Multilinguality for free, or why you should care about linking vector representations to (BabelNet) synsets

Roberto Navigli

Dipartimento di Informatica

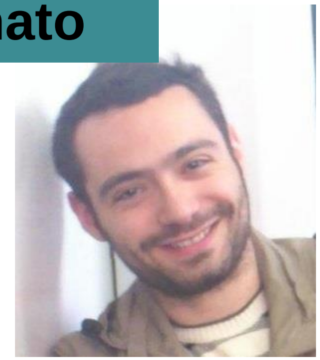
Sapienza Università di Roma

Linguistic Computing Laboratory

http://lcl.uniroma1.it

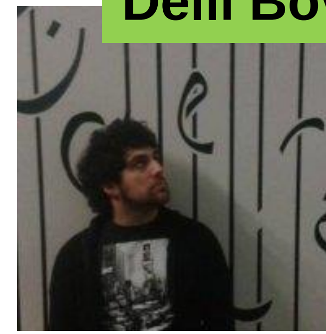
19th October 2017 – Hannover, Germany

4th Alexandria workshop

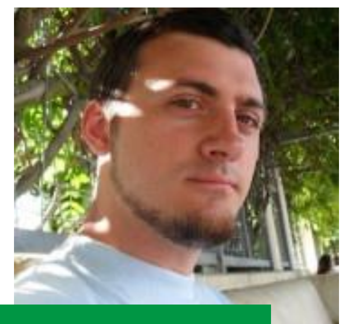


Alessandro Raganato

Claudio Delli Bovi



Taher Pilehvar



Ignacio Iacobacci

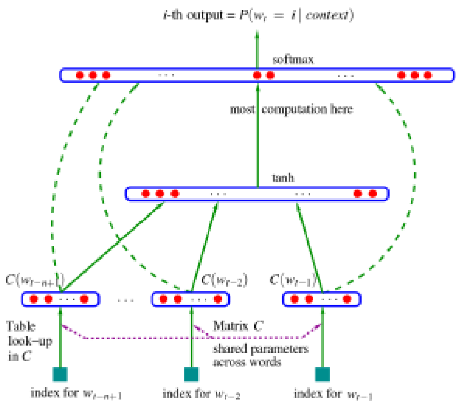
José Camacho Collados

Massimiliano Mancini

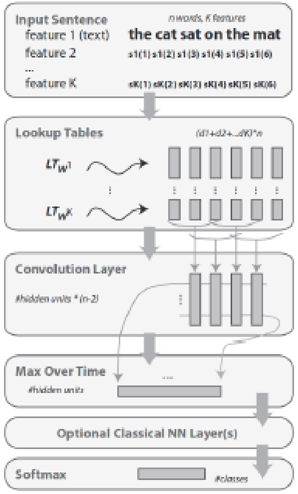
How to represent words and word senses?

- Vectors provide a representation which is easy to use, visualize and combine
  - Excellent survey (Turney and Pantel, 2010)

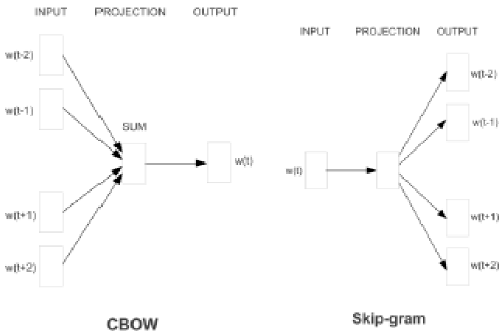
Much work on vector representations of meaning



Bengio et al. (2003)



Collobert & Weston (2008)

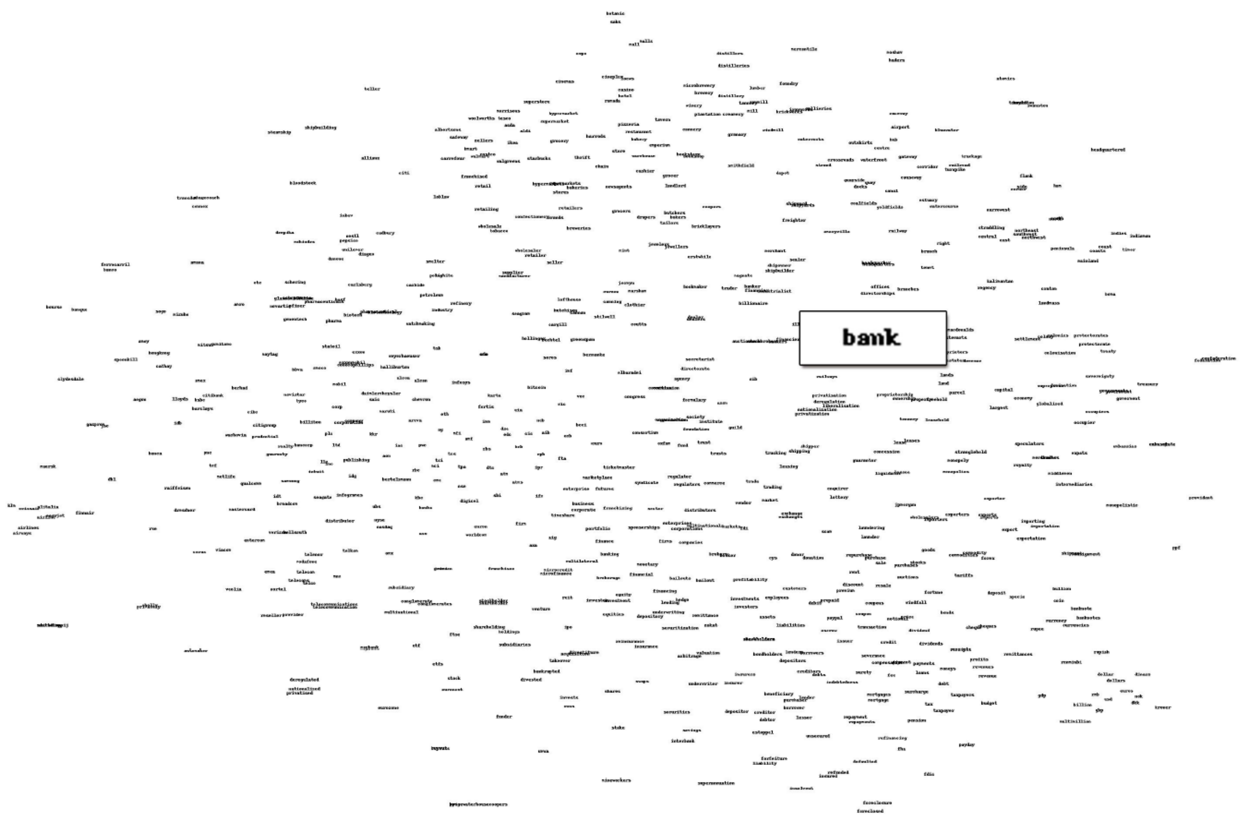


Probability and Ratio	$k = solid$	$k = gas$	$k = water$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36

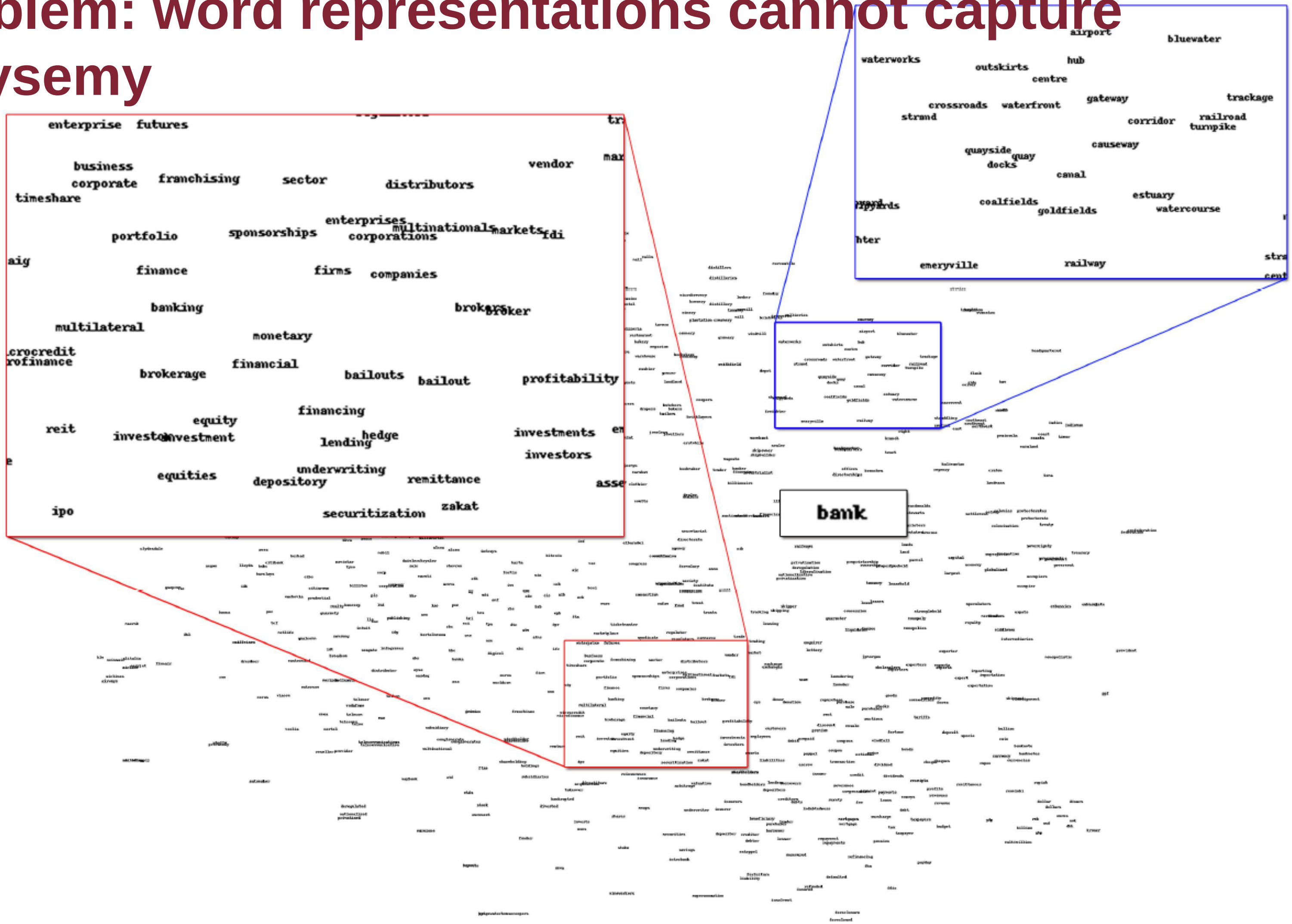
Pennington et al. (2014)

Mikolov et al. (2013)

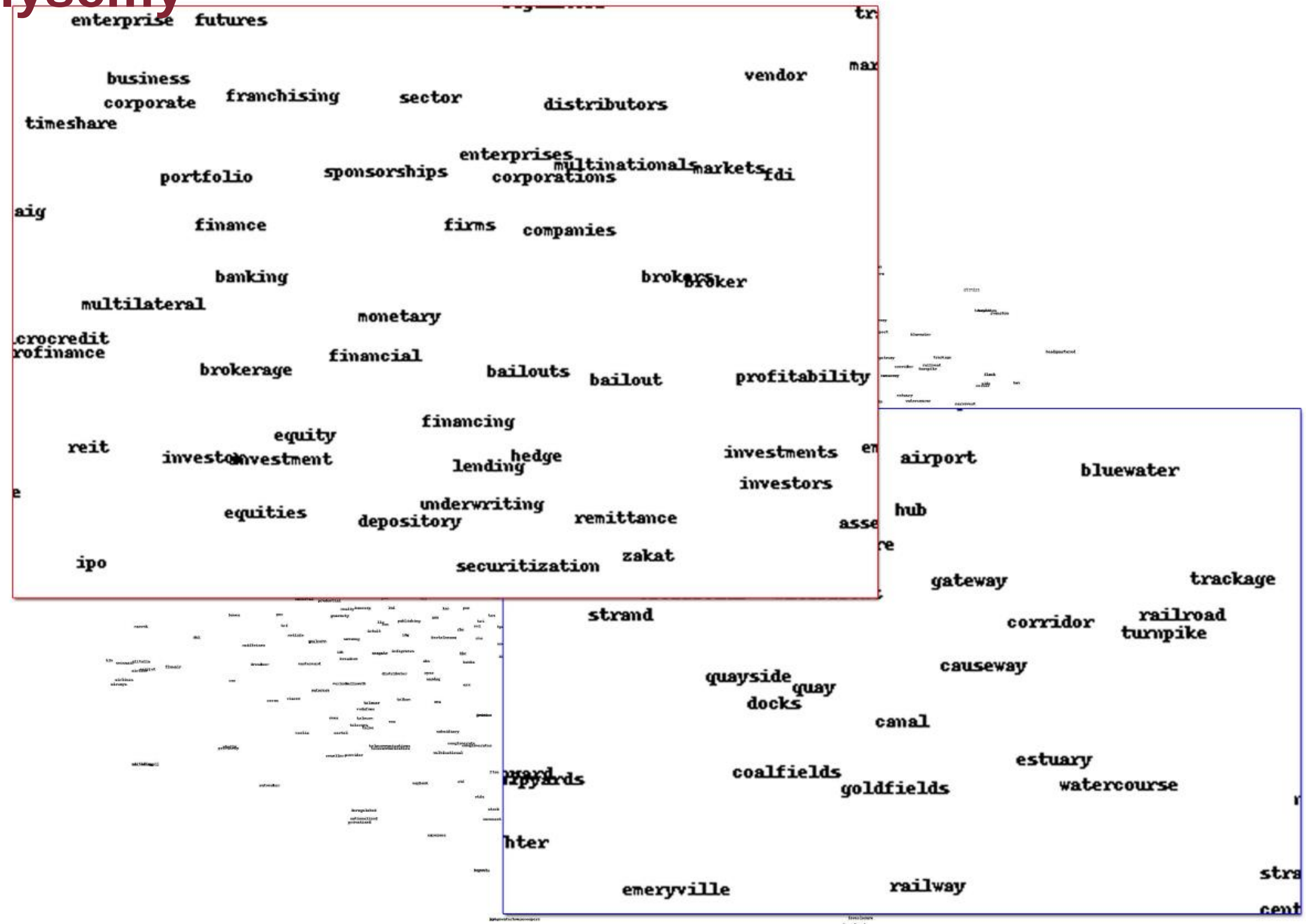
Problem: word representations cannot capture polysemy



