# Automatic Construction and Evaluation of a Large Semantically Enriched Wikipedia

Alessandro Raganato, Claudio Delli Bovi and Roberto Navigli







# WikipediA



# Problem: Wikipedia is designed for humans!

Only a fraction of linkable mentions is in fact hyperlinked:

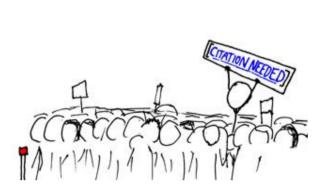
580M noun tokens, only 116M covered (~19%)

Partly due to the Wikipedia style guidelines:

- Link each concept at most once within a page
- Link only when relevant and helpful in the context







From Wikipedia, the free encyclopedia

For other uses, see Lorenzo de' Medici (disambiguation).

Lorenzo de' Medici (Italian pronunciation: [lo'rɛntso de 'mɛ:ditʃi], 1 January 1449 – 9 April 1492) was an Italian statesman and *de facto* ruler of the Florentine Republic, who was one of the most powerful and enthusiastic patrons of the Renaissance. [1][2][3] Also known as Lorenzo the Magnificent (Lorenzo il Magnifico [lo'rɛntso il maŋ'pi:fiko]) by contemporary Florentines, he was a magnate, diplomat, politician and patron of scholars, artists and poets. He is well known for his contribution to the art world by sponsoring artists such as Botticelli and Michelangelo. His life coincided with the mature phase of Italian Renaissance and his death coincided with the end of the Golden Age of Florence. [4] The fragile peace that he helped maintain among the various Italian states collapsed with his death. He is buried in the Medici Chapel in Florence.

#### Contents [hide]

- 1 Youth
- 2 Politics
- 3 Patronage
- 4 Marriage and children
- 5 Later years
- 6 In popular culture
- 7 References
- 8 Further reading
- 9 External links

#### Youth [edit]

Lorenzo's grandfather, Cosimo de' Medici, was the first member of the Medici family to combine running the Medici Bank with leading the Republic of Florence. Cosimo was one of the wealthiest men in Europe and spent a very large portion of his fortune in government and philanthropy. He was a patron of the arts and funded public works. [5] Lorenzo's father, Piero di Cosimo de' Medici, was equally at the centre of Florentine life, chiefly as an art patron and collector, while Lorenzo's uncle, Giovanni di Cosimo de' Medici, took care of the family's business interests. Lorenzo's mother, Lucrezia Tornabuoni, was a writer of sonnets and a friend to poets and philosophers of the Medici Academy. She became her son's advisor after the deaths of his father and uncle. [5]

Lorenzo, considered the brightest of the five children of Piero and Lucrezia, was tutored by a diplomat and bishop, Gentile de' Becchi, and the humanist philosopher Marsilio Ficino, [6] and he was trained in Greek by John Argyropoulos. [7] With his brother Giuliano, he participated in jousting, hawking, hunting, and horse breeding for the Palio, a horse race in Siena. At 19 he won first prize in a tournament of the sons of the leading families of Florence against the best men at arms of Italy; Niccolò Machiavelli would write that he did so "not by way of favour, but by his own valour and skill in arms". [8] His own horse was named Morello di Vento. [9]

#### Lorenzo de' Medici



Portrait by Verrocchio

Lord of Florence

Reign 2 December 1469 - 9 April 1492

Predecessor Piero the Gouty

Successor Piero the Unfortunate

Spouse(s) Clarice Orsini

Issue

Lucrezia de' Medici

Piero de' Medici

Maddalena de' Medici

Contessina Beatrice de' Medici

Giovanni de' Medici, Pope Leo X

Luisa de' Medici

Contessina de' Medici

Giuliano de' Medici. Duke of Nemours

Full name

From Wikipedia, the free encyclopedia

For other uses, see Lorenzo de' Medici (disambiguation).

