How to enable Natural Language Understanding in many languages

Roberto Navigli

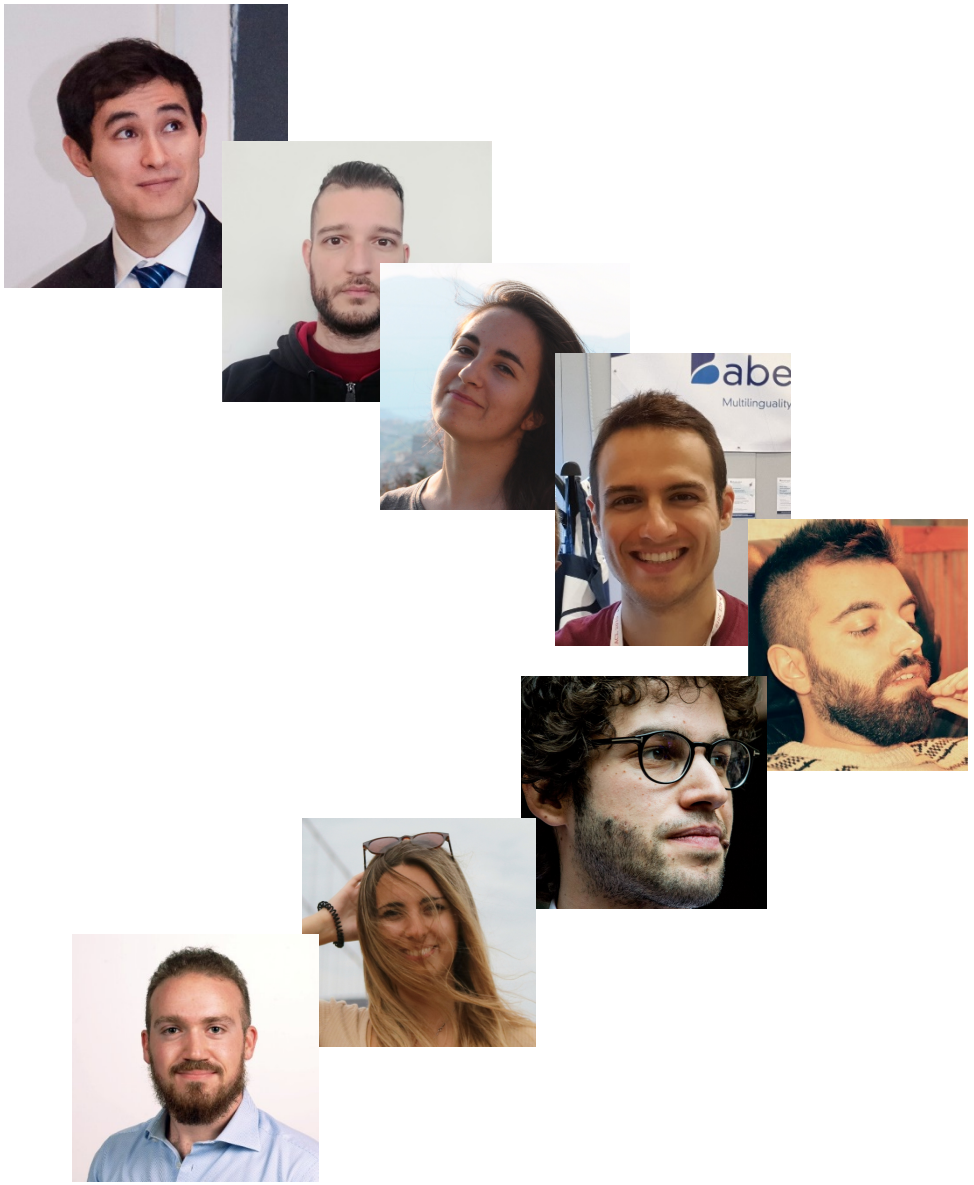
Dipartimento di Informatica

Sapienza Università di Roma

Babelscape

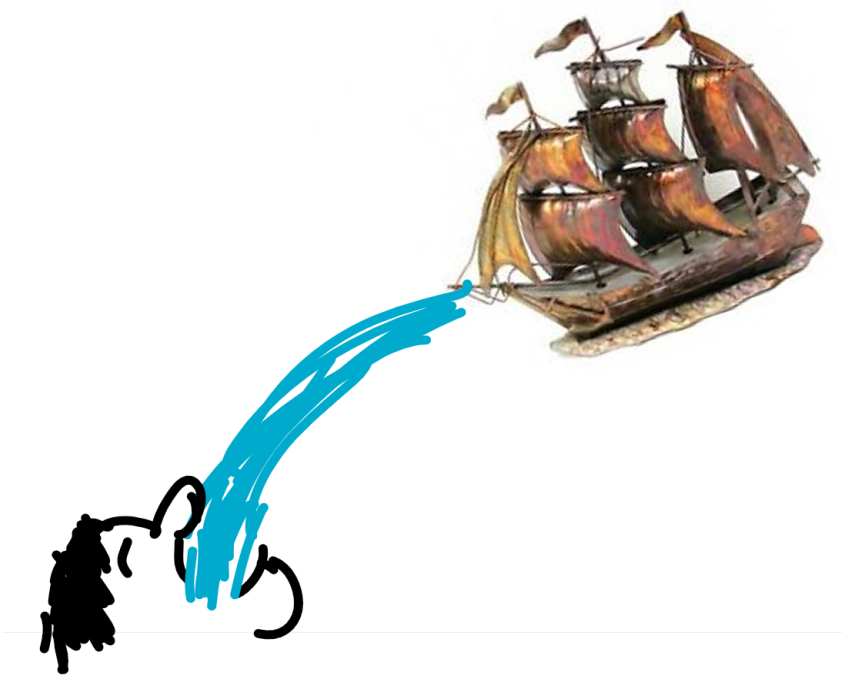
Joint work with…

- Simone Conia
- Andrea Di Fabio
- Caterina Lacerra
- Federico Martelli
- Marco Maru
- Tommaso Pasini
- Bianca Scarlini
- Federico Scozzafava



Machine Translation «does not understand»

- **EN** Is it healthy to drink from a copper vessel?
- **IT** È salutare bere da una nave di **rame**?
- **EN** Is it healthy to drink from a **copper** ship?



3 tasks to enable Natural Language Understanding at the semantic level

- Word Sense Disambiguation
  - Associating meanings with words occurring in context

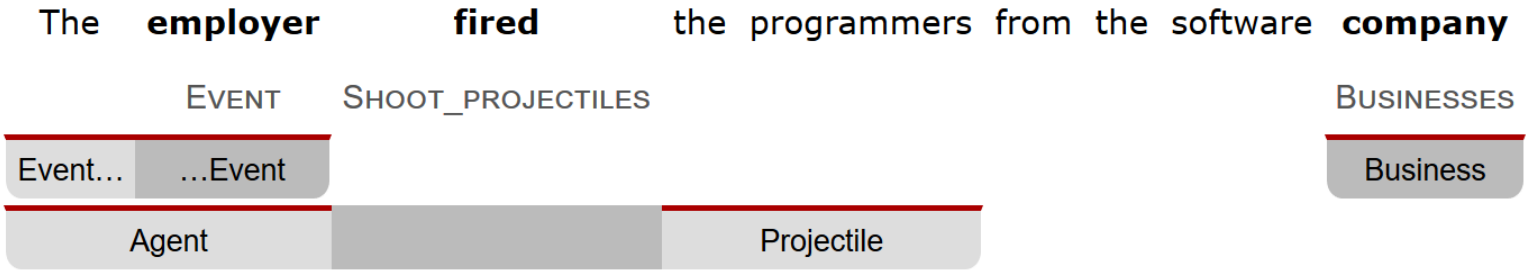
The employer fired the programmers from the software company .