Problem: word representations cannot capture polysemy



Problem: word representations cannot capture polysemy



Why should we care?

- With **word embeddings** we would have:

For distance d, $d(a,c) \leq d(a,b) + d(b,c)$.

pollen refinery

plant

(Neelakantan et al. 2014)

Why should we care?

- With **sense embeddings**, instead:

$$\text{For distance } d,\ d(a,c) \leq d(a,b) + d(b,c).$$

pollen $\longleftrightarrow$ refinery

pollen $\longleftrightarrow$ plant^1

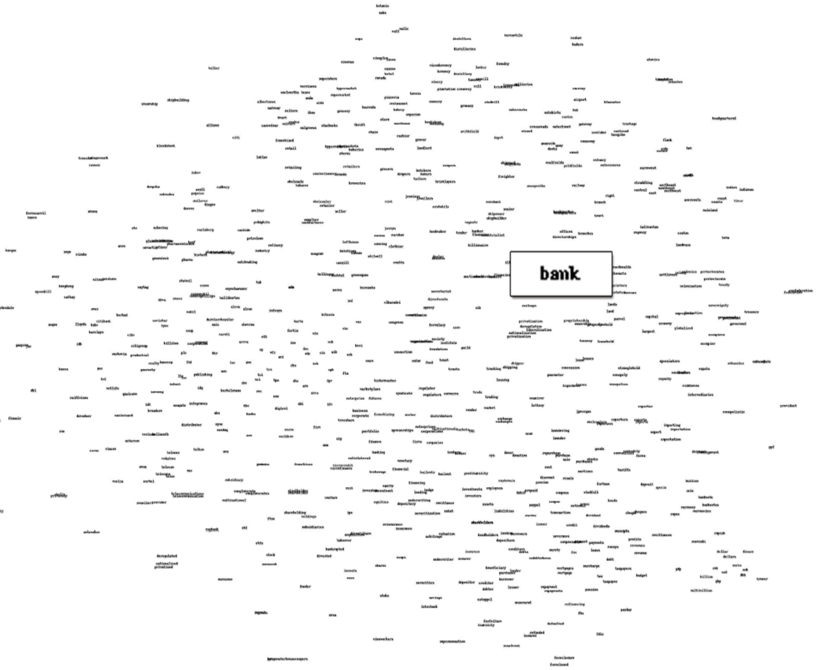
refinery $\longleftrightarrow$ plant^2

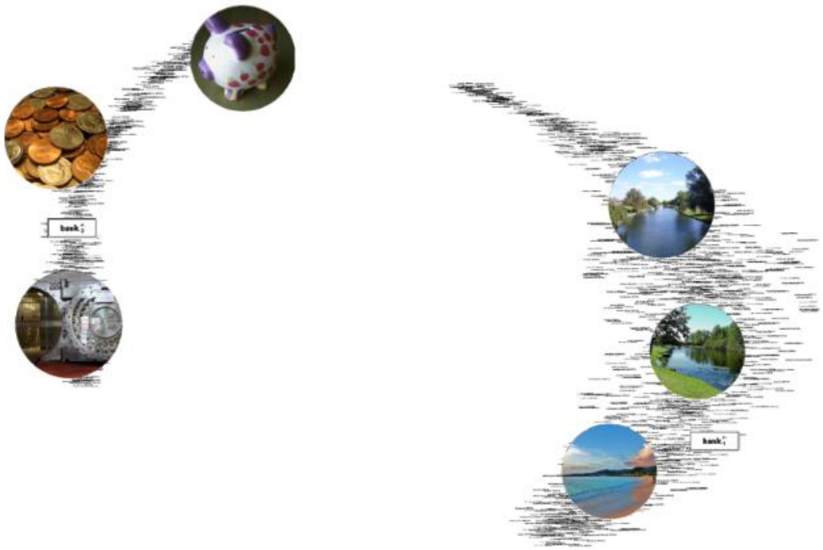
plant^1 $\longleftrightarrow$ plant^2

(Neelakantan et al. 2014)


“Language
barriers”

Solution: distinct representation for each word's meaning

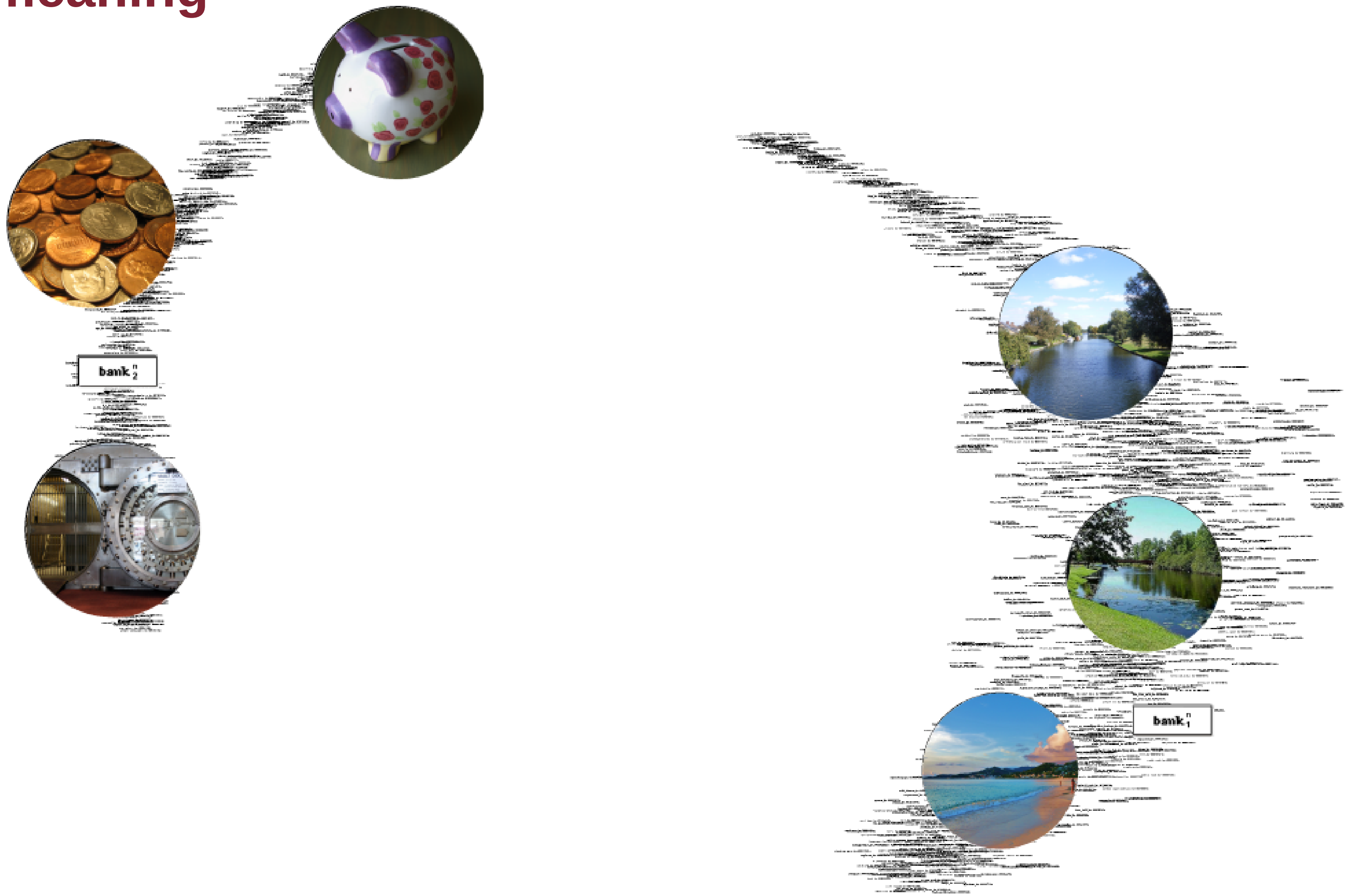




Word vector space model

Sense vector space model

Solution: distinct representation for each word's meaning



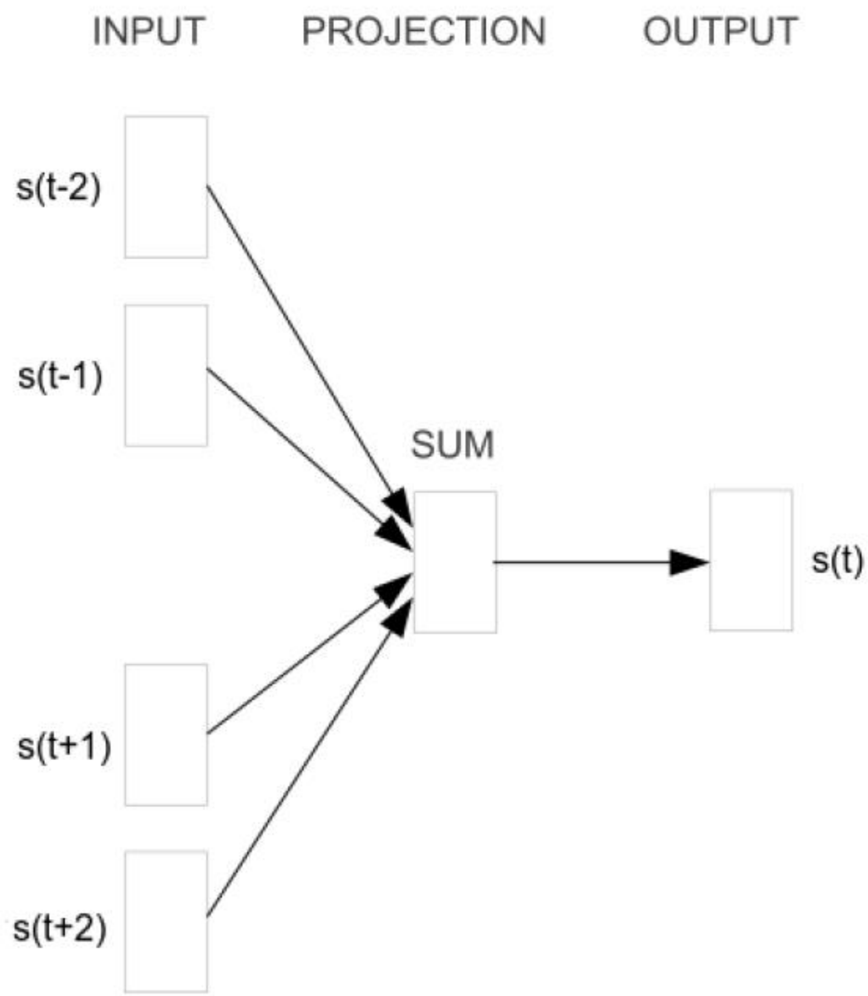
Where are we?

