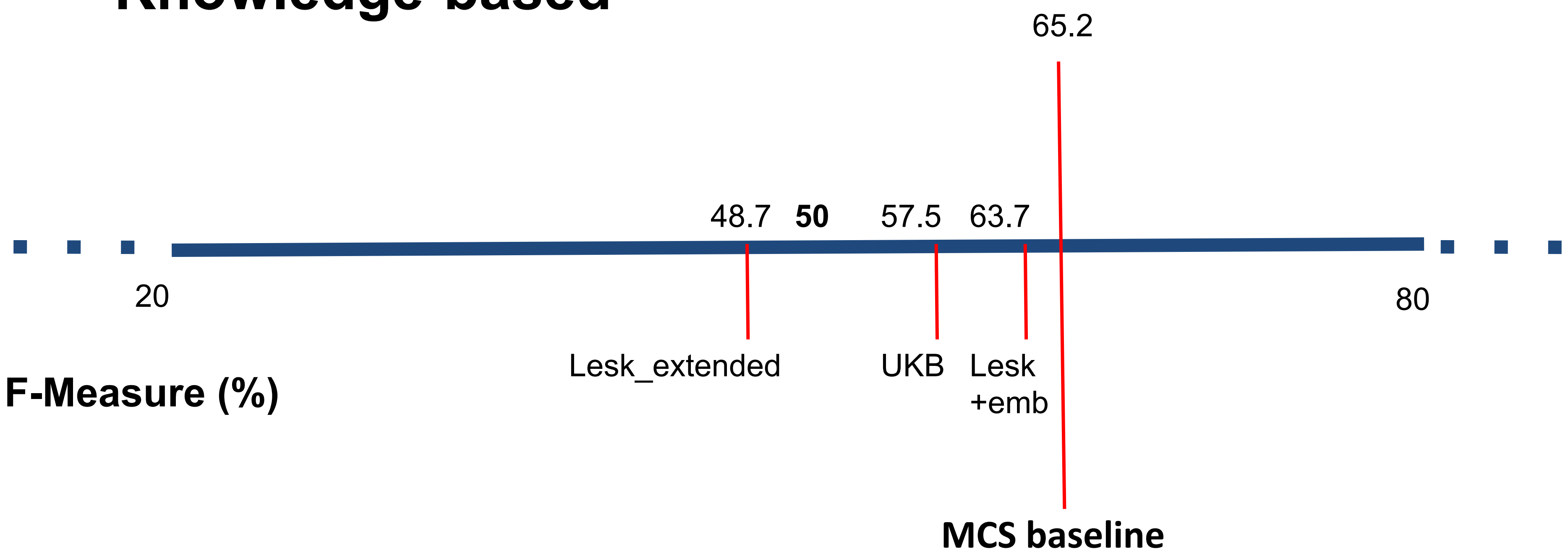
Evaluation: Results on the concatenation of all datasets

Knowledge-based



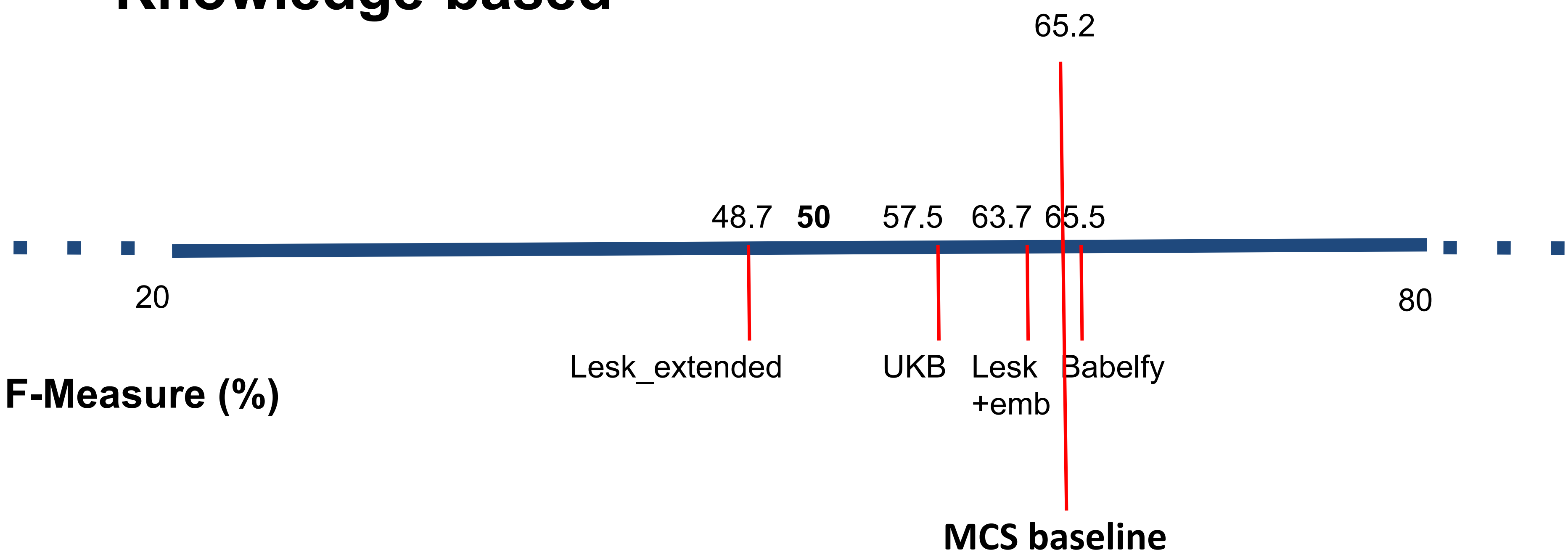
Evaluation: Results on the concatenation of all datasets