Lorenzo de' Medici (Italian pronunciation: [lo'rɛntso de 'mɛ:ditʃi], 1 January 1449 – 9 April 1492) was an Italian statesman and de facto ruler of the Florentine Republic, who was one of the most powerful and enthusiastic patrons of the Renaissance. [1][2][3] Also known as Lorenzo the Magnificent (Lorenzo il Magnifico [lo'rɛntso il maŋ'pi:fiko]) by contemporary Florentines, he was a magnate, diplomat, politician and patron of scholars, artists and poets. He is well known for his contribution to the art world by sponsoring artists such as Botticelli and Michelangelo. His life coincided with the mature phase of Italian Renaissance and his death coincided with the end of the Golden Age of Florence. [4] The fragile peace that he helped maintain among the various Italian states collapsed with his death. He is buried in the Medici Chapel in Florence.

potentially linkable

mentions!

#### Contents [hide]

- 1 Youth
- 2 Politics
- 3 Patronage
- 4 Marriage and children
- 5 Later years
- 6 In popular culture
- 7 References
- 8 Further reading
- 9 External links

#### Youth [edit]

Lorenzo's grandfather, Cosimo de' Medici, was the first member of the Medici family to combine running the Medici Bank with leading the Republic of Florence. Cosimo was one of the wealthiest men in Europe and spent a very large portion of his fortune in government and philanthropy. He was a patrol of the arts and funded public works. [5] Lorenzo's father, Piero di Cosimo de' Medici, was equally at the centre of Florentine life, chiefly as an art patron and collector, while Lorenzo's uncle, Giovanni di Cosimo de' Medici, took care of the family's business interests. Lorenzo's mother, Lucrezia Tolnabuoni, was a writer of sonnets and a friend to poets and philosophers of the Medici Academy. She became her son's advisor after the deaths of his father and uncle. [5]

Lorenzo, considered the brightest of the five children of Piero and Lucrezia, was tutored by a diplomat and bishop, Gentile de' Becchi, and the humanist philosopher Marsilio Ficino, [6] and he was trained in Greek by John Argyropoulos. [7] With his brother Giuliano, he participated in jousting, hawking, hunting, and horse breeding for the Palio, a horse race in Siena. At 19 he won first prize in a tournament of the sons of the leading families of Florence against the best men at arms of Italy; Niccolò Machiavelli would write that he did so "not by way of favour, but by his own valour and skill in arms". [8] His own horse was named Morello di Vento. [9]

#### Lorenzo de' Medici



Portrait by Verrocchio

Lord of Florence

**Reign** 2 December 1469 – 9 April 1492

Predecessor Piero the Gouty

Successor Piero the Unfortunate

Spouse(s) Clarice Orsini

Issue

Lucrezia de' Medici

Piero de' Medici

Maddalena de' Medici

Contessina Beatrice de' Medici

Giovanni de' Medici, Pope Leo X

Luisa de' Medici

Contessina de' Medici

Giuliano de' Medici. Duke of Nemours

Full name

## Our goal:

Augment Wikipedia with as much semantic information as possible

## Our goal:

Augment Wikipedia with as much semantic information as possible

### How?

- The existing Wikipedia hyperlink structure:
  - Page-to-page direct connections (Wikilinks)
  - Connections between pages and Wikipedia categories







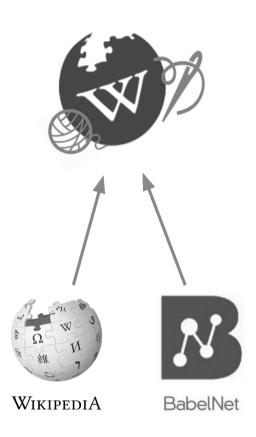
WikipediA

## Our goal:

Augment Wikipedia with as much semantic information as possible

### How?

- The existing Wikipedia hyperlink structure:
  - Page-to-page direct connections (Wikilinks)
  - Connections between pages and Wikipedia categories
- The multilingual sense inventory of BabelNet:
  - merger of many different resources (including WordNet and Wikipedia itself)
  - 14M entries and 380M semantic connections



# W 7

## The algorithm:

For each Wikipedia page **p** in **W** 



**Preprocessing** 



**Hyperlink Propagation** 



Refinement

# W 7

## The algorithm:

For each Wikipedia page **p** in **W** 



**Preprocessing** 



**Hyperlink Propagation** 

- Tokenization
- Part-of-speech tagging
- Lemmatization
- Filtering of uninformative pages