- Motivation for our work: word vs. sense representations
- Approach 1: SensEmbed (latent)
  - monolingual, but replicable in any language
- Approach 2: NASARI (explicit and latent versions)
  - monolingual and multilingual
- Approach 3: SW2V - Modeling words and senses jointly (latent)
  - in between
- Industrial applications @ Babelscape
- Conclusions

Latent representation of word senses: SensEmbed

Iacobacci, Pilehvar and Navigli (ACL 2015)

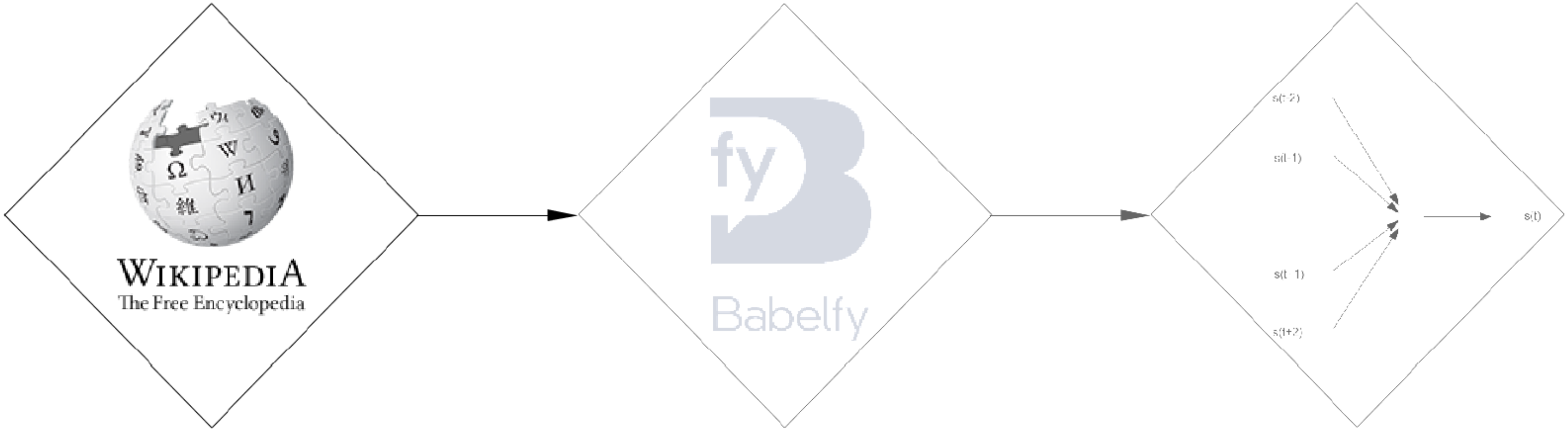
Starting point: the CBOW architecture [Mikolov et al., 2013]



Step 1: select a large corpus

...survey on the relationship between the banks and our industry , in preparation for a forthcoming forum.
...and it stands on the right bank of the Drava River , bounded by the river to the north...
... If you have dividend or receive bank or building society interest on which tax has been paid ,
...workplaces and unions. Corporations, banks and trusts controlled a great deal and , although machines...
...The critical decision for the banks will come if their own adviser sticks to his view of the costs.
countryside of high hedges and tall earth banks with trees on top. The heavily wooded area was criss-crossed...

Step 2: identify all the occurrences of a target word



...survey on the relationship between the **banks** and our industry , in preparation for a forthcoming forum.
...and it stands on the right **bank** of the Drava River , bounded by the river to the north...
... If you have dividend or receive **bank** or building society interest on which tax has been paid ,
...workplaces and unions. Corporations, **banks** and trusts controlled a great deal and , although machines...
...The critical decision for the **banks** will come if their own adviser sticks to his view of the costs.
countryside of high hedges and tall earth **banks** with trees on top. The heavily wooded area was criss-crossed...

Step 3: disambiguate each target word occurrence

...survey on the relationship between the **banks** and our industry , in preparation for a forthcoming forum.
...and it stands on the right **bank** of the Drava River , bounded by the river to the north...
... If you have dividend or receive **bank** or building society interest on which tax has been paid ,
...workplaces and unions. Corporations, **banks** and trusts controlled a great deal and , although machines...
...The critical decision for the **banks** will come if their own adviser sticks to his view of the costs.
countryside of high hedges and tall earth **banks** with trees on top. The heavily wooded area was criss-crossed...

Step 4: train CBOW with senses as targets

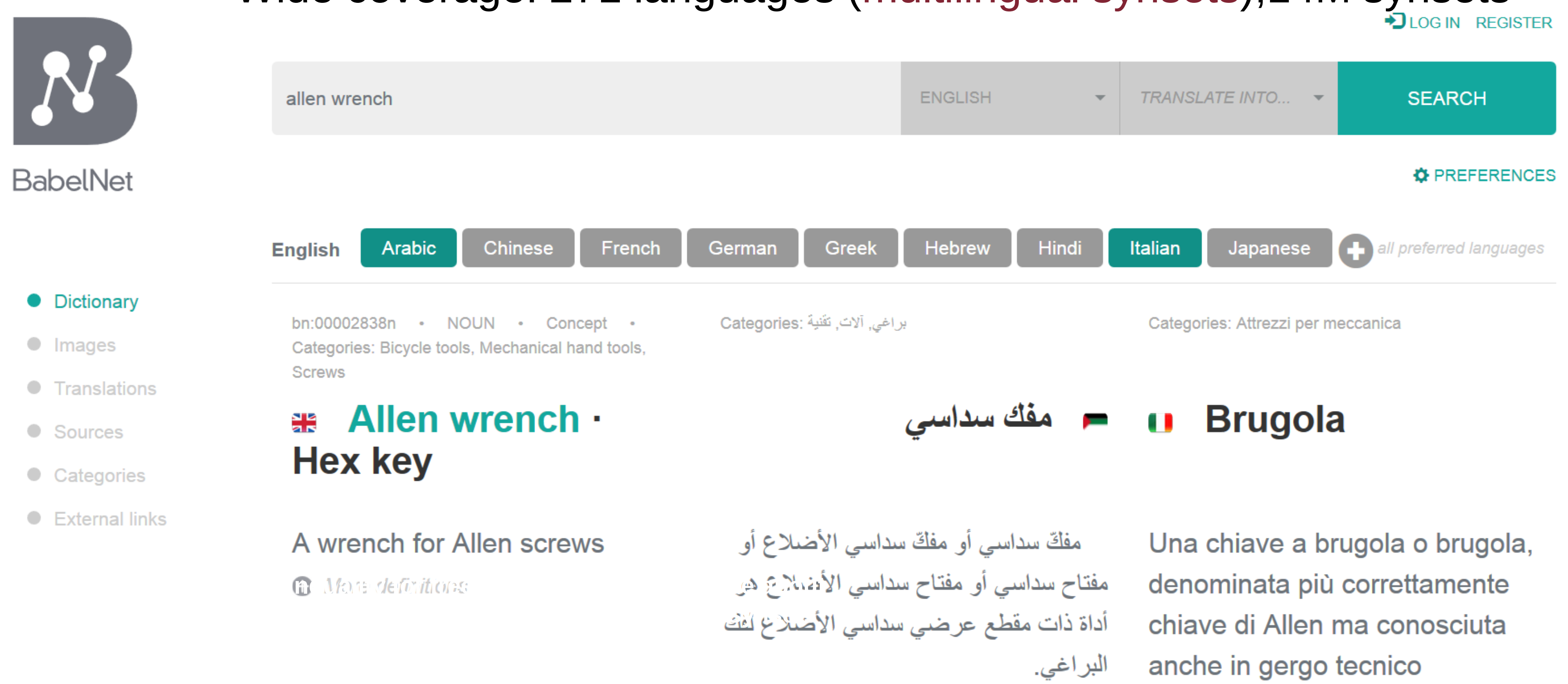
...survey on the relationship between the **banks** and our industry , in preparation for a forthcoming forum.
...and it stands on the right **bank** of the Drava River , bounded by the river to the north...
... If you have dividend or receive **bank** or building society interest on which tax has been paid ,
...workplaces and unions. Corporations, **banks** and trusts controlled a great deal and , although machines...
...The critical decision for the **banks** will come if their own adviser sticks to his view of the costs.
countryside of high hedges and tall earth **banks** with trees on top. The heavily wooded area was criss-crossed...

-2.19067 1.16642 -1.91385 -0.269672 0.712771 -0.623024 -3.20115 0.560895 0.891554 0.145258 1.26956 -0.221078 -0.0733777 2.08072 -3.30558 -0.727272 -0.902202 -1.84578 -1.38985 -0.0791954 0.989769 -1.34631 1.10242 -1.59836 -1.37341 -1.42038 0.238941 -2.98729 -0.730938 0.267584 0.0560677 -0.722721 2.23752 -2.99094 -1.45598 -0.645446 0.278277 2.28877 -0.926191 2.89934 -1.17254 1.38449 2.38617 -0.0838845 -1.80698 0.622097 0.223875 0.870654 -0.33808 -0.41957

1.16672 0.811884 -0.115492 -2.59049 -1.50286 1.2536 1.44281 0.0136615 0.131499 2.04445 -0.425782 1.29676 0.0996086 1.52687 -0.0951281 -0.715488 -0.71172 0.453871 1.08481 1.55074 0.385158 -0.116754 -0.582987 -1.56923 -0.488404 -1.07999 0.0447149 -0.733387 0.765212 2.67995 2.51105 0.192151 1.49743 2.91849 1.86901 0.23101 0.381663 1.20355 0.126758 1.57204 -0.372069 -2.45076 0.514557 -1.4028 -1.20396 0.726036 2.41265 -0.104843 2.26862 1.21729

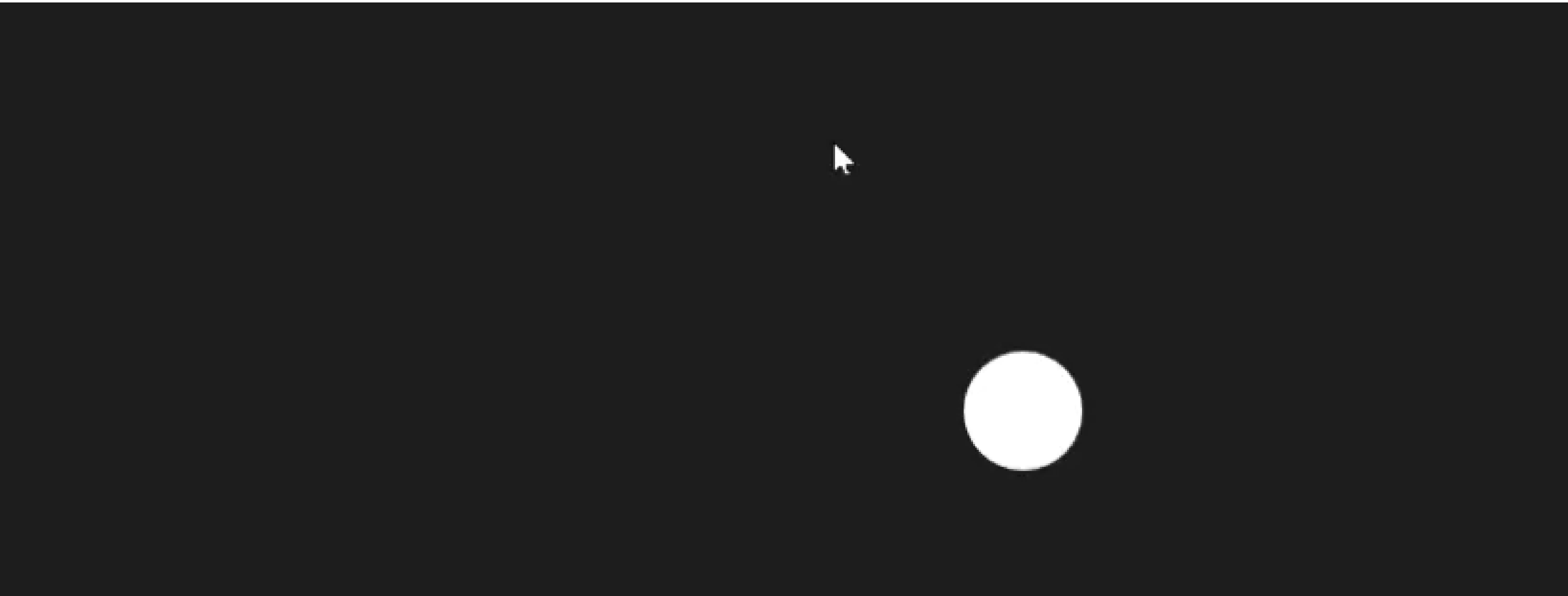
Setup – Sense inventory: BabelNet (Navigli and Ponzetto, AI Journal 2012)

- We used BabelNet, a merger of WordNet, Wikipedia, Wiktionary, OmegaWiki and other knowledge resources
- Why?
  - An extension of the lexical-semantic knowledge model of WordNet
  - Wide coverage: 271 languages (multilingual synsets),14M synsets