Knowledge-based

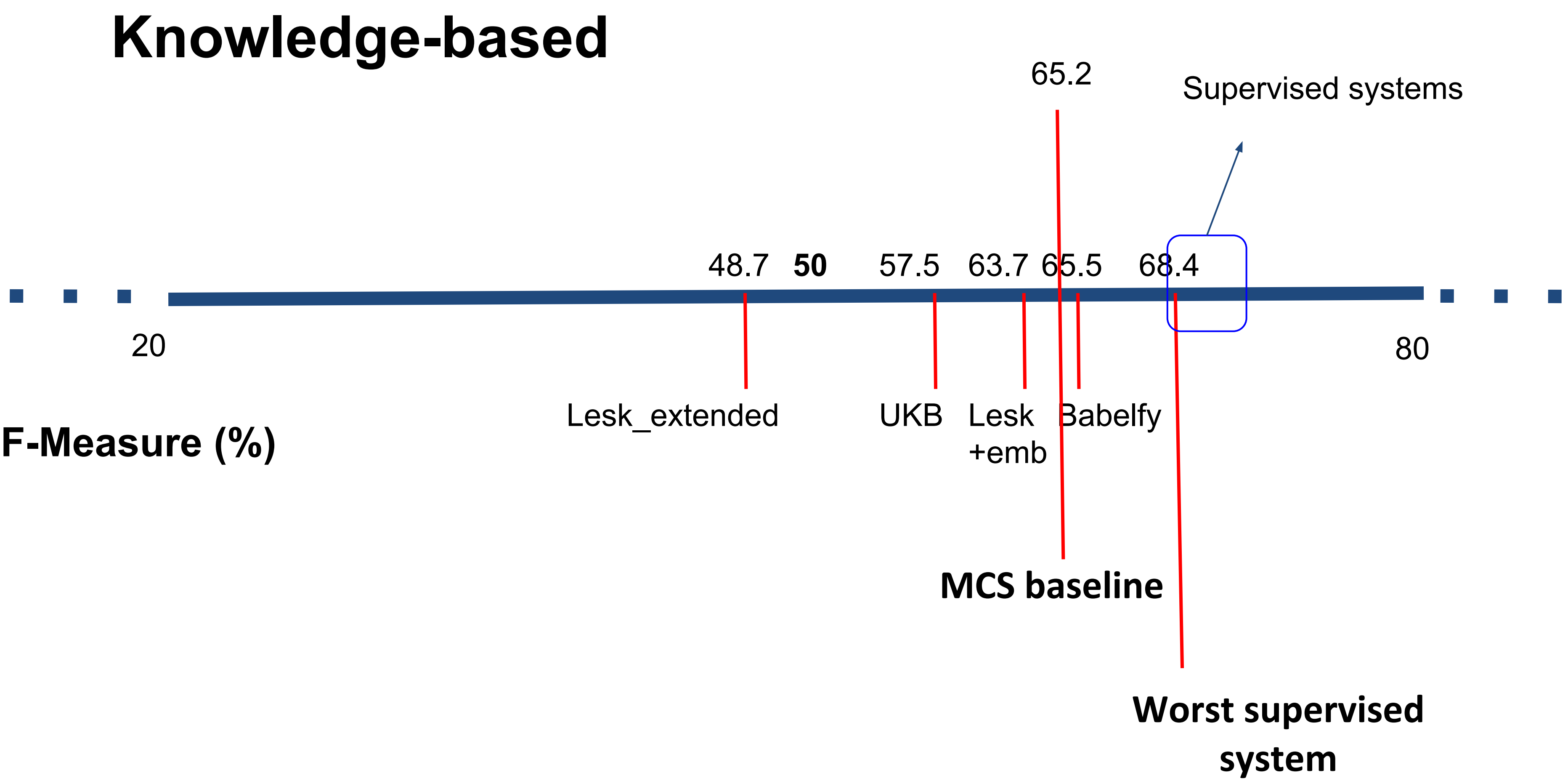


Evaluation: Results on the concatenation of all datasets

Knowledge-based



Evaluation: Results on the concatenation of all datasets

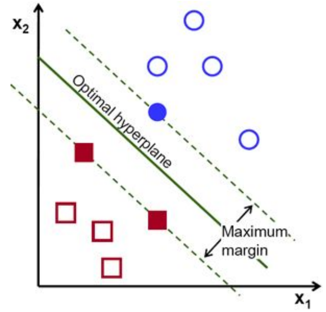


Evaluation: Comparison systems

- **Knowledge-based**
  - Lesk-extended (Banerjee and Pedersen, 2003)
  - Lesk+emb (Basile et al., 2014)
  - UKB (Agirre et al., 2014)
  - Babelfy (Moro et al., 2014)

- **Supervised**
  - IMS (Zhong and Ng, 2010)