Refinement

# W 7

## The algorithm:

For each Wikipedia page p in W



**Preprocessing** 



**Hyperlink Propagation** 



Refinement

Cascade of propagation heuristics, which

- collect a list S<sup>p</sup> of hyperlinks to be propagated across p
- scan the text of p to match any potential lexicalization (one-sense-per-page assumption)

# W 7

## The algorithm:

For each Wikipedia page **p** in **W** 



**Preprocessing** 



**Hyperlink Propagation** 



Refinement

Conservative policy to remove duplicates and overlapping mentions

Hyperlink Propagation - A bird's eye view



# Hyperlink Propagation - A bird's eye view

### **Intra-page heuristics:**

Propagate links that occur as mentions within **p** 



- Surface MentionPropagation
- Lemmatized MentionPropagation
- Person MentionPropagation



# Hyperlink Propagation - A bird's eye view

### **Intra-page heuristics:**

Propagate links that occur as mentions within **p** 

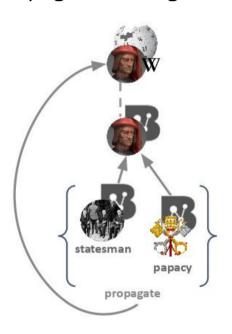


- Surface MentionPropagation
- Lemmatized Mention Propagation
- Person MentionPropagation

### **Inter-page heuristics:**

Exploit the connection of **p** with other pages or categories

- Wikipedia InlinkPropagation
- BabelNet InlinkPropagation
- CategoryPropagation
- Monosemous Content Word



From Wikipedia, the free encyclopedia

For other uses, see Lorenzo de' Medici (disambiguation).

Lorenzo de' Medici (1 January 1449 – 9 April 1492) was an Italian statesman and *de facto* ruler of the Florentine Republic during the Italian Renaissance, [1] Known as Lorenzo the Magnificent (*Lorenzo il Magnifico*) by contemporary Florentines, he was a magnate, diplomat, politician and patron of scholars, artists, and poets. He is perhap best known for his contribution to the art world, sponsoring artists such as Botticelli and Michelangelo. His life coincided with the mature phase of Italian Renaissance and his death coincided with the end of the Golden Age of Florence. [2] The fragile peace he helped maintain between the various Italian states collapsed with his death. Lorenzo de' Medici is buried in the Medici Chapel in Florence.

#### Contents [hide]

- 1 Youth
- 2 Politics
- 3 Patronage
- 4 Marriage and children
- 5 Later years
- 6 In popular culture
- 7 References
- 8 Further reading
- 9 External links

## **30** Original links

#### Youth [edit]

Lorenzo's grandfather, Cosimo de' Medici, was the first member of the Medici family to combine running the Medici Bank with leading the Republic of Florence. Cosimo was one of the wealthiest men in Europe and spent a very large portion of his fortune in government and philanthropy. He was a patron of the arts and funded public works. [3] Lorenzo's father, Piero di Cosimo de' Medici, was also at the center of Florentine life, active chiefly as an art patron and collector, while .orenzo's grandfather and uncle, Giovanni di Cosimo de' Medici took care of the family's business interests. Lorenzo's mother Lucrezia Tornabuoni was a poet and writer of sonnets and a friend to poets and philosophers of the Medici Academy. She became her son's advisor after the deaths of his father and uncle. [3]

Lorenzo, considered the brightest of the five children of Piero and Lucrezia, was tutored by a diplomat and bishop, Gentile de' Becchi and the humanist philosopher Marsilio Ficino.<sup>[4]</sup> With his brother Giuliano, he participated in jousting, hawking, hunting, and horse breeding for the Palio, a horse race in Siena. His own horse was named Morello di Vento.<sup>[5]</sup>

# Lorenzo de' Medici

Portrait by Verrocchio

Lord of Florence

Reign 2 December 1469 – 9 April 1492

Predecessor Piero the Gouty

Successor Piero the Unfortunate

Spouse(s) Clarice Orsini

Issue

Lucrezia de' Medici

Piero de' Medici

Maddalena de' Medici

Contessina Beatrice de' Medici

Giovanni de' Medici, Pope Leo X

Luisa de' Medici

Contessina de' Medici

Giuliano de' Medici. Duke of Nemours

Full name

From Wikipedia, the free encyclopedia

For other uses, see Lorenzo de' Medici (disambiguation).