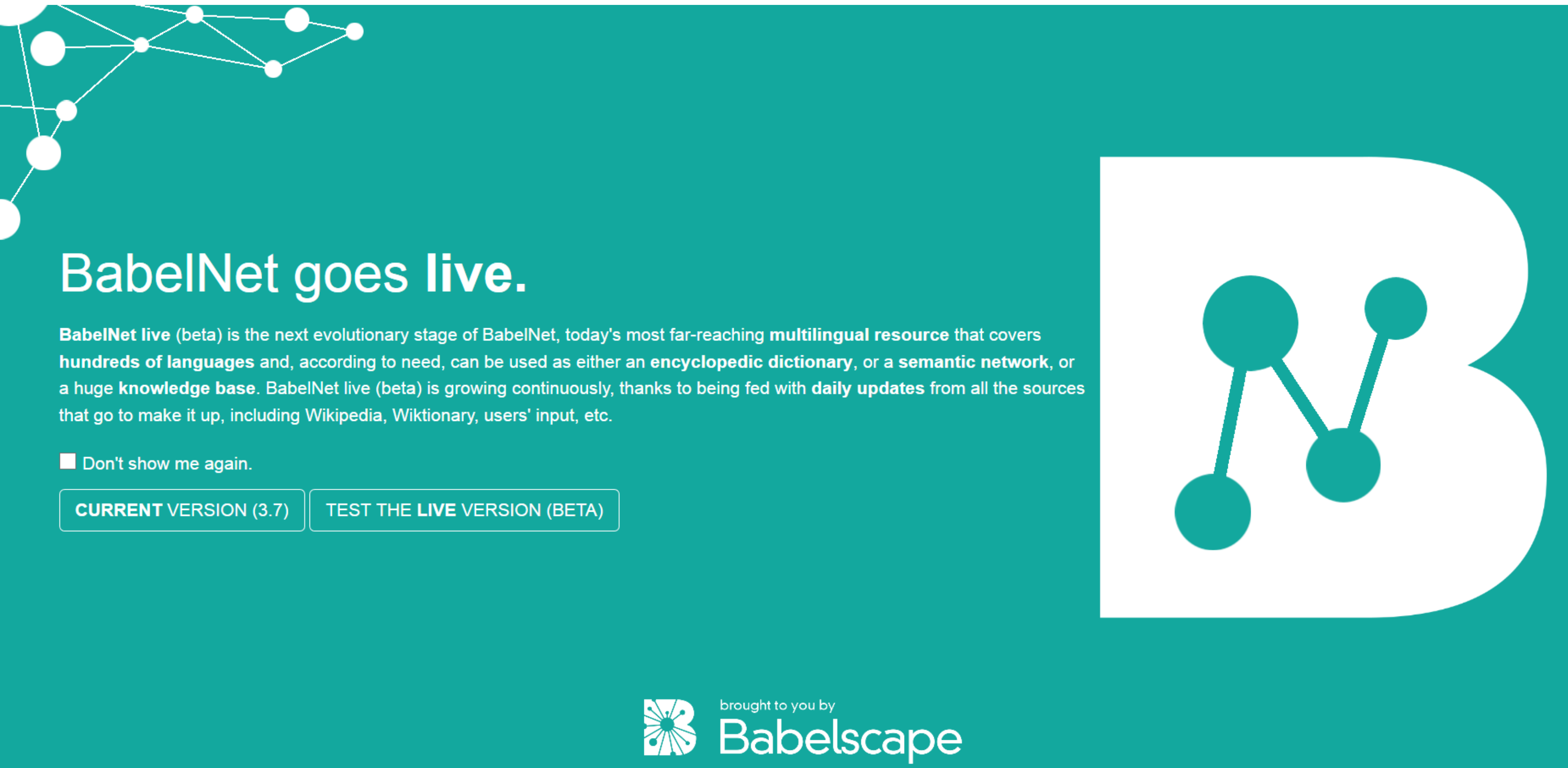
Setup – Sense inventory: BabelNet (Navigli and Ponzetto, AI Journal 2012)

- We used BabelNet, a merger of WordNet, Wikipedia, Wiktionary, OmegaWiki and other knowledge resources
- Why?
  - An extension of the lexical-semantic knowledge model of WordNet
  - Wide coverage: 271 languages (multilingual synsets),14M synsets
  - It integrates concepts (6M) and named entities (7.7M) seamlessly



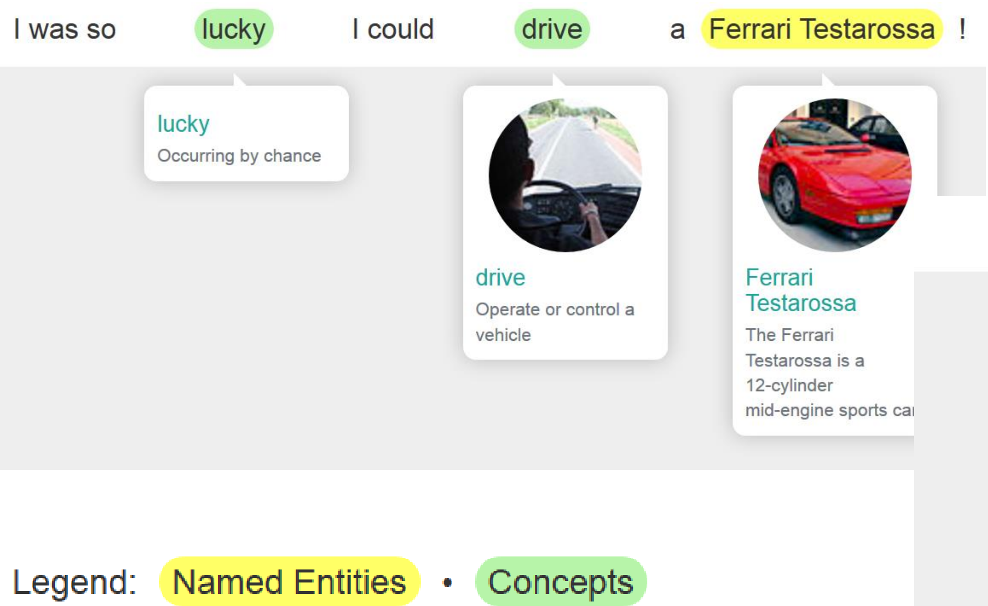
BabelNet is now live!

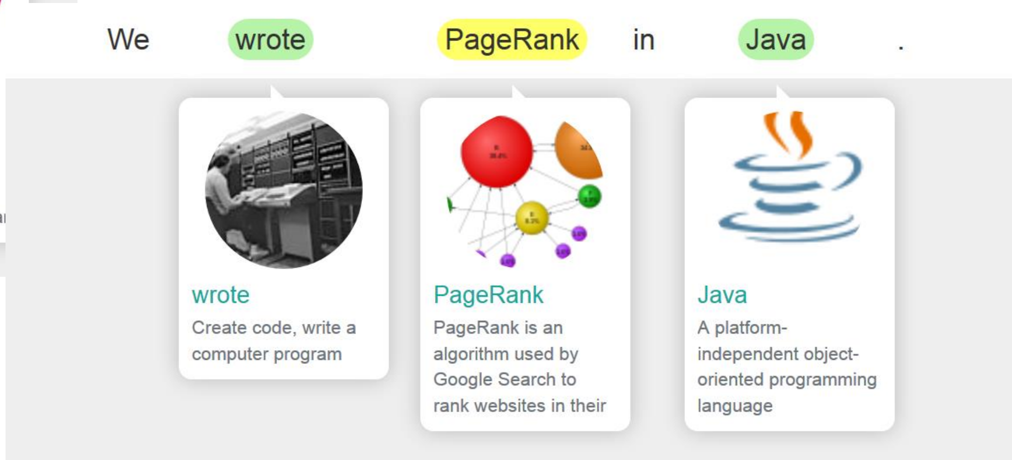
- 284 languages
- 15 million concepts and named entities
- 1.8 billion semantic relations



Setup – Disambiguation: Babelfy [Moro et al., TACL 2014]

- We used Babelfy for disambiguating the Wikipedia corpus
- Why?
  – The first (and only) system that performs Word Sense Disambiguation (common nouns, verbs, adjectives, adverbs) and Entity Linking (names) **jointly**





Setup – Disambiguation: Babelfy [Moro et al., TACL 2014]

- We used Babelfy for disambiguating the Wikipedia corpus
- Why?
  - The first (and only) system that performs Word Sense Disambiguation (common nouns, verbs, adjectives, adverbs) and Entity Linking (names) **jointly**
  - Knowledge-based: does not need millions of sentences annotated in each language (Pilehvar and Navigli, 2015)
  - Works in arbitrary languages (271 languages)
  - Can disambiguate texts written in mixed languages (language-agnostic setting)
  - [Demo on recent news]

Qualitative Evaluation

- Closest senses to different senses of ambiguous words:

$\textbf{\textit{bank}}^n_1$ (geographical)	$\textbf{\textit{bank}}^n_2$ (financial)	$\textbf{\textit{number}}^n_4$ (phone)	$\textbf{\textit{number}}^n_3$ (acting)	$\textbf{\textit{hood}}^n_1$ (gang)	$\textbf{\textit{hood}}^n_{12}$ (convertible car)
upstream^r_1	$\text{commercial\_bank}^n_1$	calls^n_1	appearing^v_6	tortures^n_5	taillights^n_1
downstream^r_1	$\text{financial\_institution}^n_1$	dialled^v_1	minor_roles^n_1	vengeance^n_1	grille^n_2
runs^v_6	national_bank^n_1	operator^n_{20}	$\text{stage\_production}^n_1$	badguy^n_1	bumper^n_2
confluence^n_1	trust_company^n_1	$\text{telephone\_network}^n_1$	$\text{supporting\_roles}^n_1$	brutal^a_1	fascia^n_2
river^n_1	savings_bank^n_1	telephony^n_1	leading_roles^n_1	execution^n_1	rear_window^n_1
stream^n_1	banking^n_1	subscriber^n_2	stage_shows^n_1	murders^n_1	headlights^n_1

Quantitative Evaluation: word similarity – results

Measure	Dataset				
	RG-65	WS-Sim	WS-Rel	YP-130	MEN
Pilehvar et al. (2013)	0.868	0.677	0.457	0.710	0.690
Zesch et al. (2008)	0.820	—	—	0.710	—
Collobert and Weston (2008)	0.480	0.610	0.380	—	0.570
Word2vec (Baroni et al., 2014)	0.840	0.800	0.700	—	0.800
GloVe	0.769	0.666	0.559	0.577	0.763
ESA	0.749	—	—	—	—
PMI-SVD	0.738	0.659	0.523	0.337	0.726
Word2vec	0.732	0.707	0.476	0.343	0.665
SENSEMBED$_{closest}$	**0.894**	0.756	0.645	**0.734**	0.779
SENSEMBED$_{weighted}$	0.871	**0.812**	**0.703**	0.639	**0.805**

- State-of-the-art performance + sense-level vectors in the same space as word vectors