  - IMS+emb (Iacobacci et al. 2016)
  - Context2Vec (Melamud et al., 2016)

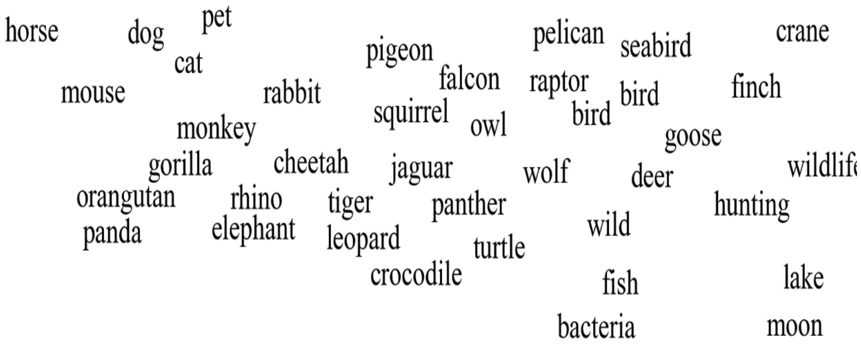
Evaluation: Comparison systems (supervised)



IMS (Zhong and Ng, 2010)

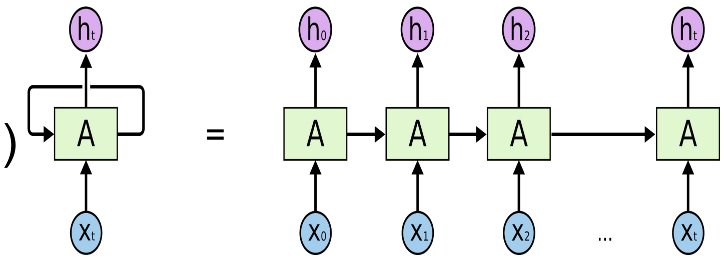
SVM classifier over a set of conventional features: surroundings words, PoS tags and local collocations.

Improvements integrating **word embeddings** as an additional feature (Taghipour and Ng, 2015; Rothe and Schütze, 2015; Iacobacci et al. 2016) -> IMS+emb.