Lorenzo de' Medici (1 January 1449 - 9 April 1492) was an Italian statesman and de facto ruler of the Florentine Republic during the Italian Renaissance. [1] Known as Lorenzo the Magnificent (Lorenzo il Magnifico) by contemporary Florentines, he was a magnate, diplomat, politician and patron of scholars, artists, and poets. He is perhap best known for his contribution to the art world, sponsoring artists such as Botticelli and Michelangelo, His life coincided with the mature phase of Italian Renaissance and his death coincided with the end of the Golden Age of Florence, [2] The fragile peace he helped maintain between the various Italian states collapsed with his death. Lorenzo de' Medici is buried in the Medici Chapel in Florence,

#### Contents [hide]

- 1 Youth
- 2 Politics 3 Patronage
- 4 Marriage and children
- 5 Later years
- 6 In popular culture
- 7 References
- 8 Further reading
- 9 External links

# **30** Original links

21 new intra-page links  $Q_w$ 



#### Youth [edit]

Lorenzo's grandfather, Cosimo de' Medici, was the first member of the Medici family to combine running the Medici Bank with leading the Republic of Florence: Cosimo was one of the wealthiest men in Europe and spent a very large portion of his fortune in government and philanthropy. He was a patron of the arts and funded public works. [3] Lorenzo's father, Piero di Cosimo de' Medici, was also at the center of Florentine life, active chiefly as an art patron and collector, while Lorenzo's grandfather and uncle, Giovanni di Cosimo de' Medici took care of the family's business interests. Lorenzo's mother Lucrezia Tornabuoni was a poet and writer of sonnets and a friend to poets and philosophers of the Medici Academy. She became her son's advisor after the deaths of his father and uncle.[3]

Lorenzo, considered the brightest of the five children of Piero and Lucrezia, was tutored by a diplomat and bishop, Gentile de' Becchi and the humanist philosopher Marsilio Ficino. [4] With his brother Giuliano, he participated in jousting, hawking, hunting, and horse breeding for the Palio, a horse race in Siena. His own horse was named Morello di Vento.[5]

#### Lorenzo de' Medici



Portrait by Verrocchio

#### Lord of Florence

2 December 1469 - 9 April 1492 Reign

Predecessor Piero the Gouty

Successor Piero the Unfortunate

Clarice Orsini Spouse(s)

#### Issue

Lucrezia de' Medici

Piero de' Medici

Maddalena de' Medici

Contessina Beatrice de' Medici

Giovanni de' Medici, Pope Leo X

Luisa de' Medici

Contessina de' Medici

Giuliano de' Medici. Duke of Nemours

#### Full name

From Wikipedia, the free encyclopedia

For other uses, see Lorenzo de' Medici (disambiguation).

Lorenzo de' Medici (1 January 1449 - 9 April 1492) was an Italian statesman and de facto ruler of the Florentine Republic during the Italian Renaissance. [1] Known as Lorenzo the Magnificent (Lorenzo il Magnifico) by contemporary Florentines, he was a magnate, diplomat, politician and patron of scholars, artists, and poets. He is perhaps best known for his contribution to the art world, sponsoring artists such as Botticelli and Michelangelo, His life coincided with the mature phase of Italian Renaissance and his death coincided with the end of the Golden Age of Florence, [2] The fragile peace he helped maintain between the various Italian states collapsed with his death. Lorenzo de' Medici is buried in the Medici Chapel in Florence.