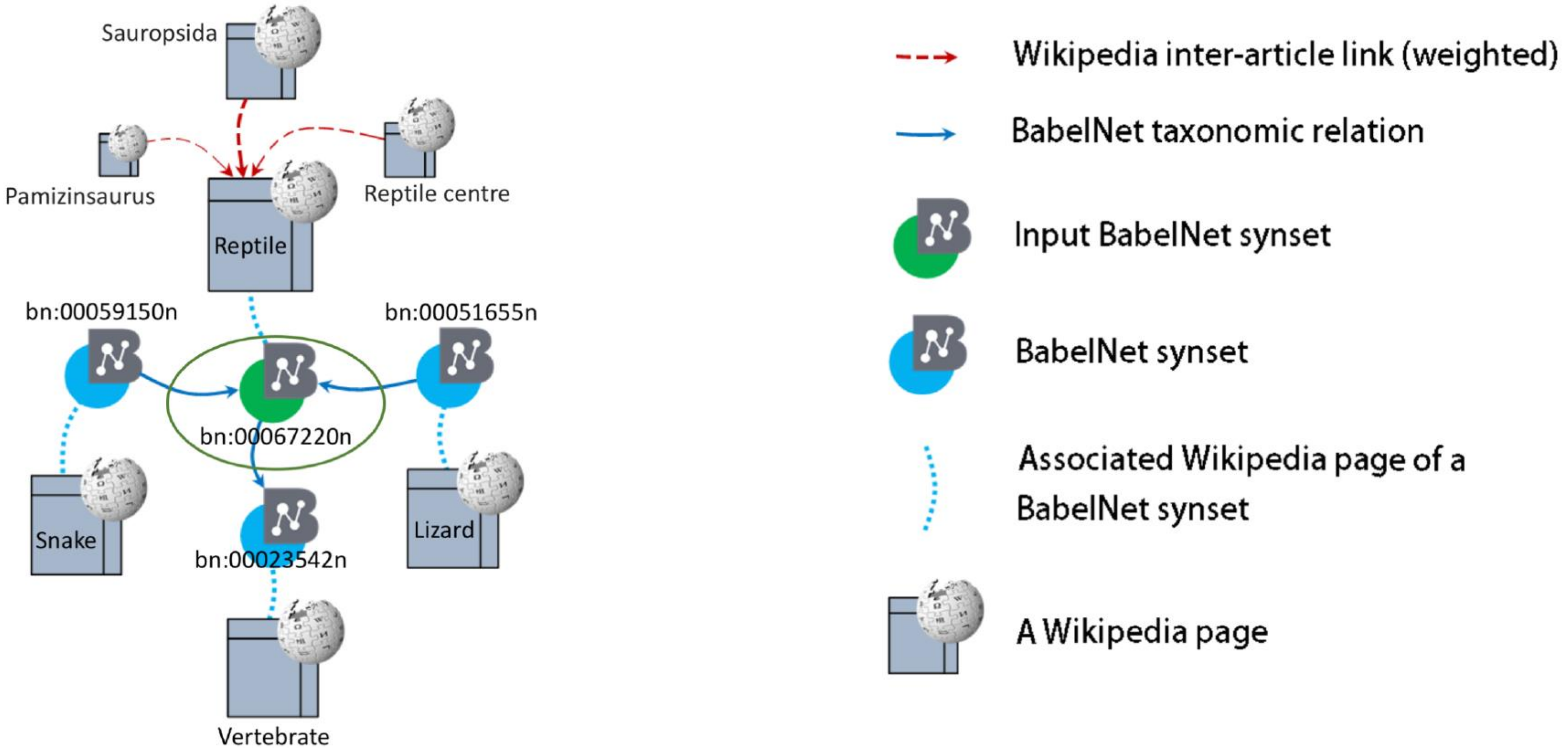
Explicit representation of concepts: NASARI

Camacho-Collados, Pilehvar and Navigli
(NAACL 2015; ACL 2015;
Artificial Intelligence Journal 2016)

Motivation

Idea 1: collect documents about a concept/entity

- For a **given concept/entity**, the initial idea is to collect a corpus of documents (Wikipedia pages) about it



Idea 2: we can create 3 different vector representations

- The collected corpus will be a **subcorpus** of a given **reference corpus** (the whole Wikipedia)
- The goal is to create a vector that represents the semantics of the **concept of interest**
- Three variants:
  - Lexical vectors (having words as components)
  - Unified vectors (**language-independent**, having BabelNet synsets as components)
  - Embedded vectors (having latent dimensions)

Calculating lexical specificity

- Given:
  - a reference corpus of T words (Wikipedia)
  - a subcorpus of t words (our set of Wikipedia pages)
- **Goal**: find a set of terms that are peculiar to the subcorpus, but not to the whole reference corpus.
- Given a word w that occurs F and f times in the corpus and subcorpus, respectively, compute the relevance of w to the subcorpus as a function of P (X ≥ f), X being a random variable following a hypergeometric distribution with parameters F, t and T.

$$spec(T, t, F, f) = -\log_{10} P(X \geq f)$$

NASARI: the lexical vector

- Conventional vector with words as dimensions
- Individual weights calculated using lexical specificity, by contrasting the frequencies in the subcorpus and the overall corpus (whole Wikipedia)
- **Pruning:** we keep only components with $P(X \geq f) \leq 0.01$
- **Example:** top-ranking components of 2 meanings of bank:

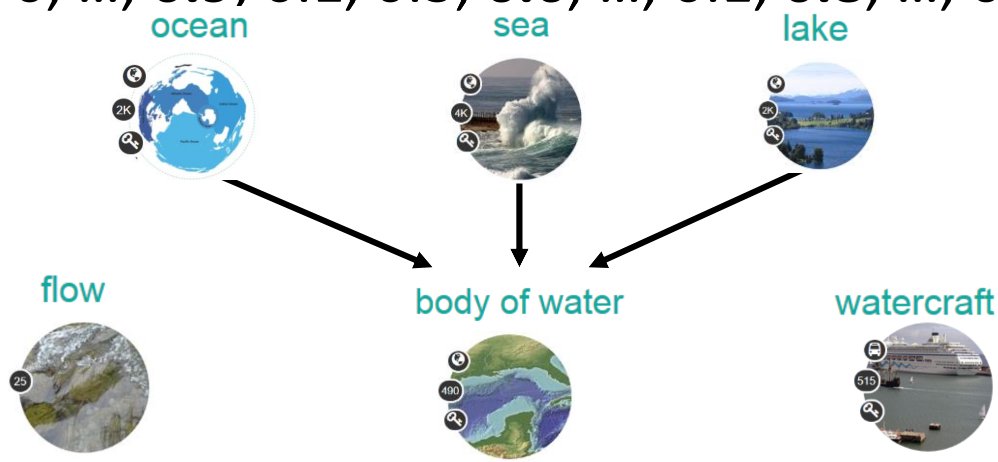
Bank (financial institution)			**Bank (geography)**		
English	French	Spanish	English	French	Spanish
bank	banque	banco	river	eau	banco
banking	bancaire	bancario	stream	castor	limnología
deposit	crédit	banca	bank	berge	ecología
credit	financier	financiero	riparian	canal	barrera
money	postal	préstamo	creek	barrage	estuarios
loan	client	entidad	flow	zone	isla
commercial_bank	dépôt	déposito	water	perchlorate	interés
central_bank	billet	crédito	watershed	humide	laguna

NASARI: the unified vector

- Cluster together similar dimensions in the lexical vector
- Then re-compute the weights for the new dimensions

$v_{lex}($  $)= (0, 0, ..., \mathbf{0.9}, 0.1, 0.3, \mathbf{0.6}, ..., 0.1, \mathbf{0.8}, ..., 0)$

lexical

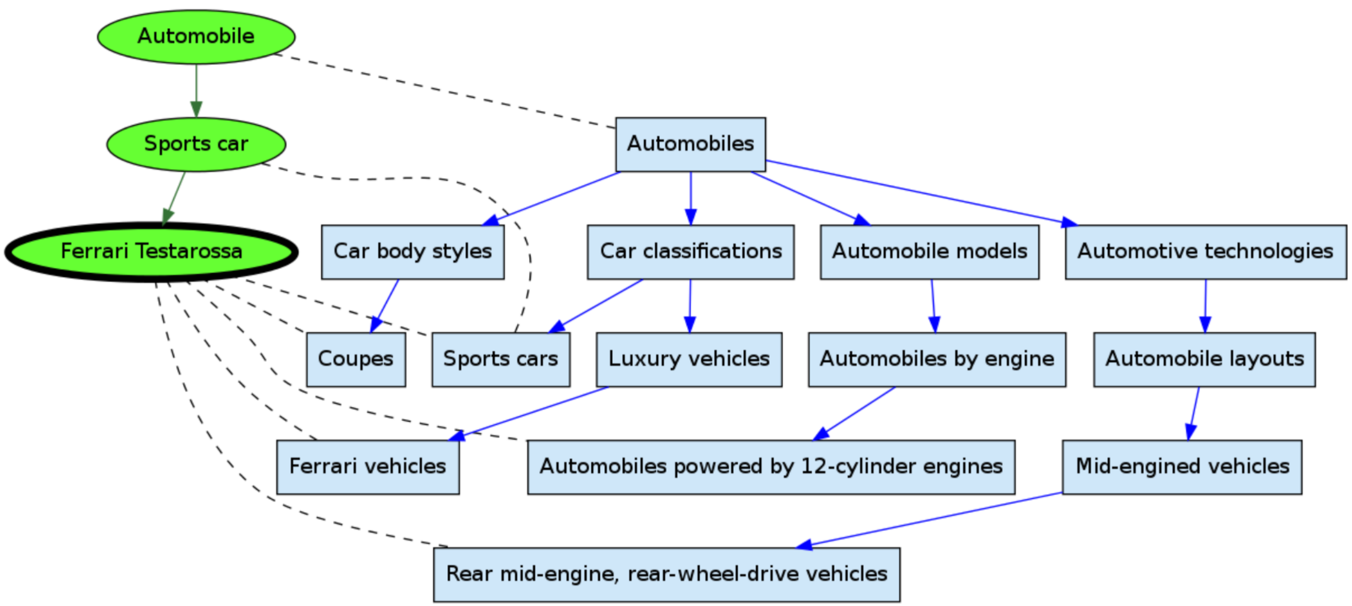


$v_{u}($  $)= (0.1, 0.3, ..., 0, 0, 0.5, 0.95, ..., 0.5, ..., 0.15)$

semantic

We use the BabelNet taxonomy

- BabelNet provides a full-fledged taxonomy: is-a relations are available for millions of concepts and named entities (**Wikipedia Bitaxonomy**, Flati et al. ACL 2014; AIJ 2016)
  - Ferrari Testarossa *is-a* sports car
  - BabelNet *is-a* semantic network & encyclopedic dictionary



NASARI: the unified vector

- Unified vectors have BabelNet synsets as dimensions
- Two key benefits:
  - Disambiguated dimensions
  - Smoothing
- Enables
  - Transfer of semantic knowledge across languages
  - Cross-lingual semantic comparison

Bank (financial institution)			Bank (geography)		
English	French	Spanish	English	French	Spanish
‡bank^2_n	‡banque^1_n	‡banco^1_n	⋆stream^1_n	eau^1_n	inclinación^9_n
reserve^2_n	•fonds^2_n	⋆Institución_financiera^1_n	river^1_n	eau$^{15}_n$	lago^1_n
⋆financial_institution^1_n	⋄dépôt^9_n	⋄depósito$^{15}_n$	‡body_of_water^1_n	excrément^1_n	‡cuerpo_de_agua^1_n
⋄deposit^8_n	∘emprunt^2_n	†Finanzas^1_n	flow^1_n	castor^1_n	⋆arroyo^1_n
banking^2_n	paiement^1_n	•dinero^2_n	course^2_n	‡étendue_d'eau^1_n	tierra$^{11}_n$
†finance^1_n	argent^2_n	∘préstamo^2_n	bank^1_n	fourrure^1_n	costa^1_n

NASARI: embedded representation

- We calculate a weighted average of the word embeddings of the lexical components of the vector for a given subcorpus T (corresponding to a concept of interest):

$$E(\mathcal{T}) = \frac{\sum_{w \in \vec{v}_{lex}(\mathcal{T})} \left(\frac{1}{rank(w, \vec{v}_{lex}(\mathcal{T}))} E(w) \right)}{\sum_{w \in \vec{v}_{lex}(\mathcal{T})} \frac{1}{rank(w, \vec{v}_{lex}(\mathcal{T}))}}$$

- **Key feature:** words and senses in the same space!
- Example of closest embedded vectors:

Bank (financial institution)		**Bank (geography)**		***bank***	
Closest senses	Cosine	Closest senses	Cosine	Closest senses	Cosine
Deposit account	0.99	Stream bed	0.98	Bank (financial institution)	0.86
Universal bank	0.99	Current (stream)	0.97	Universal bank	0.86
British banking	0.98	River engineering	0.97	British banking	0.86
German banking	0.98	Braided river	0.97	German banking	0.85
Commercial bank	0.98	Fluvial terrace	0.97	Branch (banking)	0.85
Banking in Israel	0.98	Bar (river morphology)	0.97	McFadden Act	0.85
Financial institution	0.98	River	0.97	Four Northern Banks	0.84
Community bank	0.97	Perennial stream	0.96	State bank	0.84

Experiments (Camacho-Collados et al., AI Journal 2016)

- Word similarity
- Cross-lingual similarity
  - RG-65 in English, German and French
- Word Sense Disambiguation (WSD)
  - Multilingual WSD
- Domain labeling
  - "BabelDomains: Large-Scale Domain Labeling of Lexical Resources" (Camacho-Collados and Navigli, EACL 2017)
- Sense clustering

Cross-lingual Word similarity

English	r	ρ	French	r	ρ	German	r	ρ	Spanish	r	ρ
NASARI	0.81	0.78	NASARI	**0.82**	0.73	NASARI	0.69	0.65	NASARI	**0.85**	0.79
NASARI$_{lexical}$	0.80	0.78	NASARI$_{lexical}$	0.80	0.70	NASARI$_{lexical}$	0.69	0.67	NASARI$_{lexical}$	**0.85**	0.79
NASARI$_{unified}$	0.80	0.76	NASARI$_{unified}$	**0.82**	**0.76**	NASARI$_{unified}$	**0.71**	**0.68**	NASARI$_{unified}$	0.82	0.77
NASARI$_{embed}$	**0.82**	**0.80**	–	–	–	–	–	–	NASARI$_{embed}$	0.79	0.77
SOC-PMI	0.61	–	SOC-PMI	0.19	–	SOC-PMI	0.27	–	–	–	–
PMI	0.41	–	PMI	0.34	–	PMI	0.40	–	–	–	–
LSA-Wiki	0.65	0.69	LSA-Wiki	0.57	0.52	–	–	–	–	–	–
Wiki-wup	0.59	–	–	–	–	Wiki-wup	0.65	–	–	–	–
Word2Vec	–	0.73	Word2Vec	–	0.47	Word2Vec	–	0.53	Best-Word2Vec	0.80	**0.80**
Retrofitting	–	0.77	Retrofitting	–	0.61	Retrofitting	–	0.60	–	–	–
NASARI$_{poly\text{-}embed}$	0.74	0.77	NASARI$_{poly\text{-}embed}$	0.60	0.69	NASARI$_{poly\text{-}embed}$	0.46	0.52	NASARI$_{poly\text{-}embed}$	0.68	0.74
Polyglot-embed	0.51	0.55	Polyglot-embed	0.38	0.35	Polyglot-embed	0.18	0.15	Polyglot-embed	0.51	0.56
IAA	0.85$^{\diamond}$	-	IAA	-	-	IAA	0.81	-	IAA	0.83	-

Spearman (ρ) and Pearson (r) correlation performance of different systems on multilingual editions of the RG-65 datasets.