Evaluation: Comparison systems (supervised)

Context2Vec (Melamud et al., 2016)

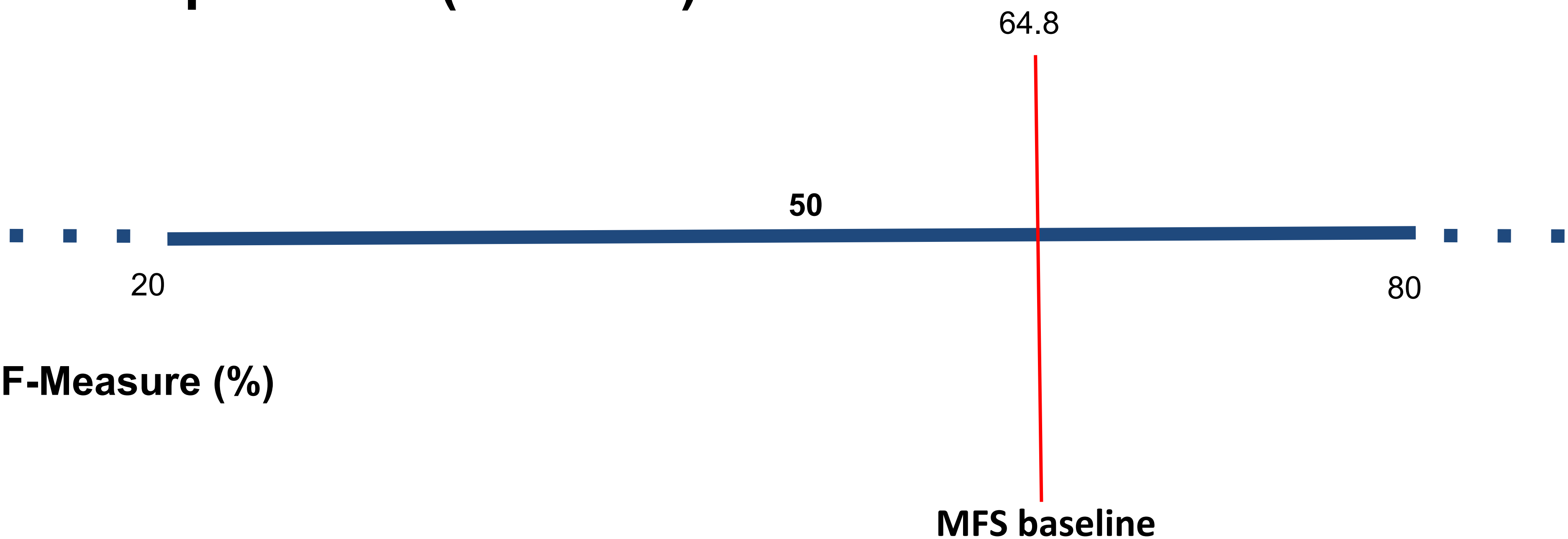


Three steps:

- First, a **bidirectional LSTM** is trained on an unlabeled corpus.
- Then, this model is used to **learn an output (context) vector for each sense annotation** in the sense-annotated training corpus.
- Finally, the **sense annotation whose context vector is closer to the target word's context vector** is selected as the intended sense.

