SEW: Automatic Construction and Evaluation of a Large Semantically Enriched Wikipedia

http://lcl.uniroma1.it/sew

Alessandro Raganato
raganato@di.uniroma1.it bn:17381127n

Claudio Delli Bovi
dellibovi@di.uniroma1.it bn:17381128n

Roberto Navigli
navigli@di.uniroma1.it bn:09353187n

What is it?

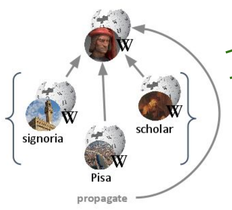
- **SEW** (**S**emantically **E**nriched **W**ikipedia) is a sense-annotated corpus automatically built from **Wikipedia** by exploiting its hyperlink structure along with the wide-coverage sense inventory of **BabelNet**.
- **SEW** constitutes both a large-scale Wikipedia-based semantic network and a sense-tagged dataset with more than **200 million** annotations of over **4 million** different concepts and named entities.

Preprocessing

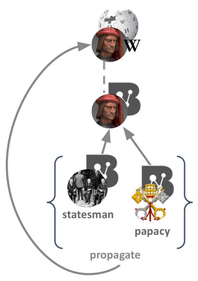
- **Tokenization**
- **Part-of-speech tagging**
- **Lemmatization**
- **Filtering of uninformative pages** (e.g. '`List of`', disambiguation pages, common surnames)

Inter-page Hyperlink Propagation

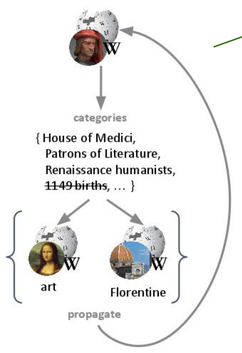
- Wikipedia Inlink Propagation (WIL)



- BabelNet Inlink Propagation (BIL)



- Category Propagation (CP)