Comparison systems:

- SOC-PMI and PMI (Joubarne and Inkpen, 2011) – 1st and 2nd order co-occ.
- Retrofitting (Faruqui et al., 2015)
- Wiki-wup (Ponzetto and Strube, 2015)
- LSA-Wiki (Granada et al., 2014)
- Polyglot-embed (Al-Rfou et al., 2013) – emb. on wikipedias in many languages

Understanding text: Word Sense Disambiguation

kiwi
(target word)

“we are fond of fruit such as **kiwi** and banana”
(context)

Resources

Word Sense Disambiguation system

output sense

Multilingual Word Sense Disambiguation

- **Sense choice:** the best sense is given by the NASARI vector closest to the text vector:

$$\hat{s} = \underset{s \in \mathcal{L}_w}{\operatorname{argmax}}\, WO(\vec{v}_{lex}(\mathcal{T}), \overrightarrow{\text{NASARI}}_{lex}(s))$$

- **Dataset:** the Wikipedia sense inventory for the SemEval-2013 all-words multilingual WSD task (Navigli et al. 2013) – from 1242 to 1039 annotated instances
- **Evaluation measure:** F1-measure
- **Results:**

System	English	French	Italian	German	Spanish	Average
NASARI	86.3	**76.2**	83.7	**83.2**	82.9	**82.5**
MUFFIN	84.5	71.4	81.9	83.1	**85.1**	81.2
Babelfy	**87.4**	71.6	**84.3**	81.6	83.8	81.7
UMCC-DLSI	54.8	60.5	58.3	61.0	58.1	58.5
MFS	80.2	74.9	82.2	83.0	82.1	79.3

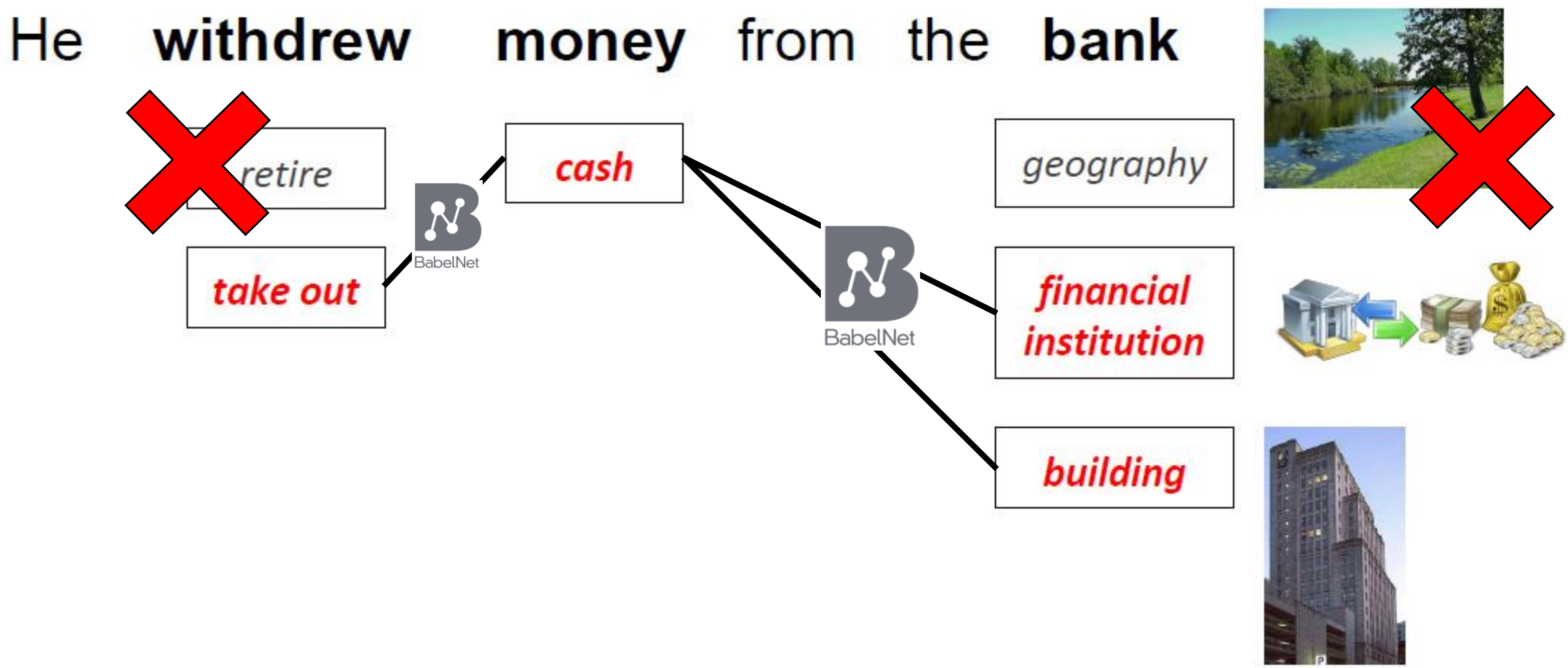
Latent representation of words AND senses together: SW2V

Mancini, Camacho, Iacobacci and Navigli
(CoNLL 2017)

Objective

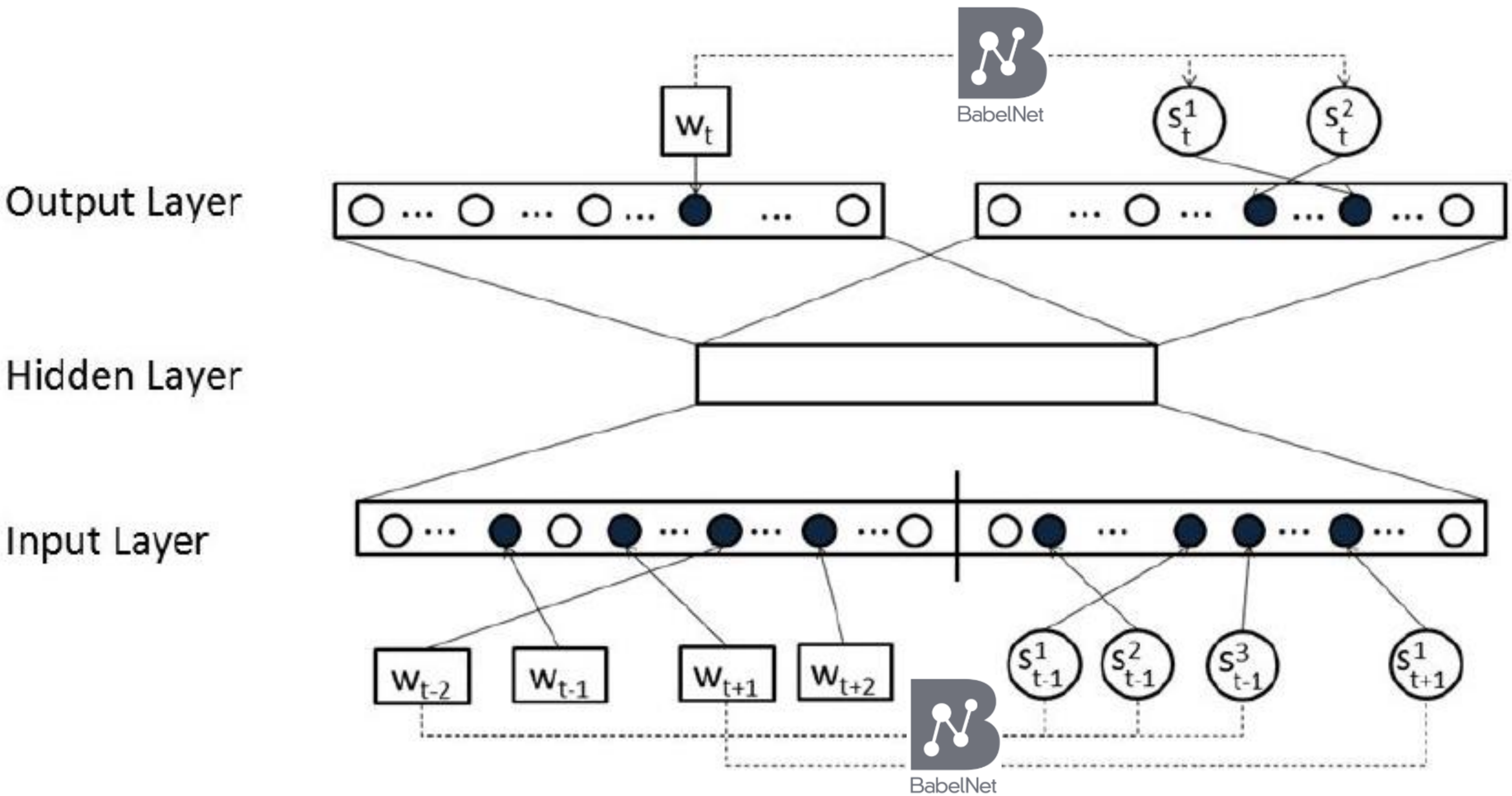
- Other approaches model either senses (SensEmbed) or obtain embeddings as a result of postprocessing word embeddings
- **Goal:** modeling words and senses in the same vector space
- **How:** exploiting the explicit relationships between words and senses available in BabelNet *for the words in context*

Example


He withdrew money from the bank
retire
take out
BabelNet
cash
BabelNet
geography
financial institution
building

Extending Word2Vec with senses

$$E=-\log(p(w_t|W^t,S^t)) - \sum_{s\in St} \log(p(s|W^t,S^t))$$



Words and their associated senses used in the input and output layers

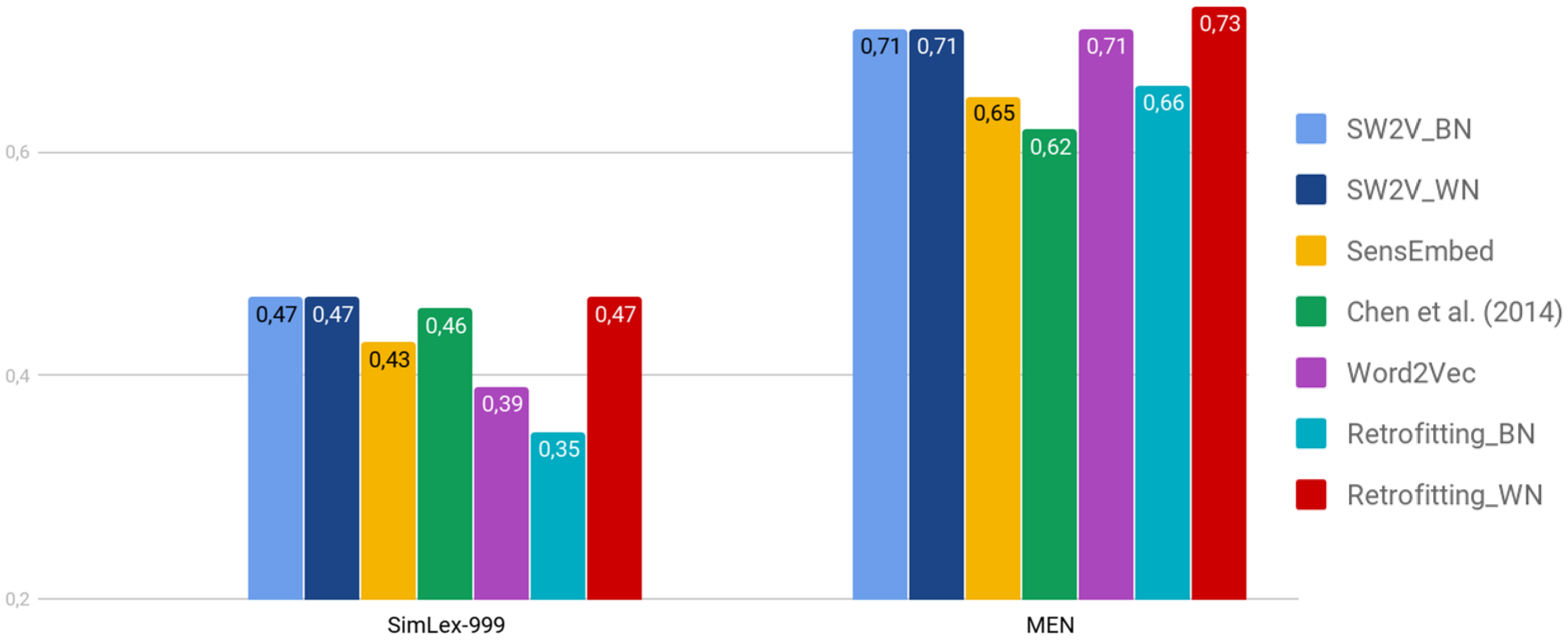
Words+Senses as input and output in SW2V