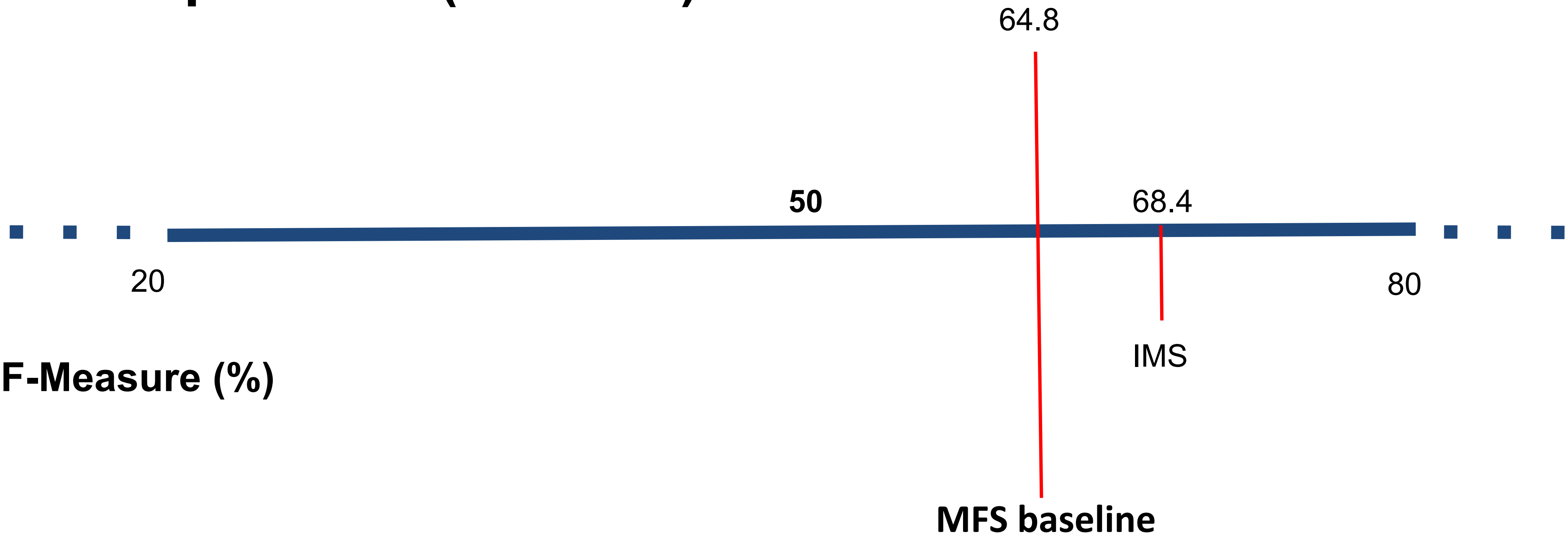
Evaluation: Results on the concatenation of all datasets

Supervised (SemCor)



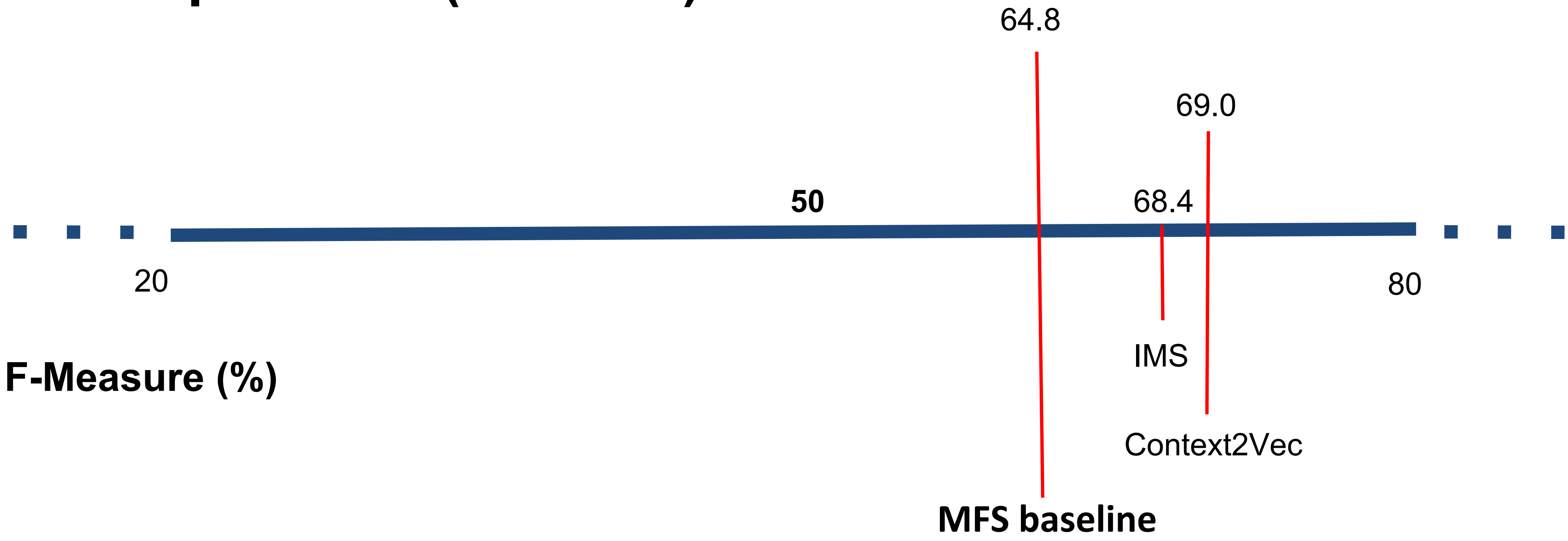
Evaluation: Results on the concatenation of all datasets