#### Contents [hide]

- 1 Youth
- 2 Politics
- 3 Patronage
- 4 Marriage and children
- 5 Later years
- 6 In popular culture
- 7 References
- 8 Further reading
- 9 External links

**30** Original links

21 new intra-page links &

31 new inter-page links **Q**<sub>xy</sub>



#### Youth [edit]

Lorenzo's grandfather, Cosimo de' Medici, was the first member of the Medici family to combine running the Medici Bank with leading the Republic of Florence: Cosimo was one of the wealthiest men in Europe and spent a very large portion of his fortune in government and philanthropy. He was a patron of the arts and funded public works. [3] Lorenzo's father, Piero di Cosimo de' Medici, was also at the center of Florentine life, active chiefly as an art patron and collector, while Lorenzo's grandfather and uncle, Giovanni di Cosimo de' Medici took care of the family's business interests. Lorenzo's mother Lucrezia Tornabuoni was a poet and writer of sonnets and a friend to poets and philosophers of the Medici Academy. She became her son's advisor after the deaths of his father and uncle.[3]

Lorenzo, considered the brightest of the five children of Piero and Lucrezia, was tutored by a diplomat and bishop, Gentile de' Becchi and the humanist philosopher Marsilio Ficino. [4] With his brother Giuliano, he participated in jousting, hawking, hunting, and horse breeding for the Palio, a horse race in Siena. His own horse was named Morello di Vento.[5]

## Lorenzo de' Medici



Portrait by Verrocchio

Lord of Florence

2 December 1469 - 9 April 1492 Reign

Predecessor Piero the Gouty

Piero the Unfortunate

Clarice Orsini Spouse(s)

Issue

Lucrezia de' Medici

Piero de' Medici

Maddalena de' Medici

Contessina Beatrice de' Medici

Giovanni de' Medici, Pope Leo X

Luisa de' Medici

Contessina de' Medici

Giuliano de' Medici. Duke of Nemours

Full name



Statistics - Another bird's eye view

# Statistics - Another bird's eye view

#### Total number of sense annotations

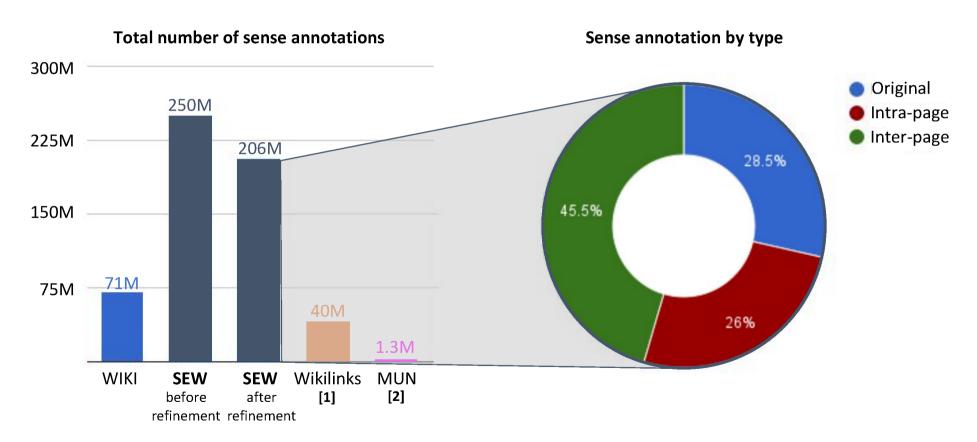


[1] S. Singh, A. Subramanya, F. Pereira, and A. McCallum. *Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia*. Technical Report UM-CS-2012-015.

[2] K. Taghipour and H. Ng. One million sense-tagged instances for word sense disambiguation and induction. CoNLL, 2015.