3 tasks to enable Natural Language Understanding at the semantic level

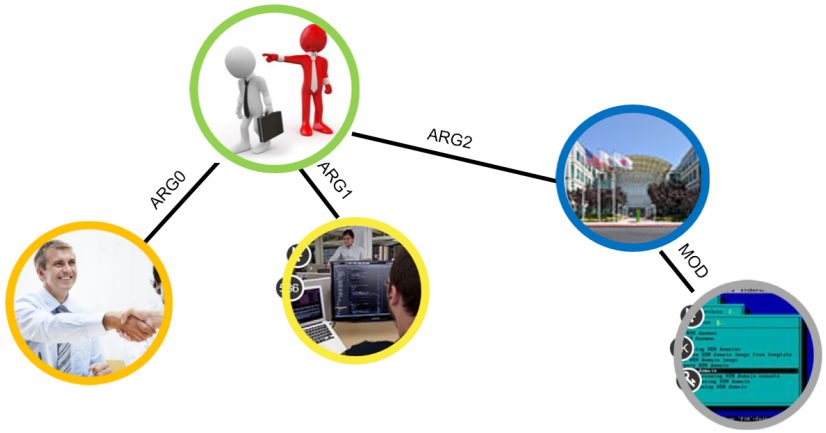
- Word Sense Disambiguation
- Semantic Role Labeling
  - «Shallow semantic parsing» which performs predicate-argument annotations



3 tasks to enable Natural Language Understanding at the semantic level

- Word Sense Disambiguation
- Semantic Role Labeling
- Semantic Parsing
  - Transforming the text into a structured semantic representation

The employer fired the programmers from the software company .



A Key Goal of AI – Multilingual Machine Reading

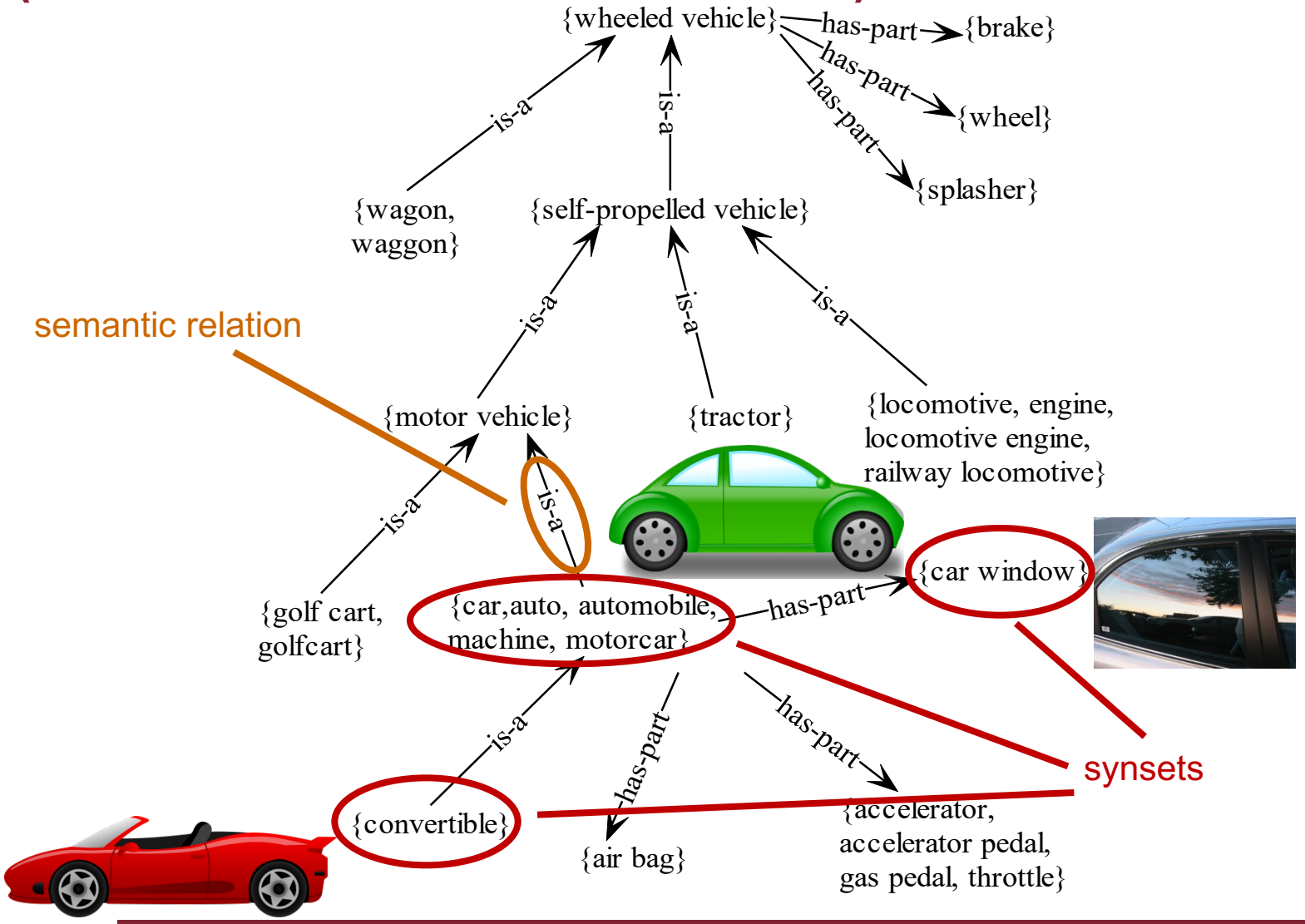
- Why? Machines could potentially **"read" the entire Web**
- Answer all kinds of questions, summarize, translate, etc.

Key question: deriving meaning from natural language by **overcoming its inherent complexities**