Supervised (SemCor)



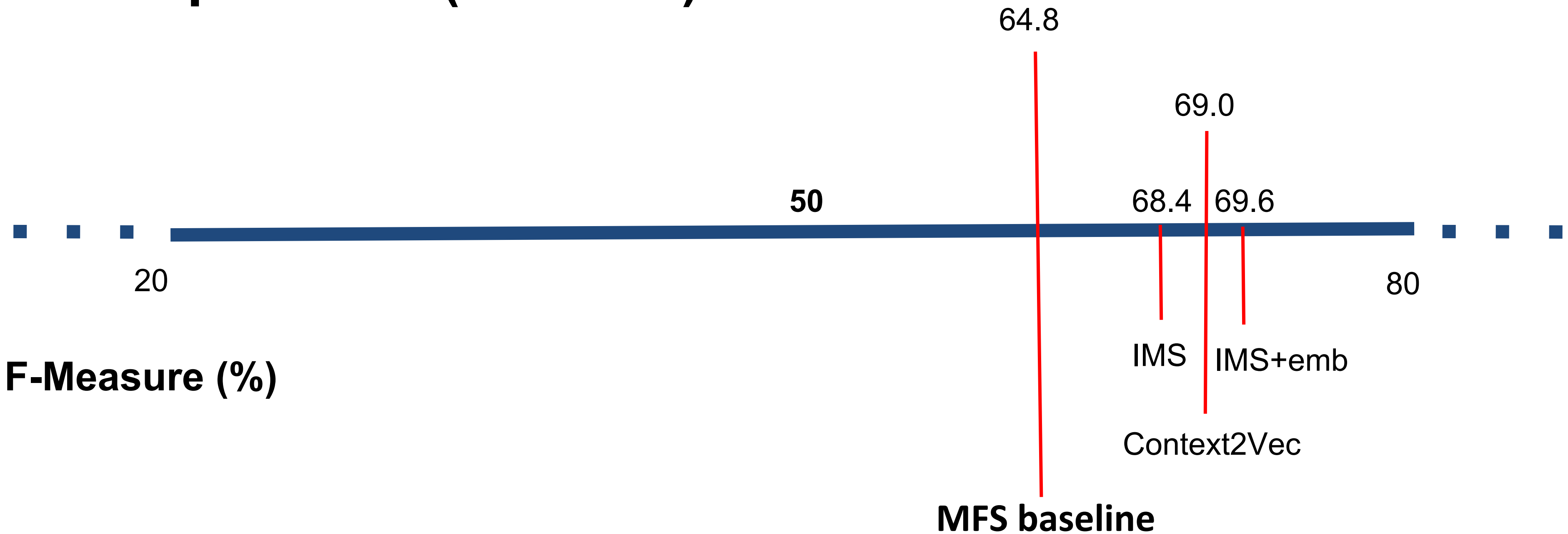
Evaluation: Results on the concatenation of all datasets

Supervised (SemCor)