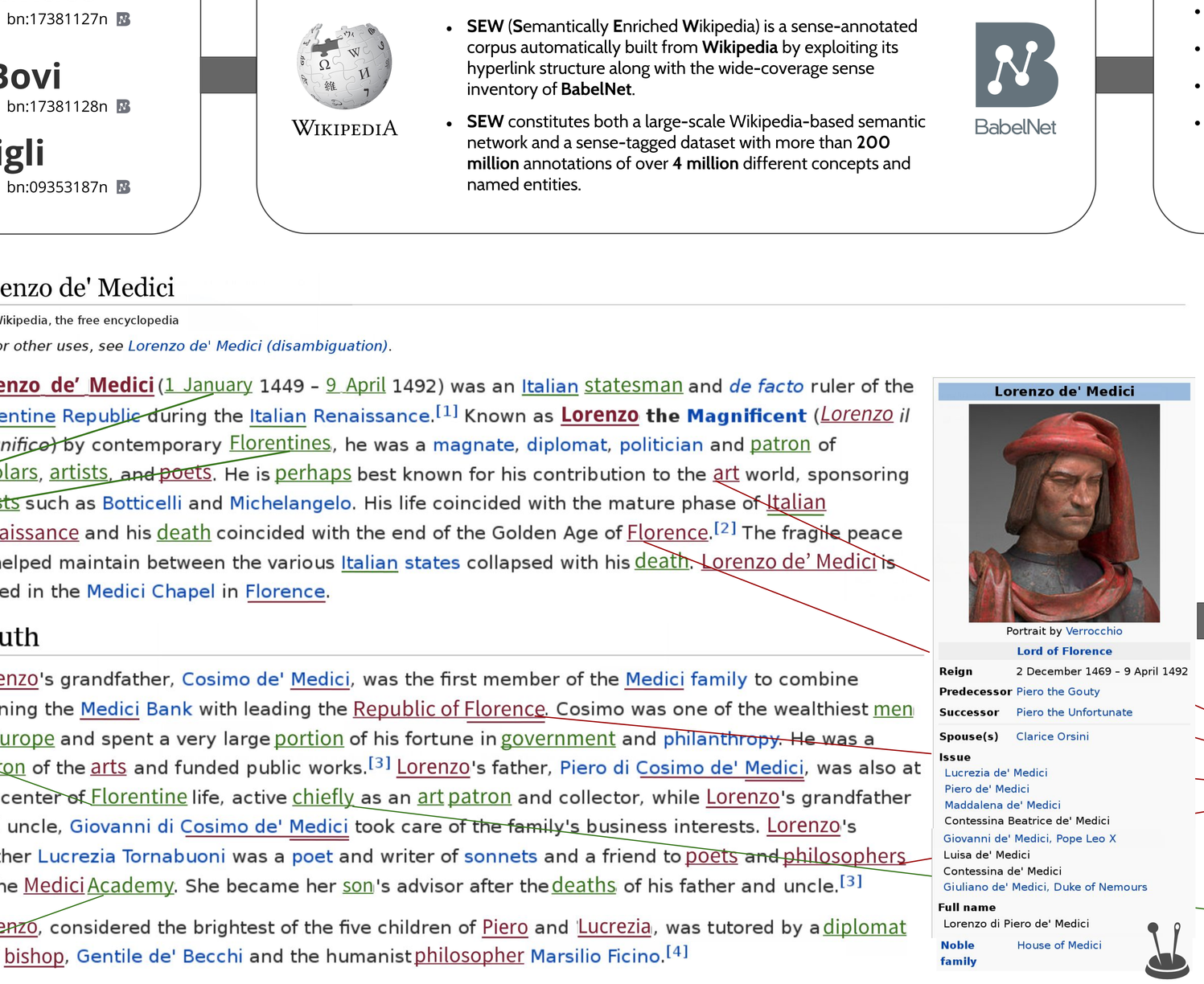


Category entropy:

$$H(c) = -\sum_{h \in S^c} f^c(h) \log_2 f^c(h)$$

Set of hyperlinks propagated through category **c**

Normalized frequency counts of hyperlinks in S^c



Lorenzo de' Medici

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.

Youth

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]

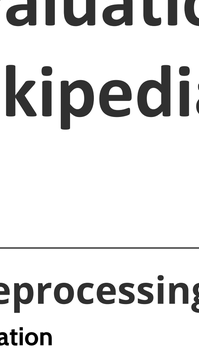
Lorenzo, considered the brightest of the five children of Piero and Lucrezia, was tutored by a diplomat and bishop, Gentile de' Becchi and the humanist philosopher Marsilio Ficino.[4]

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	Lorenzo di Piero de' Medici
Noble family	House of Medici

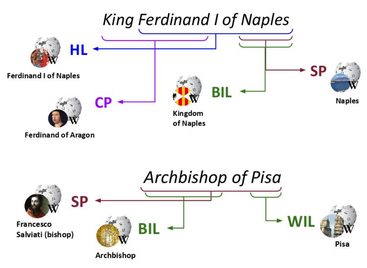
Intra-page Hyperlink Propagation



- Surface Mention Propagation (SP)
- Lemmatized Mention Propagation (LP)
- Person Mention Propagation (PP)
- Monosemous Content Word (MP)

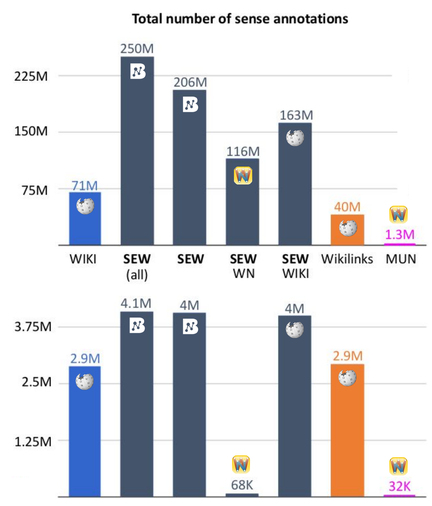
Refinement

Conservative policy to deal with **overlapping mentions** and **duplicates**:



1. Prefer **intra-page** annotations over **inter-page** ones
2. if the mention is still ambiguous after 1. prefer the **longest match**
3. if the mention is still ambiguous after 1. and 2. **remove all annotations**

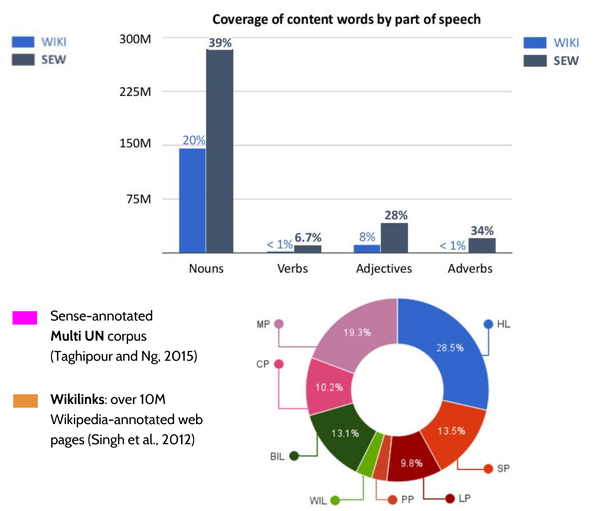
Statistics



Total number of distinct sense covered

Sense-annotated **Multi UN** corpus (Taghipour and Ng, 2015)

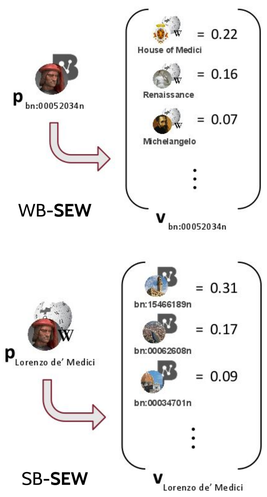
Wikilinks: over 10M Wikipedia-annotated web pages (Singh et al., 2012)



Sense annotation by type

Evaluation #2

Using **SEW** to build vector representations for **BabelNet senses** and **Wikipedia pages**:



Pearson (**r**) and Spearman (**ρ**) correlation figures in the **word similarity** task:

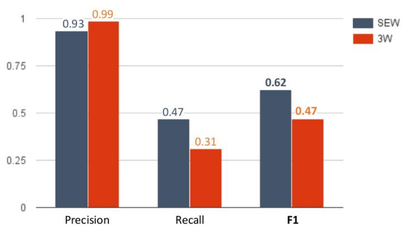
		WB-SEW RC	WB-SEW LS	SB-SEW RC	SB-SEW LS	WB-HL RC	WB-HL LS	SB-HL RC	SB-HL LS
WS-Sim	r	**0.65**	0.64	0.50	0.57	0.58	0.58	0.53	0.52
	ρ	0.69	**0.70**	0.56	0.57	0.59	0.61	0.49	0.51
SimLex-666	r	**0.38**	**0.38**	0.26	0.34	0.32	0.32	0.28	0.31
	ρ	0.40	**0.41**	0.33	0.36	0.31	0.32	0.27	0.27

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

		WB-SEW RC	WB-SEW LS	WB-HL RC	WB-HL LS	Word2Vec original	Word2Vec retro	Polyglot
EN	r	0.673	**0.674**	0.619	0.614	-	-	0.51
	ρ	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
	ρ	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
	ρ	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
	ρ	**0.815**	0.812	0.764	0.759	-	-	0.56

Evaluation #1

Intrinsic Evaluation - Annotation Quality:
SEW's pipeline on a hand-labeled evaluation set of **2000** Wikipages (Noraset et al. 2014)



Extrinsic Evaluation - Entity Linking:
training **IMS** (Zhong and Ng, 2010), a supervised WSD system, on **SEW**

	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**