# Statistics - Another bird's eye view







## **Experiments**

Intrinsic Evaluation - Annotation Quality

We compared our sense annotations against those discovered by **3W** (Noraset et al. 2014), a Wikipedia-specific system designed to add automatically high-precision hyperlinks





## **Experiments**

Intrinsic Evaluation - Annotation Quality

We compared our sense annotations against those discovered by **3W** (Noraset et al. 2014), a Wikipedia-specific system designed to add automatically high-precision hyperlinks



 Extrinsic Evaluation - Entity Linking and Semantic Similarity

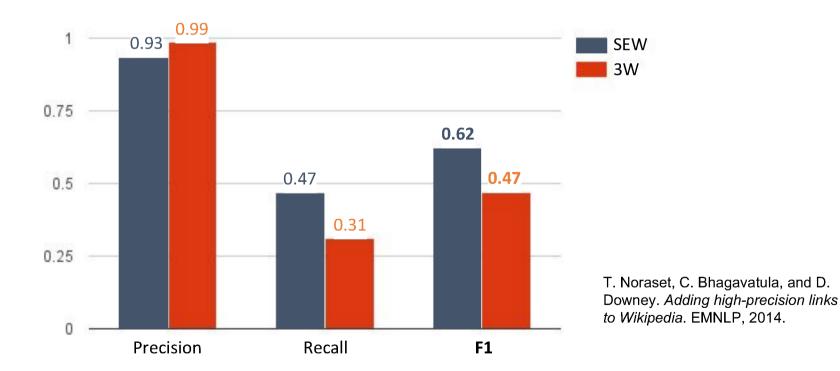
We used **SEW** both as **training set** for Entity Linking, and as **semantic network** to develop Wikipedia-based vector representations for Semantic Similarity





## **Intrinsic Evaluation**

Hand-labeled evaluation set of **2000** Wikipedia pages (Noraset et al. 2014):





## **Extrinsic Evaluation #1 - Entity Linking**

### Benchmark system:

**IMS** (Zhong and Ng 2010), a state-of-the-art supervised system for Word Sense Disambiguation in English based on SVMs

• IMS + SEW IMS trained on SEW

• IMS + HL IMS trained only on the original hyperlinks (baseline #1)

• MFS Most Frequent Sense provided by BabelNet (baseline #2)

#### Datasets:

- SemEval 2013, task 12
- SemEval 2015, task 13

- . MSNBC
- . AIDA-CoNLL



# **Extrinsic Evaluation #1 - Entity Linking**

	SemEval-2013	SemEval-2015	MSNBC	AIDA-CoNLL
IMS+SEW	0.810	0.882	0.789	0.726
IMS+HL	0.775	0.758	0.695	0.712
MFS	0.802	0.857	0.620	0.535
UMCC-DLSI	0.548	-	-	- ·
Babelfy	0.874	-	·-	0.821
DFKI	-	0.889	-	-
SUDOKU	-	0.870	-	-
Wikifier	-	-	0.812	0.724
M&W	-	-	0.685	0.823



**Extrinsic Evaluation #2 - Semantic Similarity** 



# **Extrinsic Evaluation #2 - Semantic Similarity**

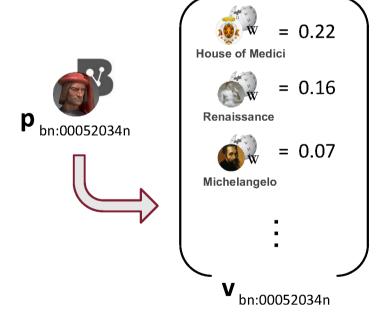
We used **SEW** to build Wikipedia-based vector representations for **BabelNet senses**:



## **Extrinsic Evaluation #2 - Semantic Similarity**

We used **SEW** to build Wikipedia-based vector representations for **BabelNet senses**:

 WB-SEW: associates a vector v with each sense in BabelNet, defined in the space of Wikipedia pages;

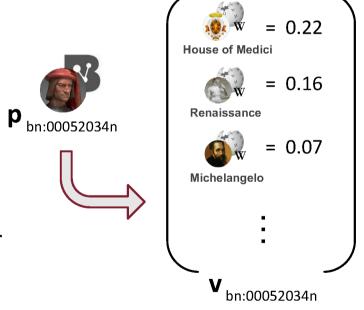




## **Extrinsic Evaluation #2 - Semantic Similarity**

We used **SEW** to build Wikipedia-based vector representations for **BabelNet senses**:

- WB-SEW: associates a vector v with each sense in BabelNet, defined in the space of Wikipedia pages;
- Each component of v is an estimated frequency based on either raw count (RC) or lexical specificity (LS)