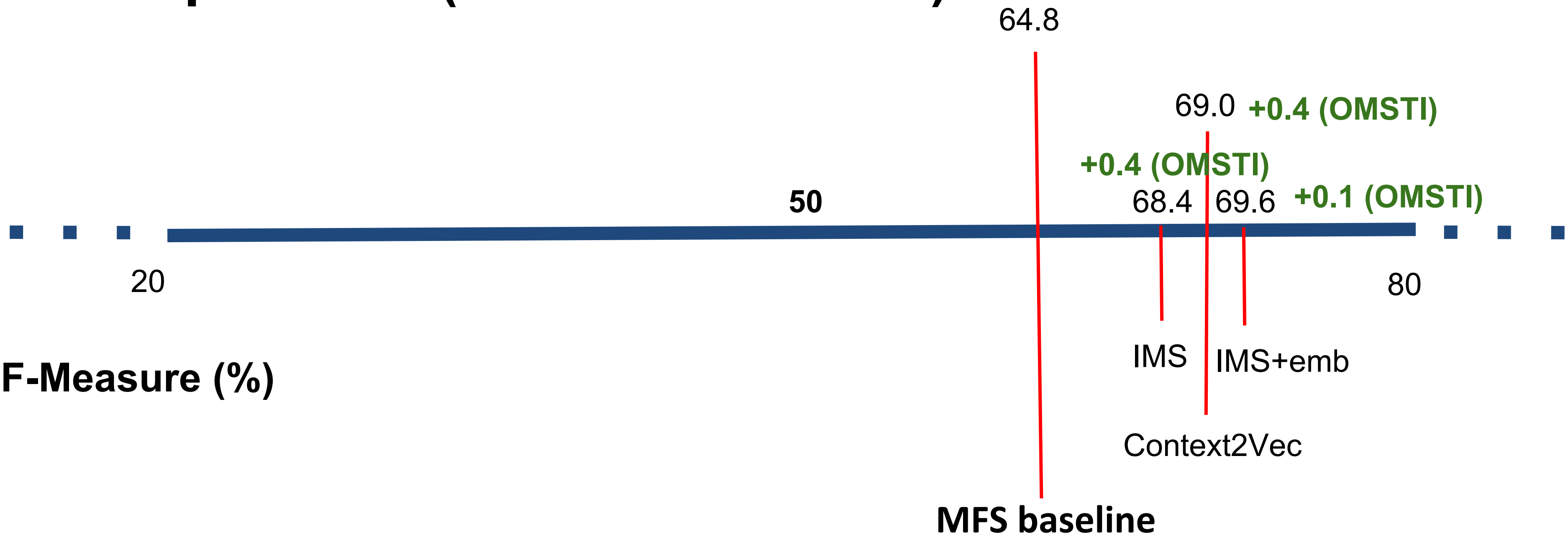
Evaluation: Results on the concatenation of all datasets

Supervised (SemCor)



Evaluation: Results on the concatenation of all datasets

Supervised (SemCor + OMSTI)



Training corpus

The automatically-constructed OMSTI **helps to improve the results of the supervised systems** trained on SemCor only.

Research direction -> (semi)automatic construction of sense-annotated datasets in order to overcome the **knowledge-acquisition bottleneck**.

Knowledge-based vs. Supervised

Supervised systems clearly outperform knowledge-based systems.

Supervised systems seem to better capture **local contexts**:

In sum, at both the federal and **state** government levels at least part of the seemingly irrational behavior voters display in the voting booth may have an exceedingly rational explanation.

Knowledge-based systems

Competitive for nouns, but underperform in other PoS tags.

The **Most Common Sense (MCS) baseline is still hard to beat**.

Only Babelfy and UKB* manage to outperform this baseline but…

- Babelfy uses the MCS baseline as a back-off strategy.
- The configuration of UKB which outperforms the baseline integrates all the sense distribution from SemCor.

Bias towards the Most Frequent Sense (MFS)

All IMS-based systems answer **over 75% of the times with the MFS**. Context2Vec is slightly less affected (73.1% on average).

The MFS bias is also present in graph-based systems, confirming the findings of previous studies: Calvo and Gelbukh (2015), Postma et al. (2016).

Evaluation: Analysis

Low overall performance on verbs

All systems **below 58%**.

Verbs are extremely fine-grained in WordNet: **10.4 number of senses per verb** on average on all datasets (4.8 in nouns and lower in adjectives and adverbs).

For example, the verb ***keep*** has 22 meaning in WordNet, 6 of them denoting *possession*.

Conclusion

We presented a **unified evaluation framework for all-words Word Sense Disambiguation**, including standardized training and testing data.

This **eases the task of researchers** to evaluate their systems and ensures a **fair comparison**.

Conclusion

We presented a **unified evaluation framework for all-words Word Sense Disambiguation**, including standardized training and testing data.

This **eases the task of researchers** to evaluate their systems and ensures a **fair comparison**.

Two potential research directions based on semisupervised learning:

- Exploiting large amounts of unlabeled corpora for learning accurate word embeddings or training **neural language models**

- **(Semi)Automatic construction of high-quality sense-annotated corpora**

Conclusion

We presented a **unified evaluation framework for all-words Word Sense Disambiguation**, including standardized training and testing data.

This **eases the task of researchers** to evaluate their systems and ensures a **fair comparison**.

Two potential research directions based on semisupervised learning:

- Exploiting large amounts of unlabeled corpora for learning accurate word embeddings or training **neural language models**
- **(Semi)Automatic construction of high-quality sense-annotated corpora**

http://lcl.uniroma1.it/wsdeval

Thank you!

All the data available at

http://lcl.uniroma1.it/wsdeval