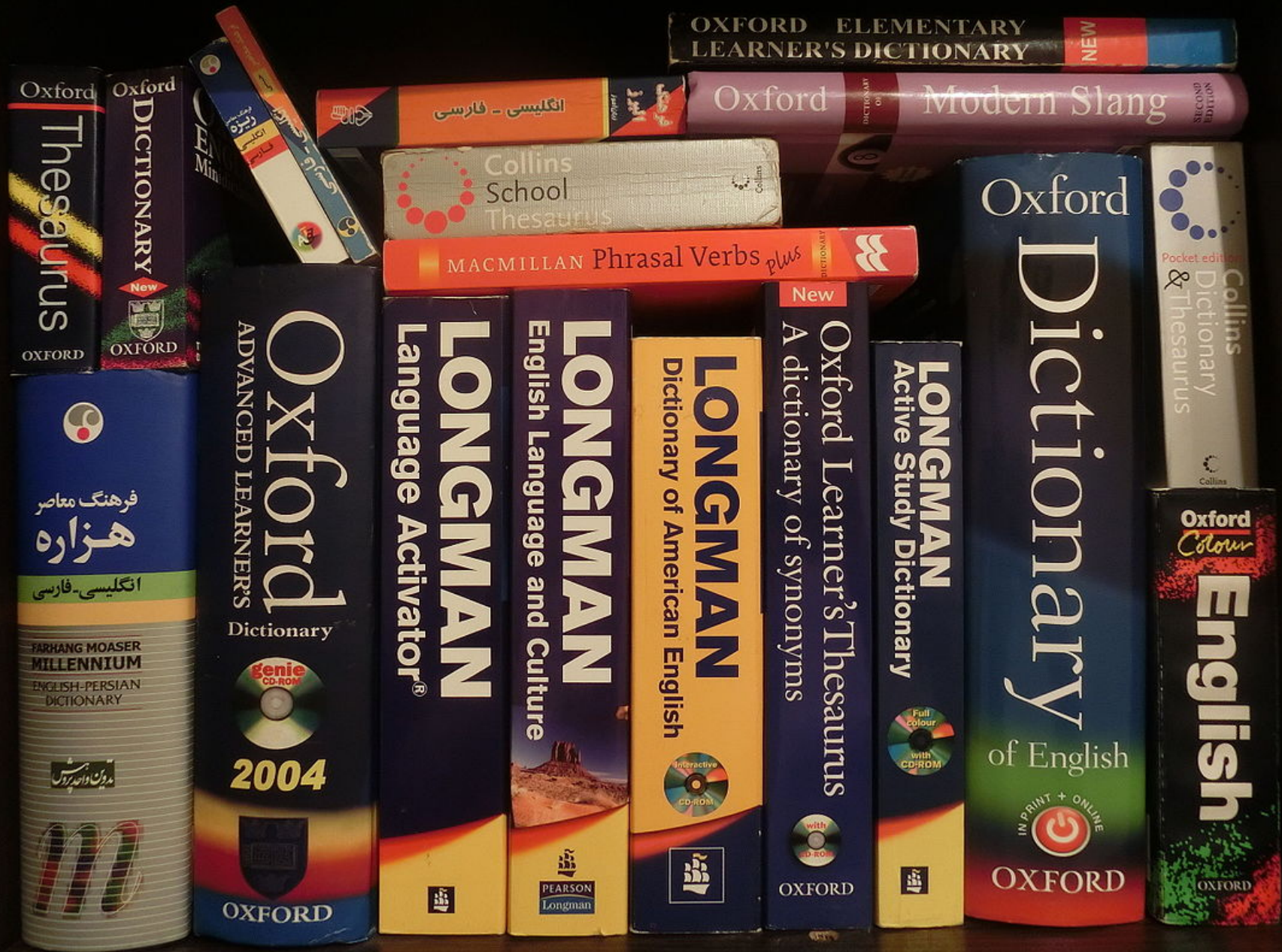
Word Sense Disambiguation

The mouse is eating the cheese .

WordNet is THE sense inventory for English (Miller et al. 1990; Fellbaum, 1998)



The resource diaspora

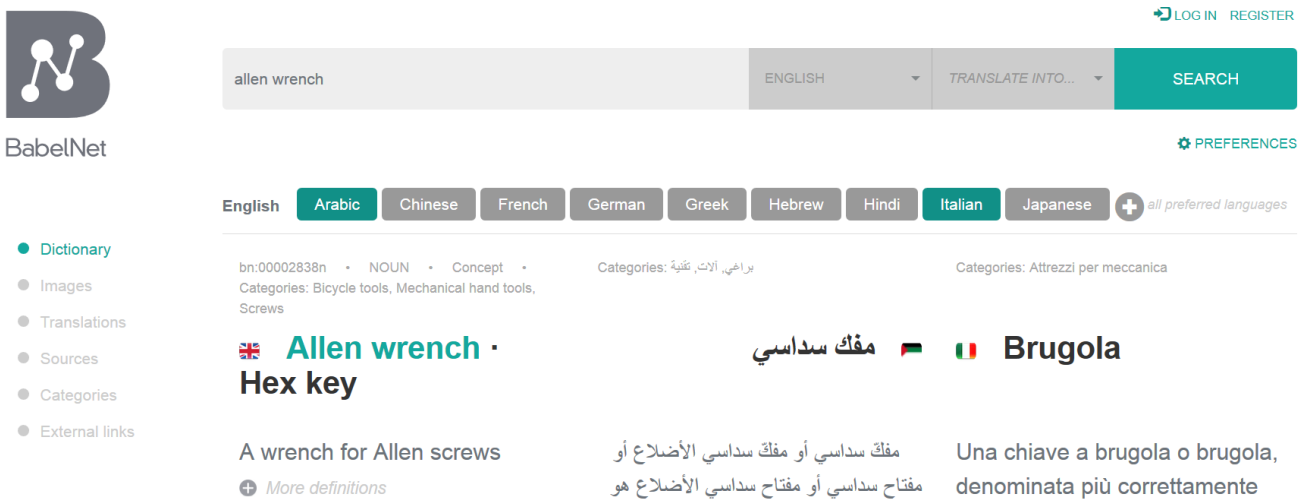


Word senses: which sense inventory to scale multilingually?

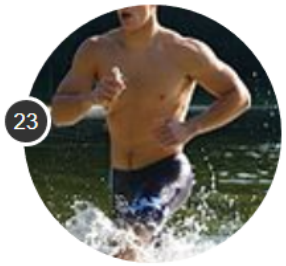
- **BabelNet** (Navigli and Ponzetto, ACL 2010; AIJ 2012): a merger of WordNet, multilingual wordnets, Wikipedia, Wikidata, Wiktionary and many other resources

Word senses: which sense inventory to scale multilingually?

- **BabelNet** (Navigli and Ponzetto, ACL 2010; AIJ 2012): a merger of WordNet, multilingual wordnets, Wikipedia, Wikidata, Wiktionary and many other resources



Verb



run

Move fast by using one's feet, with one foot off the ground at any given time

ID: 00093170v | Concept

هَرْوَلَ, جَرَى, رَكَضَ

奔跑, 跑

courir

rennen

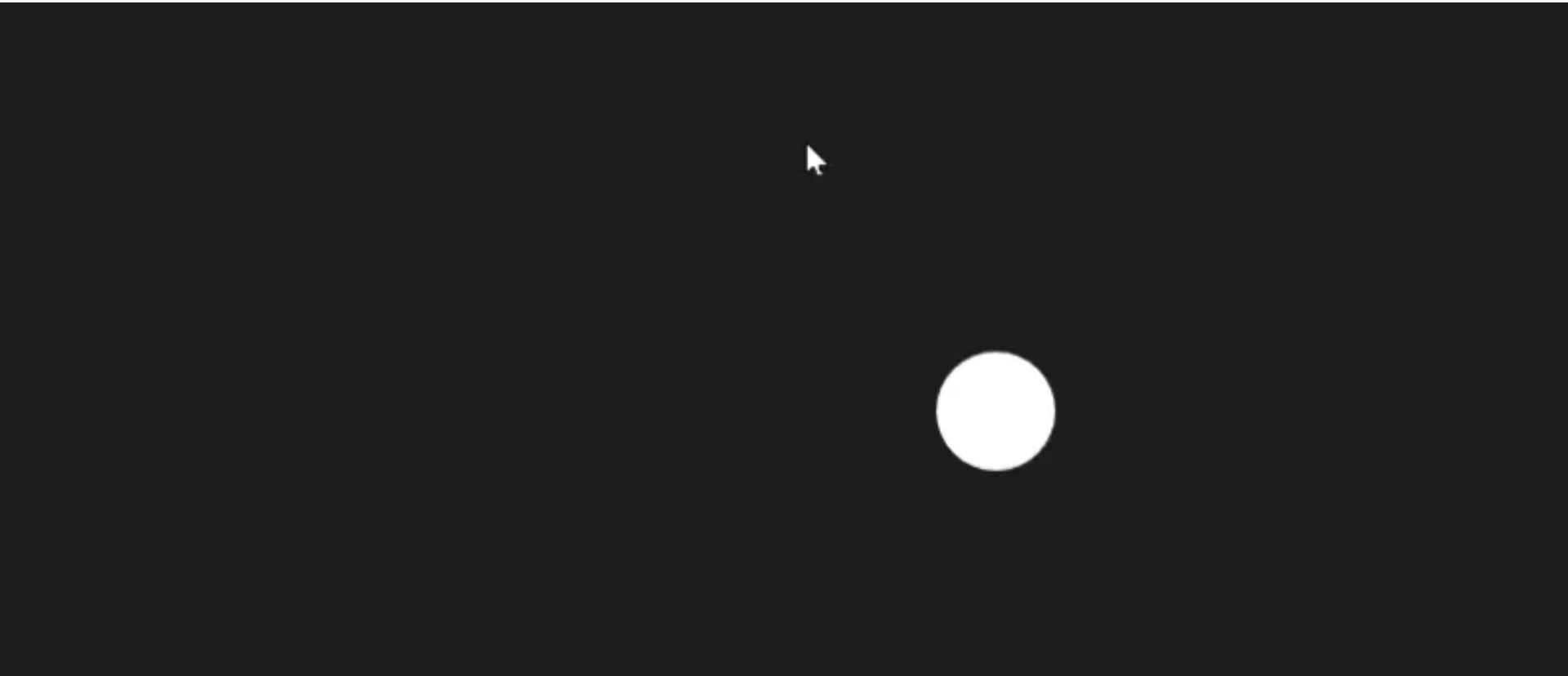
τρέχω, κινούμαι

רץ

correre

Why do we need BabelNet?