- The **best configuration** is with senses only as input and words+senses as output
  - On: WS-Sim and RG-65

		Output											
		Words				Senses				Both			
		WS-Sim		RG-65		WS-Sim		RG-65		WS-Sim		RG-65	
		r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
Input	Words	0.49	0.48	0.65	0.66	0.56	0.56	0.67	0.67	0.54	0.53	0.66	0.65
	Senses	0.69	0.69	0.70	0.71	0.69	0.70	0.70	**0.74**	**0.72**	**0.71**	**0.71**	**0.74**
	Both	0.60	0.65	0.67	0.70	0.62	0.65	0.66	0.67	0.65	**0.71**	0.68	0.70

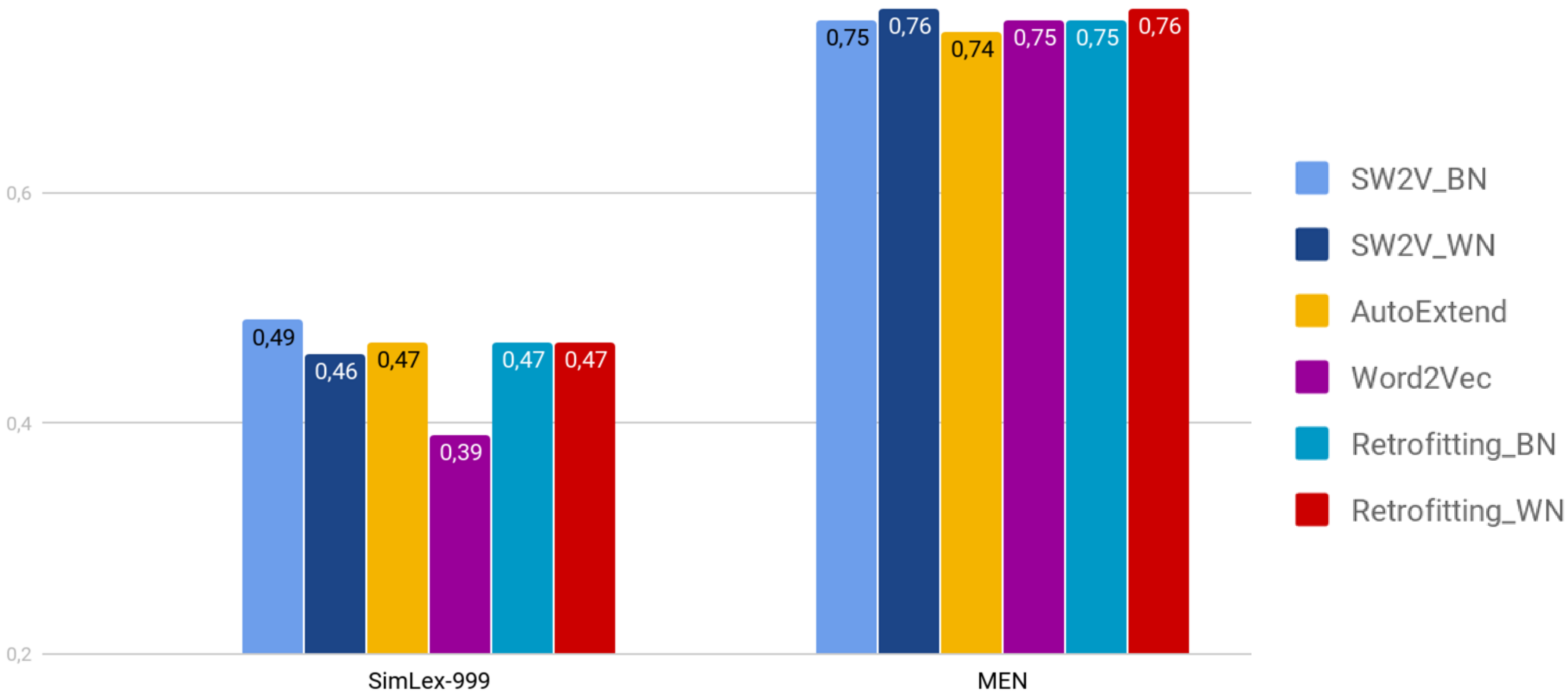
Evaluation: word similarity

- All models using Wikipedia corpus (Pearson correlation)



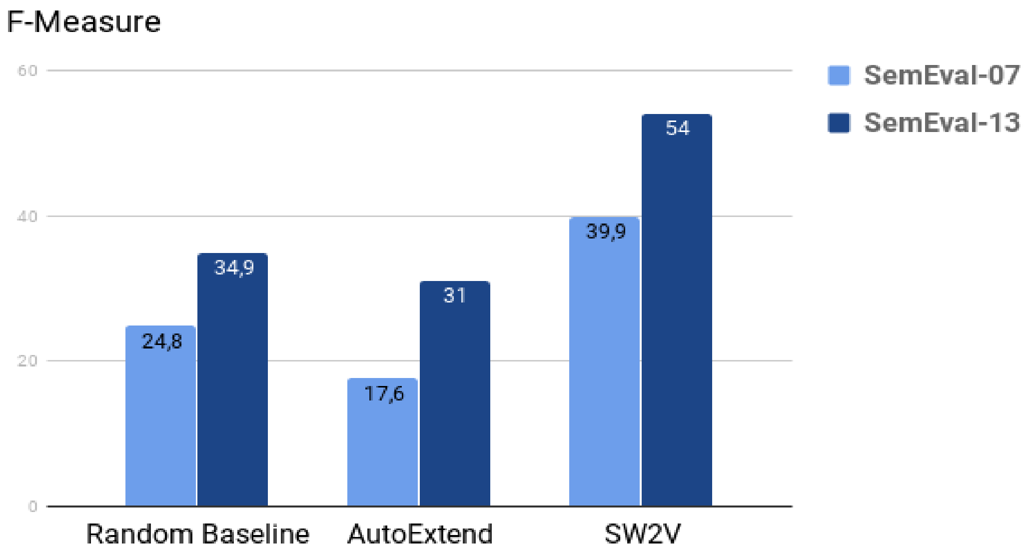
Evaluation: word similarity

- All models using UMBC corpus (Pearson correlation)



Evaluation: Most Frequent Sense for WSD

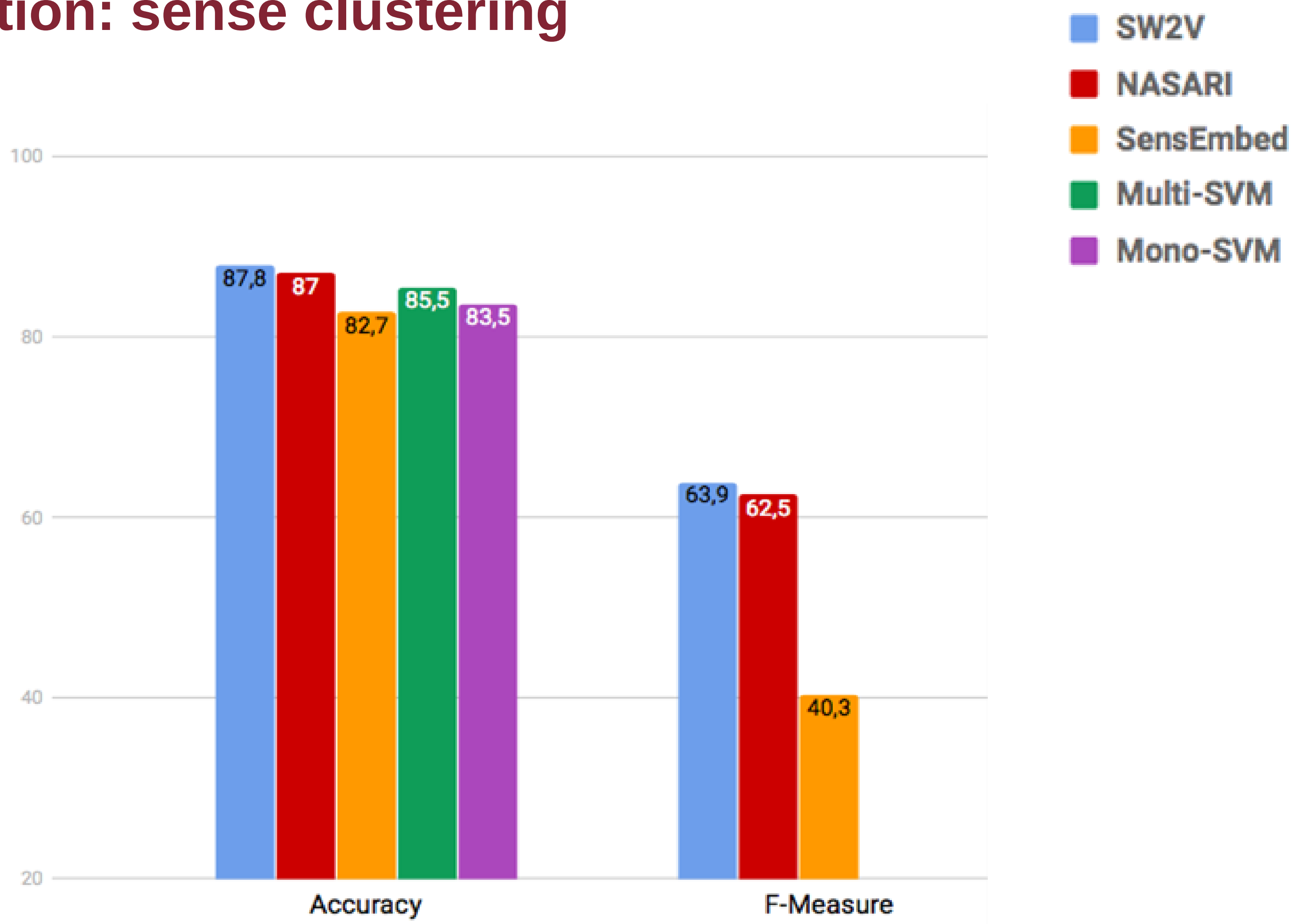
- **Evaluation:** use closeness of sense vectors to word vectors to determine sense frequency
  - We can calculate the Most Frequent Sense for each word
- **Test:** Semeval-2007 and Semeval-2013 all-words WSD