# **Extrinsic Evaluation #2 - Semantic Similarity**

**Pearson** (r) and **Spearman** ( $\rho$ ) correlation in the word similarity task:

		WB-SEW		WB-HL		
		RC	LS	RC	LS	Baseline: only original
WS-Sim	r	0.65	0.64	0.58	0.58	hyperlinks (HL)
	$\rho$	0.69	0.70	0.59	0.61	
SimLex-666	r	0.38	0.38	0.32	0.32	
	$\rho$	0.40	0.41	0.31	0.32	



## **Extrinsic Evaluation #2 - Semantic Similarity**

**Pearson** (r) and **Spearman** ( $\rho$ ) correlation in the word similarity task:

		WB-SEW		WB-HL		
		RC	LS	RC	LS	Baseline: only original
WS-Sim	r	0.65	0.64	0.58	0.58	hyperlinks (HL)
	$\rho$	0.69	0.70	0.59	0.61	
SimLex-666	r	0.38	0.38	0.32	0.32	
	ho	0.40	0.41	0.31	0.32	

**ADW** (Pilehvar et al. 2013) on WS-Sim:

r = 0.63  $\rho = 0.69$ 

**ESA** (Gabrilovich and Markovitch, 2007) on WS-Sim: r = 0.40  $\rho = 0.47$ 



**Extrinsic Evaluation #2 - Semantic Similarity (cont.)** 



## **Extrinsic Evaluation #2 - Semantic Similarity (cont.)**

**Pearson** (r) and **Spearman** ( $\rho$ ) for multilingual word similarity on RG-65:

		WB-SEW		WB-HL		Word2Vec		Polyglot
		RC	LS	RC	LS	original	retro	
EN	r	0.673	0.674	0.619	0.614	-	-	0.51
<b>L</b> AT	$\rho$	0.608	0.620	0.592	0.592	0.73	0.77	0.55
FR	r	0.808	0.811	0.773	0.778	-	-	0.38
	$\rho$	0.755	0.759	0.693	0.681	0.47	0.61	0.35
DE	r	0.639	0.639	0.584	0.580	-	-	0.18
DE	$\rho$	0.689	0.695	0.637	0.615	0.53	0.60	0.15
ES	r	0.811	0.804	0.757	0.740	_	-	0.51
	$\rho$	0.815	0.812	0.764	0.759	-	-	0.56



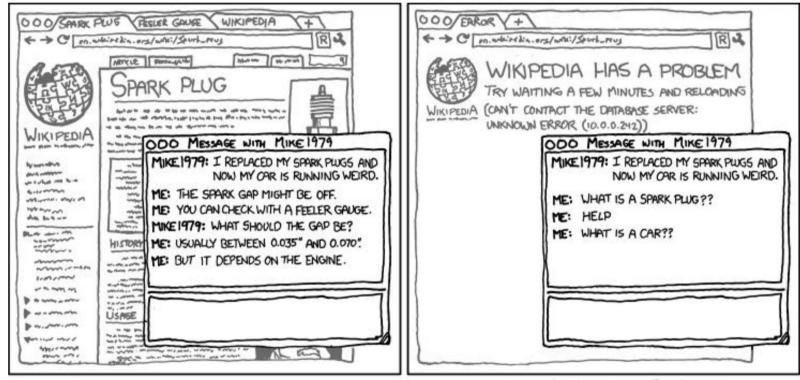
## **Conclusion**

- **SEW**: A Semantically Enriched Wikipedia with more than **200M** annotations of over **4M** different word senses and named entities from the very same, wide-coverage sense inventory of BabelNet
- The cascade of propagation heuristics used for SEW is surprisingly simple, fully automatic and self contained (no need for training, validation or tuning any off-the-shelf system)
- Our extensive evaluation, beside proving the quality of SEW's annotations, sets important performance baselines for multiple tasks and datasets

# Thank you!

(and see you at the poster session this evening)





WHEN WIKIPEDIA HAS A SERVER OUTAGE, MY APPARENT IQ DROPS BY ABOUT 30 POINTS.



http://lcl.uniroma1.it/sew