- Multilinguality: the same concept is expressed in tens of languages
- Coverage: 284 languages and 16 million entries!
- Concepts and named entities together: dictionary and encyclopedic knowledge is semantically interconnected



Two ERC projects aimed at automatic text understanding

MOUSSE: ERC Consolidator Grant (2017-2022)

The current limits of Word Sense Disambiguation

- Lack of high-quality knowledge
  - Limited to WordNet, BabelNet, etc.
- Limited to one language
  - English, English, English?
- Fine granularity of the sense inventory
  - WordNet, WordNet, WordNet

The current limits of Word Sense Disambiguation

- Lack of high-quality knowledge (Maru et al., EMNLP 2019)
  - Limited to WordNet, BabelNet, etc.
- Limited to one language
  - English, English, English?
- Fine granularity of the sense inventory
  - WordNet, WordNet, WordNet

What is missing in WordNet and BabelNet?

- WordNet is manually curated, but contains mostly paradigmatic relations (e.g. convertible –is-a-> car)
- BabelNet is very rich, but contains a huge number of unclassified relations (e.g. ERC –> Brussels, screen –> pixel)

- To disambiguate, we need syntagmatic relations
  - open -> door
  - open -> business
  - play -> piano
  - play -> act
  - play -> game
  - door -> window

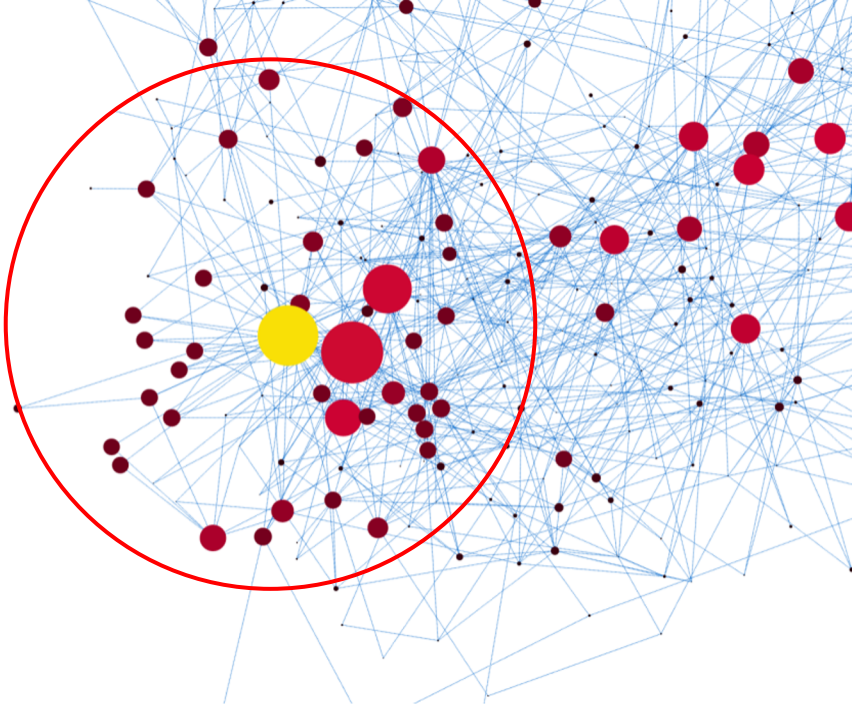
SyntagNet: addressing the lack of syntagmatic knowledge (Maru et al., EMNLP 2019)

- We extracted cooccurrences of nouns and verbs on a large scale from Wikipedia and the British National Corpus

$$score(w_1, w_2) = log_2(1 + n_{w_1 w_2}) \frac{2n_{w_1 w_2}}{n_{w_1} + n_{w_2}}$$

- We disambiguated the cooccurrences with our state-of-the-art WSD system
  - Babelfy (Moro et al., TACL 2014)
- We validated the top-ranking cooccurrences manually
- Minimum agreement between 3 annotators: **κ = 0.71** (substantial agreement)
  - Disagreement instances are valid alternative tags, not factual errors

Knowledge-based WSD



- Personalized PageRank can be used to perform disambiguation of an ambiguous word in context
- The probability of a word sense depends on its direct and indirect interconnections to other senses of words in context
- The quality of disambiguation depends on the **underlying semantic network** which connects meanings

Comparison between different Lexical Knowledge Bases

		English					
resource	**#relations**	**Sens2**	**Sens3**	**Sem07**	**Sem13**	**Sem15**	**All**
WNG (WordNet+PWNG)	671,779	69.2	65.9	54.9	66.8	70.7	67.1
WNG+KnowNet20	520,682	67.2	65.8	53.8	67.3	71.5	66.6
WNG+deepKnowNet95d	522,880	66.9	64.9	53.6	66.9	71.6	66.2
WNG+BabelNet 4.0	9,447,341	67.5	64.1	53.0	67.6	66.9	65.6
WNG+eXtended WordNet	551,551	67.7	65.7	52.3	67.6	71.0	66.7
WNG+ColWordNet	8,424	69.2	65.9	54.1	66.7	70.7	67.1
WNG+SyntagNet	88,019	**71.2**	**71.6**	**59.6**	**72.4**	**75.6**	**71.5**

Underlined results show statistically significant differences from the baseline (χ^2 test, $p < 0.05$)

Comparison with the state of the art in WSD

system	Sens2	Sens3	Sem07	Sem13	Sem15	All
LSTMLP•	**73.8**	71.8	**63.5**	69.5	72.6	**71.5**
IMSC2V$_{+PR}$∞	**73.8**	71.9	63.3	68.2	72.8	71.2
fastSense△	73.5	**73.5**	62.4	66.2	73.2	71.1
UKB+SyntagNet	71.2	71.6	59.6	**72.4**	**75.6**	**71.5**

• Yuan et al., 2016

† Raganato et al. (2017b)

⋆ Gutiérrez Vázquez et al. (2010)

◇ best SUDOKU-RUNk (Manion, 2015)

○ Pasini and Navigli (2017)

∞ Melacci et al. (2018)

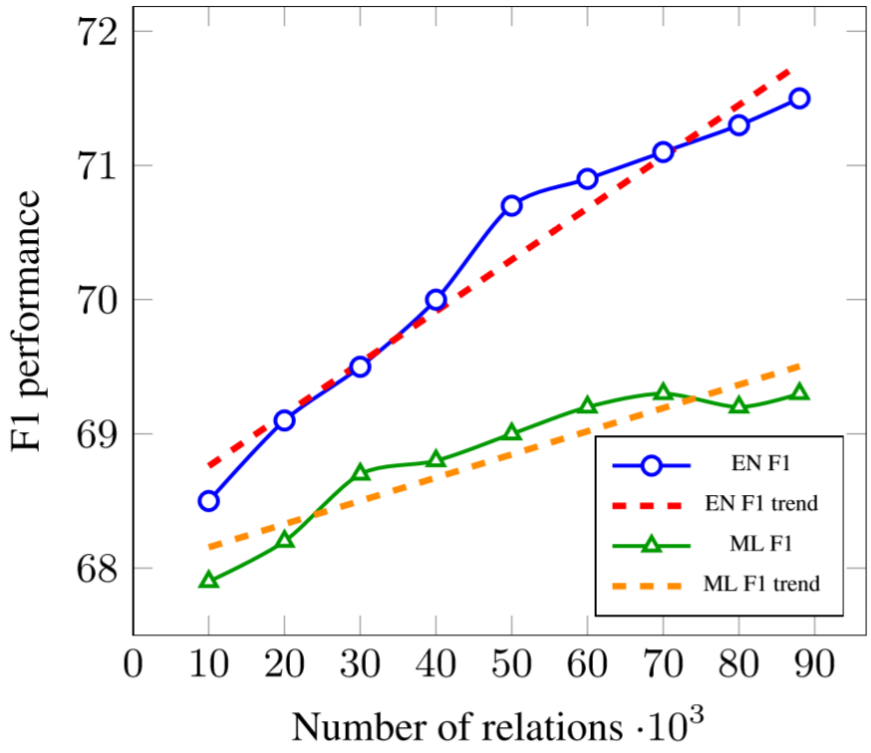
△ Uslu et al. (2018)

‡ result obtained by aggregating the outputs of the best systems across languages

Underlined results show statistically significant differe[illegible] from results (χ^2 test, $p < 0.05$)

Syntagmatic relations rival (and surpass) neural supervised WSD!