Evaluation: sense clustering

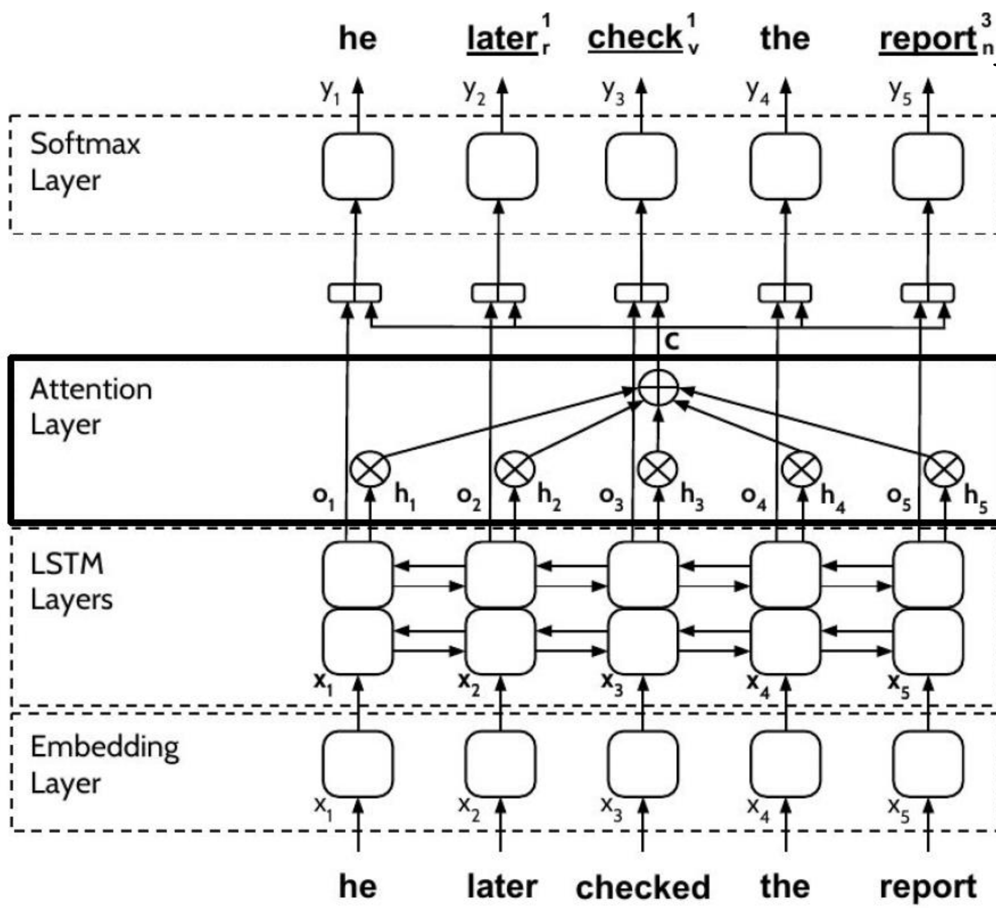
- Now we can perform semantic tasks
- **Goal:** tackle the fine granularity of sense inventories
- **Evaluation datasets** from Dandala et al. (2013)
  - Highly ambiguous words from past SemEval competitions

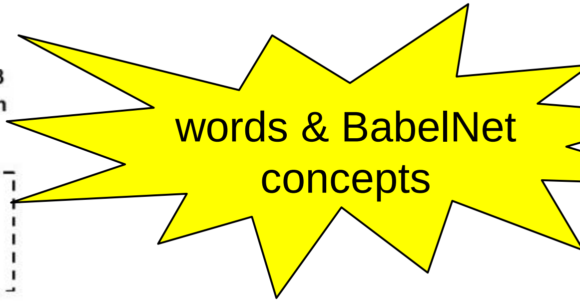
Evaluation: sense clustering



Neural Models for Word Sense Disambiguation (Raganato, Delli Bovi, Navigli, EMNLP 2017)

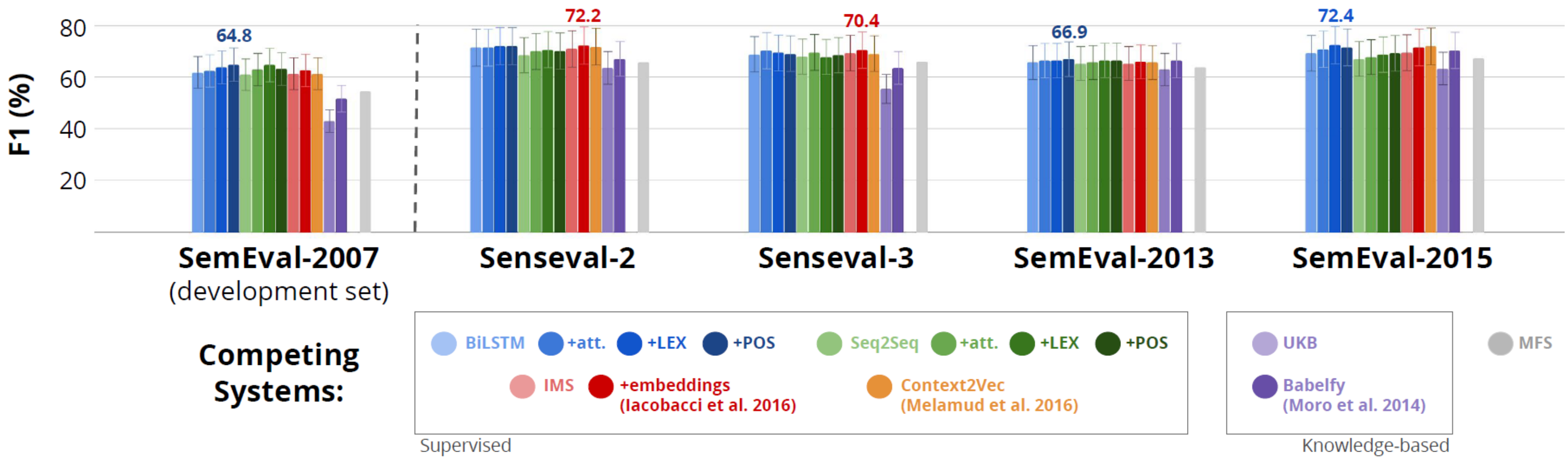
- Sequence labeling:





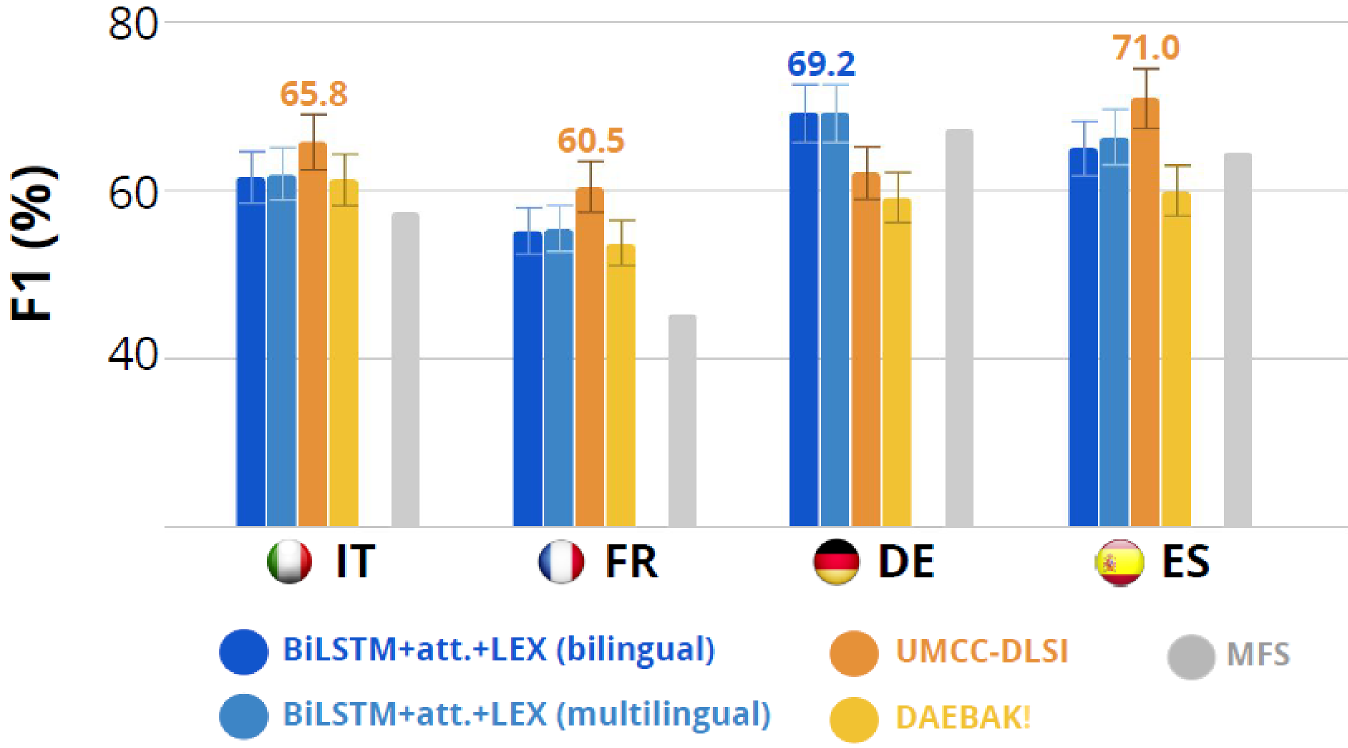
Neural Models for Word Sense Disambiguation (Raganato, Delli Bovi, Navigli, EMNLP 2017)

- Training on English (SemCor sense annotated data)
- Testing on all English Senseval & SemEval test sets



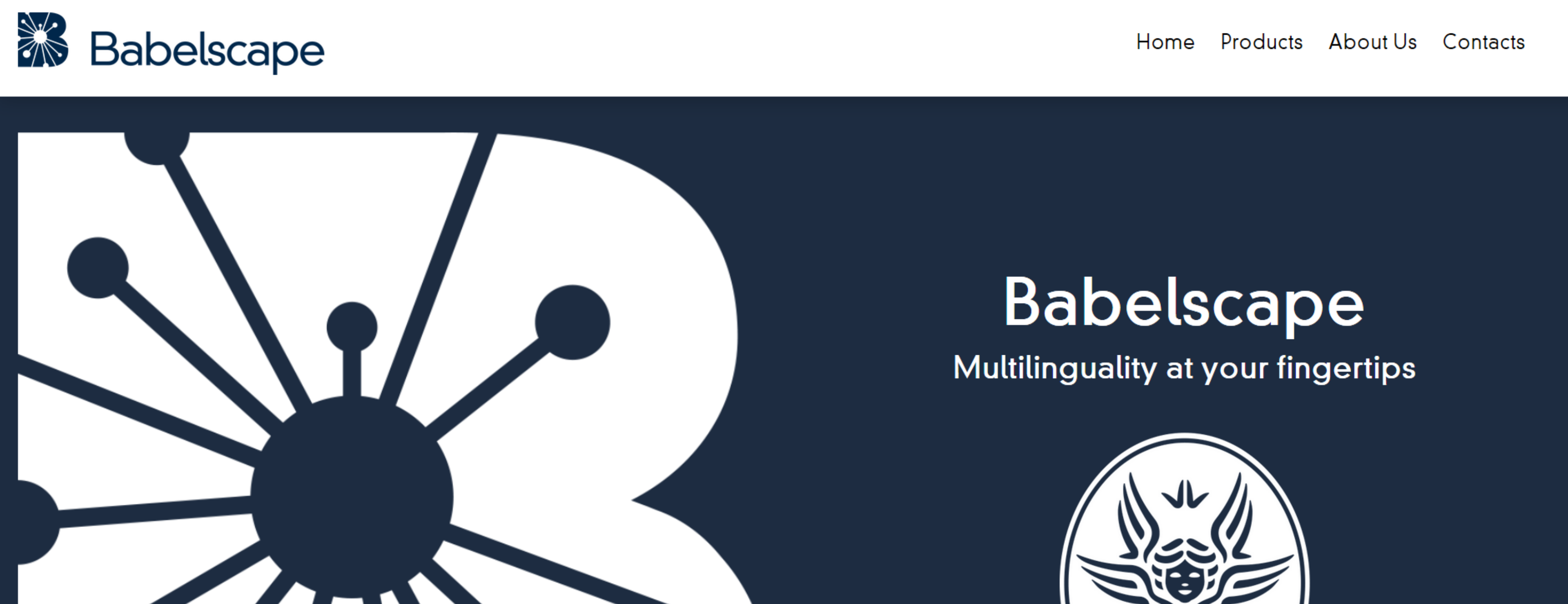
Neural Models for Word Sense Disambiguation (Raganato, Delli Bovi, Navigli, EMNLP 2017)

- Training on English (SemCor sense annotated data)
- **Testing on arbitrary languages** (!) – SemEval 2013
  - Using multilingual embeddings to encode words in the same space



The future of BabelNet and related technologies

- The MultiJEDI ERC project is now over (but: the MOUSSE ERC grant just started)
  - moving to sentence representations
- However, much work still to be done in this direction
- We created a Sapienza startup, Babelscape, with the **key objective** of making BabelNet sustainable
- Income is reinvested in BabelNet and subsequent projects



Wrapping up

- We advocated for **linking to BabelNet**
  - **SensEmbed:** lcl.uniroma1.it/sensembed
  - **NASARI:** lcl.uniroma1.it/nasari
- Monolingual vs. multilingual:
  - **Monolingual** (but no limit to which language can be used: SensEmbed, NASARI lexical/embedded)
  - **Inherently multilingual** (NASARI unified vector)
- Explicit vs. latent:
  - Explicit vectors provide **human-readable** components (NASARI lexical and unified)
  - Latent vectors are **more compact**, less sparse and **faster to process** (SensEmbed, NASARI embedded)
- Enable semantic, "translatable" output
- Move from **one language to another** seamlessly

Thanks or…

merci

(grazie)

MultiJEDI (Starting Grant, 2011-2016) + **MOUSSE** (Consolidator Grant, 2017-2022)

Roberto Navigli

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