We are far from reaching a plateau!



- More relation edges are expected to provide further performance increase
- English growth steeper because of cleaner sense inventory

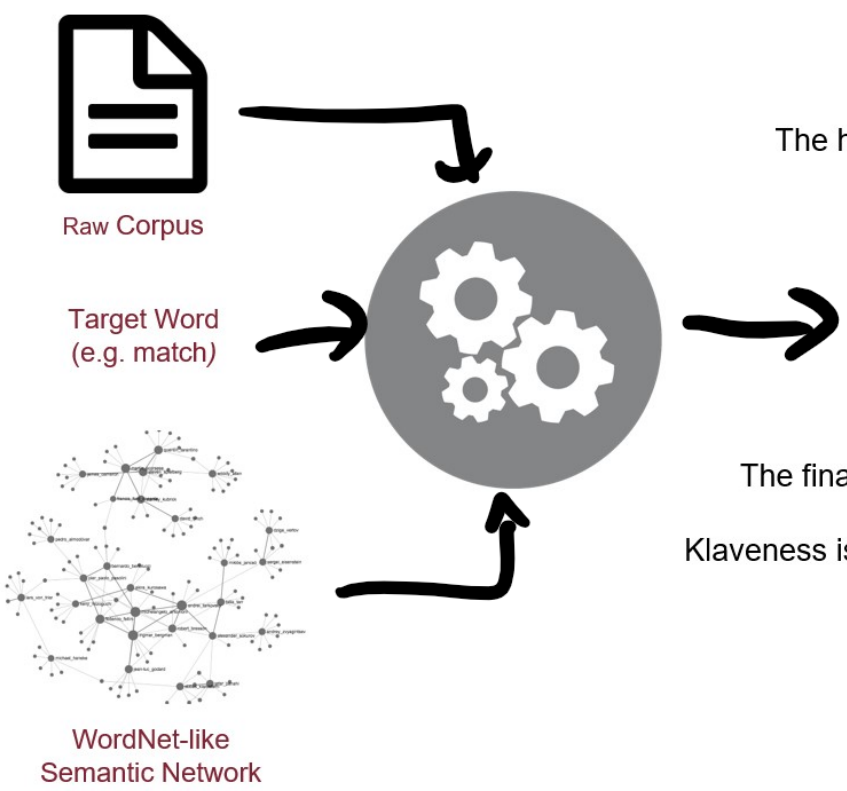
The current limits of Word Sense Disambiguation

- Lack of high-quality knowledge
  - Limited to WordNet, BabelNet, etc.
- Limited to one language (Scarlini et al., AAAI 2020)
  - English, English, English
- Fine granularity of the sense inventory
  - WordNet, WordNet, WordNet


“Language barriers”

Addressing the lack of sense-annotated data

- Goal: overcome the **knowledge acquisition bottleneck**
  - i.e. acquire **large training datasets for supervised WSD**



The head of the **match** consisted of a mixture of potassium

Wooden **matches** are packaged in matchboxes

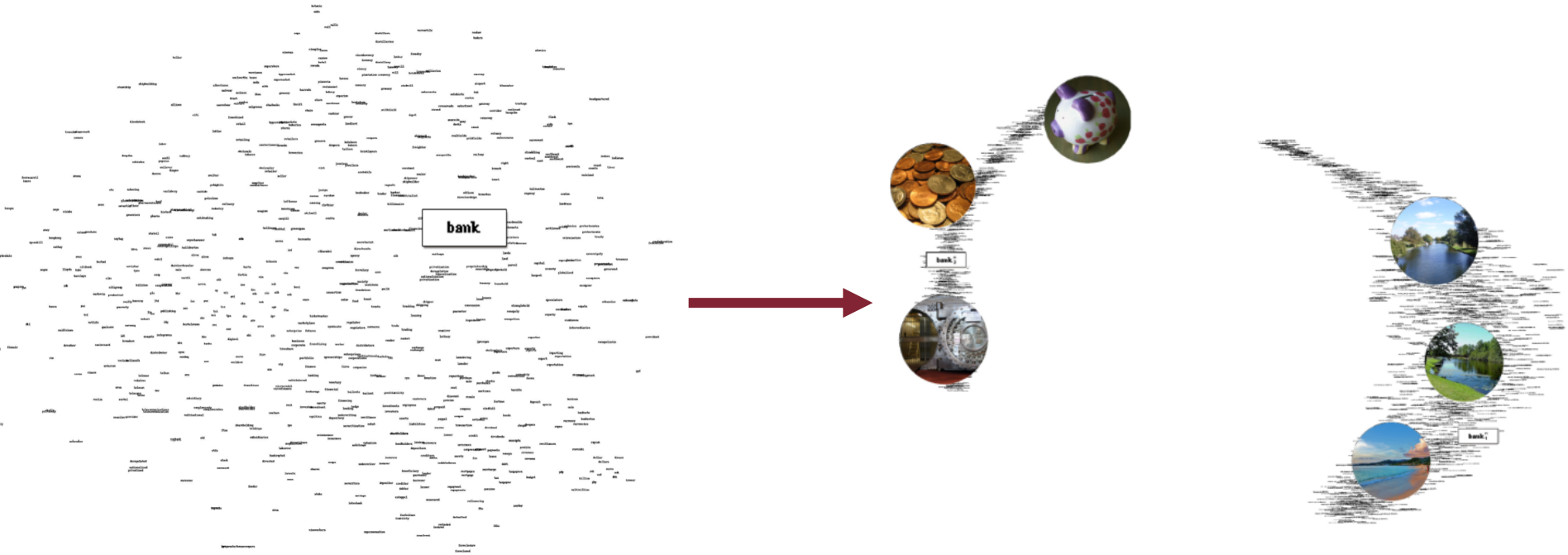
The **match** takes place in a ring

The final tournament **matches** were held in four stadium venues

Klaveness is a Norwegian lawyer and footballer who played 73 **matches**

Training Corpus

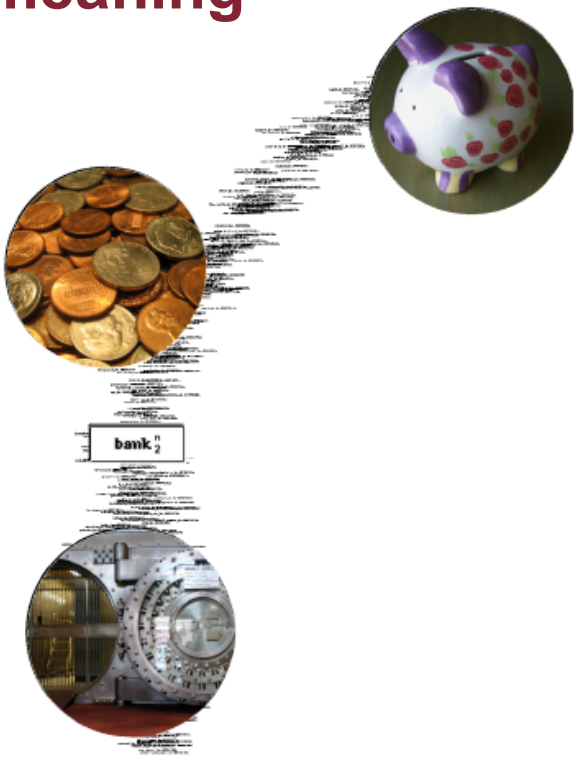
Solution: move from lexical to semantic

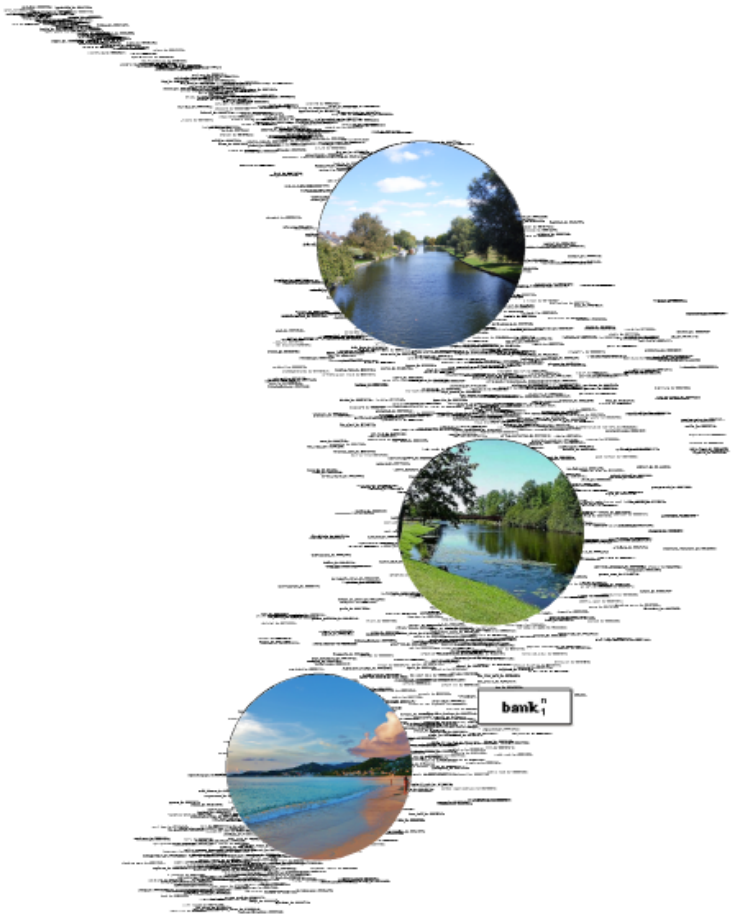


Word vector space model

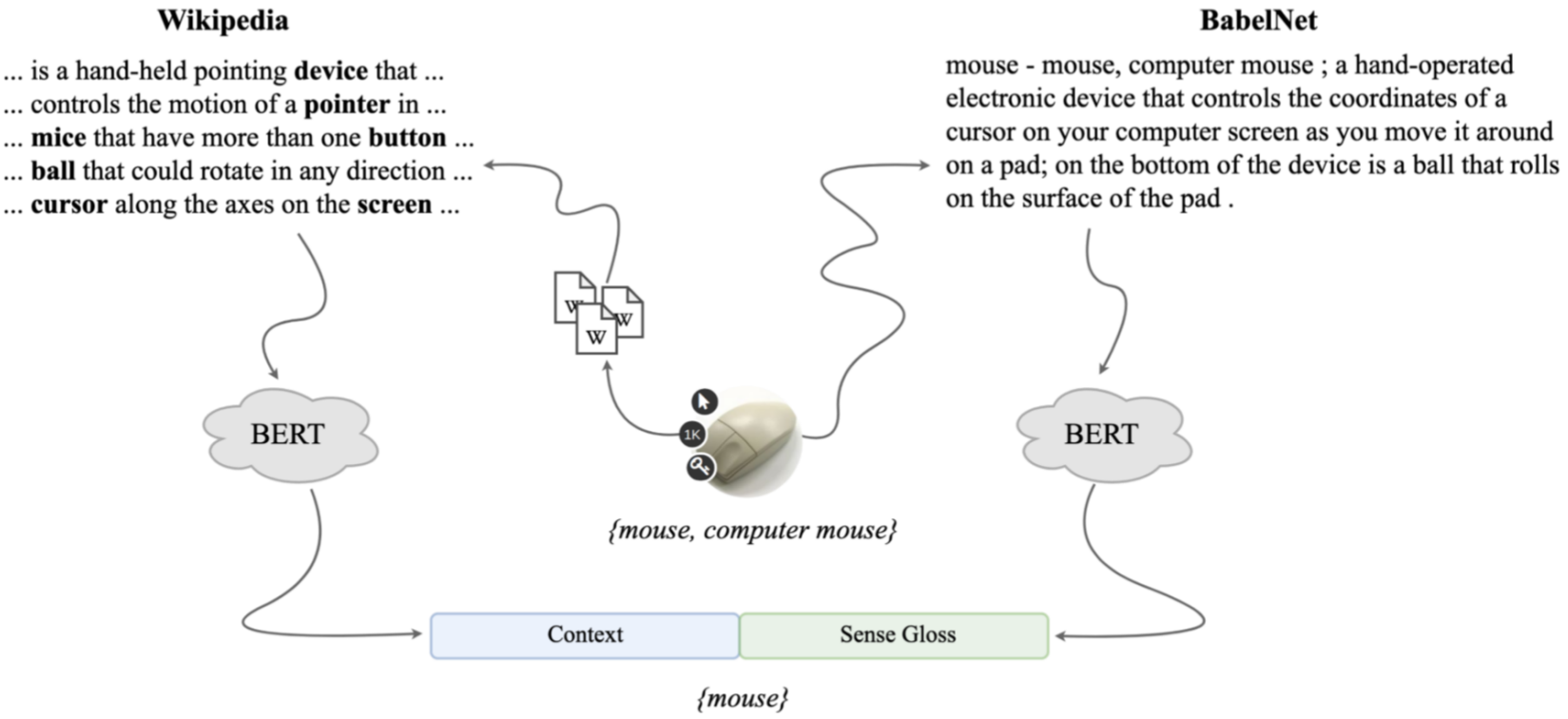
Sense vector space model

Solution: distinct representation for each word's meaning





Sense representations out of contextualized embeddings (AAAI 2020, SensEmBERT)

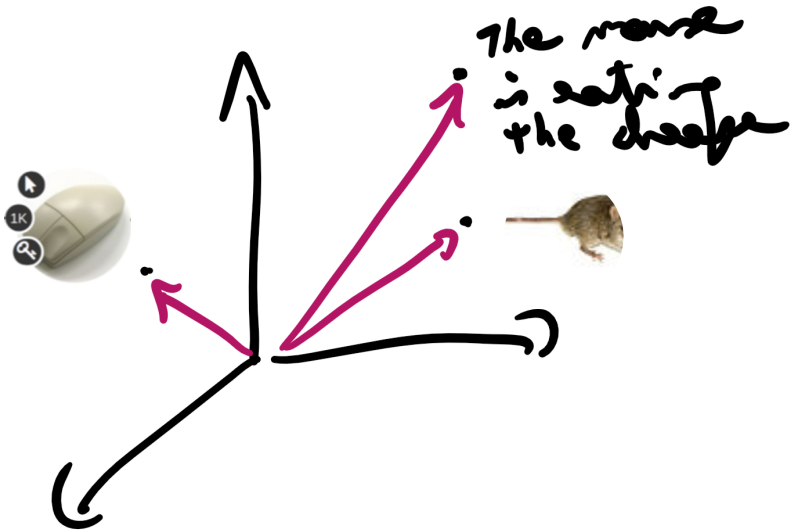


$$v_s = \frac{\sum_{w_i \in W_s} rank(w_i)^{-1} \; v_{w_i}}{\sum_{w_i \in W_s} rank(w_i)^{-1}}$$

How to perform Word Sense Disambiguation?

- Very, very simple: 1-nearest neighbour!

- For a given sentence, e.g. *The **mouse** is eating the cheese*, compare the representations of the meanings of **mouse** with the context vector representation:



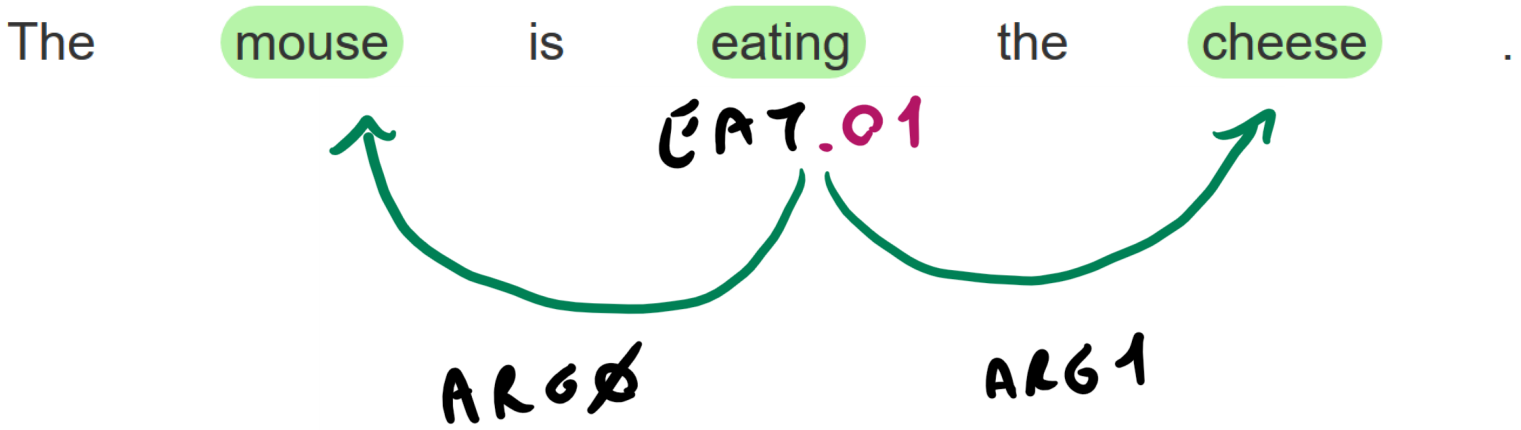
Knowledge-based+Neural achieves 80% accuracy in Word Sense Disambiguation!

	Model	Senseval-2	Senseval-3	SemEval-07	SemEval-13	SemEval-15	ALL
KB	MFS	72.1	72.0	65.4	63.0	66.3	67.6
	Lesk$_{ext}$+emb (2014)	74.6	72.7	66.0	66.2	67.8	69.8
	UKB$_{gloss}$ (2014)	70.6	58.4	56.6	59.0	62.3	62.1
	Babelfy (2014)	74.0	66.7	61.0	66.4	69.9	68.6
Sup	IMS+emb (2016)	79.0	74.6	71.1	65.9	72.1	71.9
	Bi-LSTM (2017)	78.6	72.7	71.1	66.4	73.3	71.6
	HCAN (2018a)	78.3	73.2	70.9	68.5	73.8	72.6
	EWISE$_{ConvE}$ (2019)	-	-	-	69.4	-	74.0
Sup$_{context}$	context2vec (2016)	78.0	73.1	66.7	65.6	71.6	71.0
	LSTM-LP (2016)	79.6	76.3	71.7	69.5	72.8	-
	BERT k-NN (2019)	71.7	73.0	72.9	65.6	68.4	69.3
	BERT k-NN + MFS (2019)	81.4	76.3	73.6	71.8	74.0	75.5
	LMMS (2019)	82.6	77.8	76.7	75.4	76.6	77.9
Ours	SENSEMBERT	79.9	70.0	74.2	75.0	**79.7**	75.7
	SENSEMBERT$_{sup}$	**84.0**	**80.8**	**80.5**	**78.5**	79.5	**80.5**

The current limits of Word Sense Disambiguation

- Lack of high-quality knowledge
  - Limited to WordNet, BabelNet, etc.
- Lack of sense-annotated data
  - Limited to SemCor, MASC, Senseval, SemEval, not much more
- Fine granularity of the sense inventory (Lacerra et al., AAAI 2020)
  - WordNet, WordNet, WordNet
  - **No time in this overview talk, sorry!**

Dependency-based Semantic Role Labeling



Roleset id: eat.01

Roles:

Arg0-PAG: *consumer, eater* (vnrole: 39.1-1-agent)
Arg1-PPT: *meal* (vnrole: 39.1-1-patient)

What are the issues with the current frame inventories?

	Cluster type	#	Argument roles	#	Meaning units	#
FrameNet	Frames	1,224	Frame elements	10,542	Lexical units	5,200
VerbNet	Levin's classes	329	Thematic roles	39	Senses	6,791
PropBank	Verbs	5,649	Proto-roles	6	Framesets	10,687
WordNet	-	-	-	-	Synsets	13,767

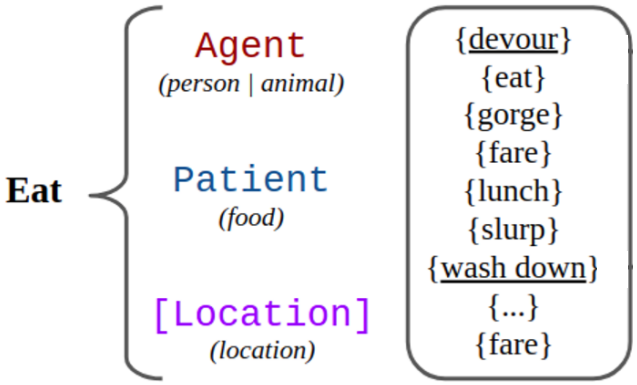
- Low coverage of verbal meanings (FrameNet, VerbNet)
- Low coverage of verbs (all resources, compared to 11.5k distinct verbs in WordNet)
- Unintelligibile argument roles (ARG0, ARG1, PropBank)
- Syntactic organization of verbs (VerbNet)
- Language (English!) specificity (all resources)

VerbAtlas: a new resource for SRL and Semantic Parsing (Di Fabio et al., EMNLP 2019)

- A **manually-crafted inventory of verbs** and **argument structures** with:
  1. full coverage of the verbal lexicon
  2. prototypical argument structures for each cluster of synsets that define a semantically-coherent frame
  3. explicit, cross-domain semantic roles
  4. refined semantic information and selectional preferences for the frame argument structure
  5. scalability: linkage to WordNet and, as a result, to BabelNet
  6. reusability: full mapping to PropBank framesets
  7. effectiveness: ability to improve over PropBank on the CoNLL-2009 dataset

VerbAtlas vs. PropBank

VerbAtlas



PropBank

Roleset id: eat.01

Roles:

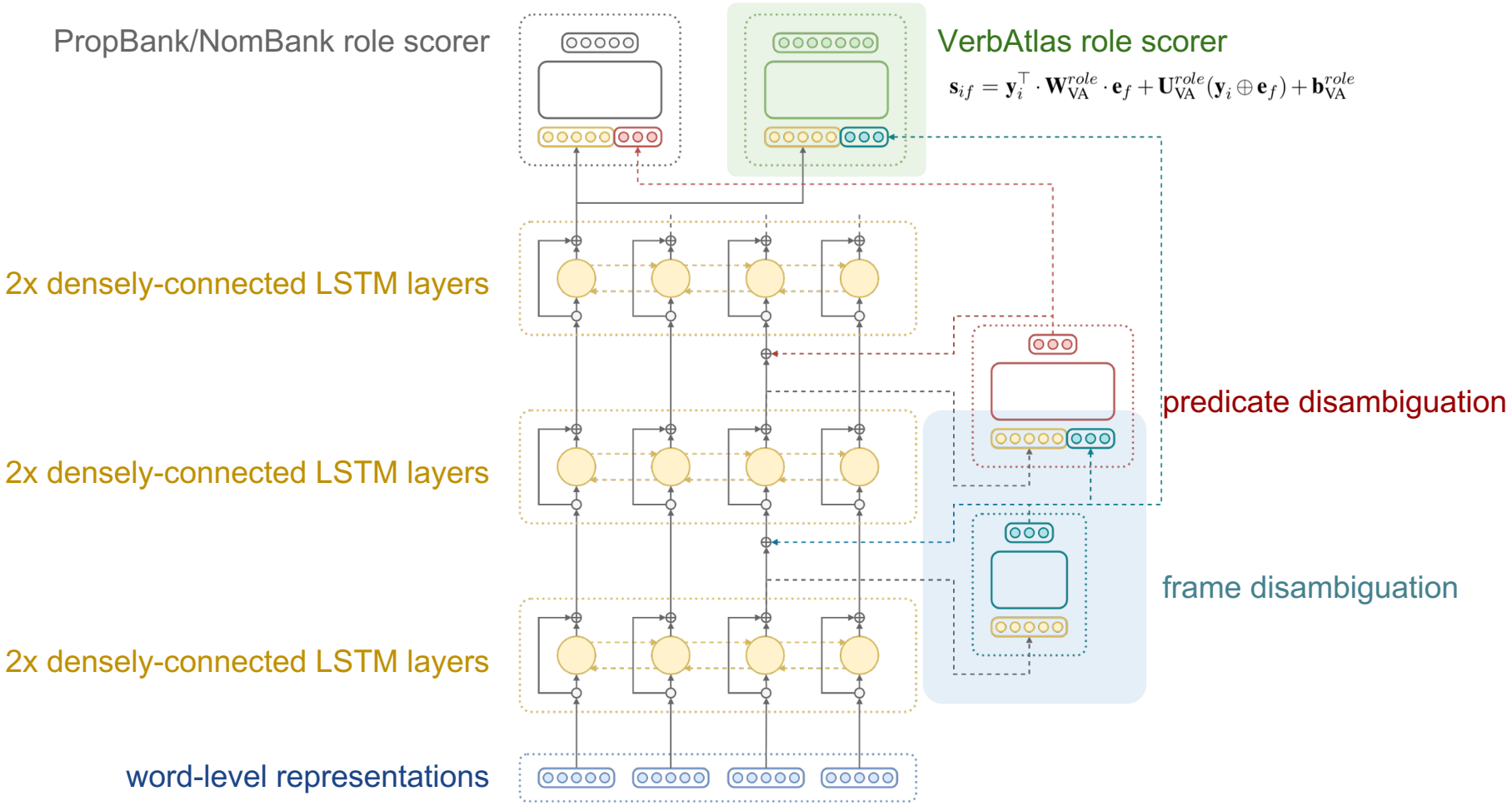
Arg0-PAG: *consumer, eater* (vnrole: 39.1-1-agent)
Arg1-PPT: *meal* (vnrole: 39.1-1-patient)

Roleset id: wash_down.04 ,

Roles:

Arg0-PAG: *Agent of swallowing*
Arg1-PPT: *Thing being swallowed*
Arg2-MNR: *Liquid enabling swallowing*

1. **A Simple SRL Architecture based on Cai et al. (2018)**
2. **Aiding predicate disambiguation with frame disambiguation**
3. **Achieving a deeper understanding thanks to finer-grained semantically-coherent roles**



Results on the CoNLL-2009 dataset

In-domain

Syntax-aware system	P	R	F_1
Roth and Lapata (2016)	88.1	85.3	86.7
Marcheggiani and Titov (2017)	89.1	86.8	88.0
He et al. (2018)	89.7	**89.3**	89.5
Li et al. (2018)	**90.3**	**89.3**	**89.8**
Syntax-agnostic system	P	R	F_1
Marcheggiani et al. (2017)	88.7	86.8	87.7
He et al. (2018)	89.5	87.9	88.7
Cai et al. (2018)	89.9	89.2	89.6
This work	**90.5**	**89.5**	**90.0**

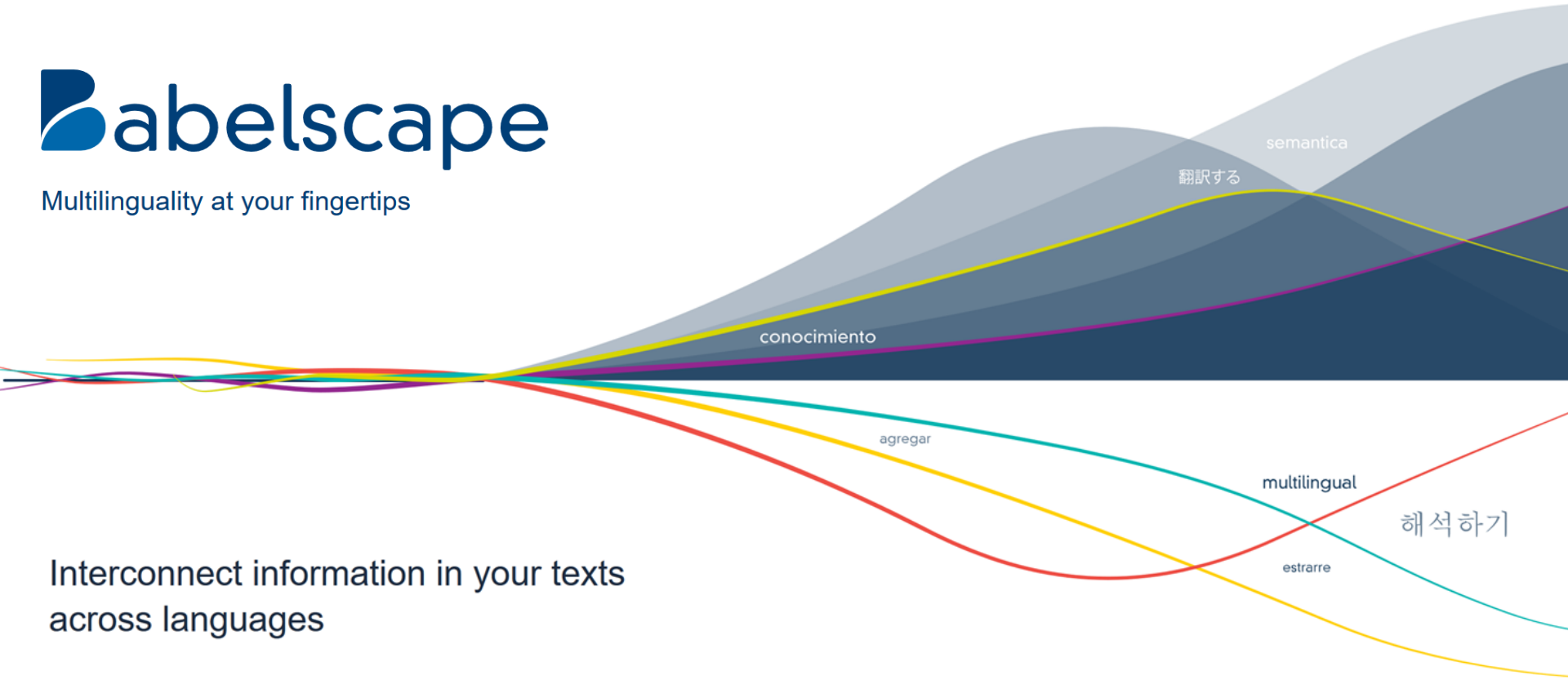
Out-of-domain

Syntax-agnostic system	P	R	F_1
Marcheggiani et al. (2017)	79.4	76.2	77.7
He et al. (2018)	**81.7**	76.1	78.8
Cai et al. (2018)	79.8	78.3	79.0
This work	81.1	**78.4**	**79.7**

- VerbAtlas adds important information already within the domain, and provides robustness across domains

Babelscape: the future of BabelNet and related technologies

- We created a Sapienza spinoff, Babelscape, with the **key objective** of making BabelNet and related technologies sustainable
- Income is reinvested in BabelNet and subsequent projects
- Working for EUIPO and big companies



Demos

- Let's now look at demos of our technologies

Conclusions

- Natural Language Understanding is hampered by the lack of knowledge, annotated data and suitable resources
  - Especially if we want to go multilingual or scale across domains
- We presented approaches to **relieve the knowledge acquisition bottleneck**
  - take the tasks of disambiguation and semantic role labeling beyond the current state of the art
  - make the systems and their outputs more understandable, explainable and scalable
- All of the approaches can **scale across languages thanks to BabelNet**

Thanks or…

ERC MOUSSE Consolidator Grant, contract no. 726487

SAPIENZA
UNIVERSITÀ DI ROMA

Roberto Navigli

Linguistic Computing Laboratory
http://lcl.uniroma1.it
@RNavigli