Sew-Embed at SemEval-2017 Task 2: Language-Independent Concept Representations from a Semantically Enriched Wikipedia

SAPIENZA UNIVERSITÀ DI ROMA
DIPARTIMENTO DI INFORMATICA

http://lcl.uniroma1.it/sew

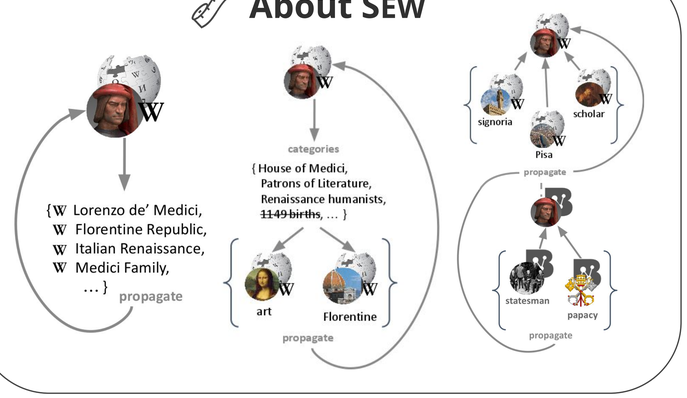
Claudio Delli Bovi
dellibovi@di.uniroma1.it bn:17381128n

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

SEW: What is it?

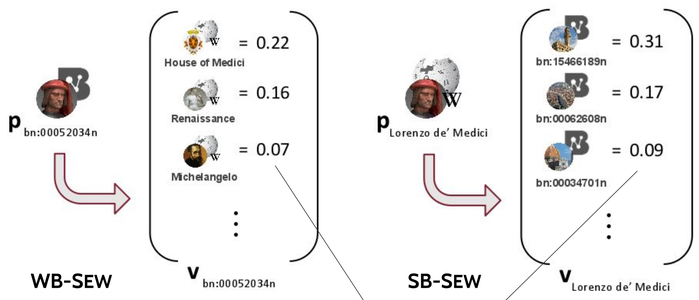
- **SEW** (**S**emantically **E**nriched **W**ikipedia) [6] is a sense-annotated corpus automatically built from **Wikipedia** by exploiting its hyperlink structure along with the wide-coverage sense inventory of **BabelNet** [5].
- **SEW** constitutes both a large-scale Wikipedia-based semantic network and a sense-tagged dataset with more than **200 million** annotations of over **4 million** different concepts and named entities.

About SEW



SEW: Explicit Vector Representations

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



Lexical Specificity (LS) [3]:

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

Word Similarity:

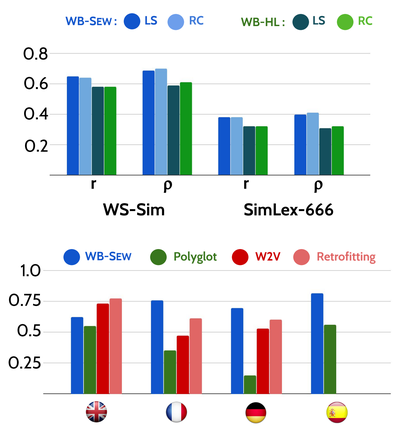
$$\max_{s_1 \in S_{w_1},\, s_2 \in S_{w_2}} \sigma(\vec{s_1}, \vec{s_2})$$

Given two words w_1 and w_2 select the *closest* pair of senses

Weighted Overlap:

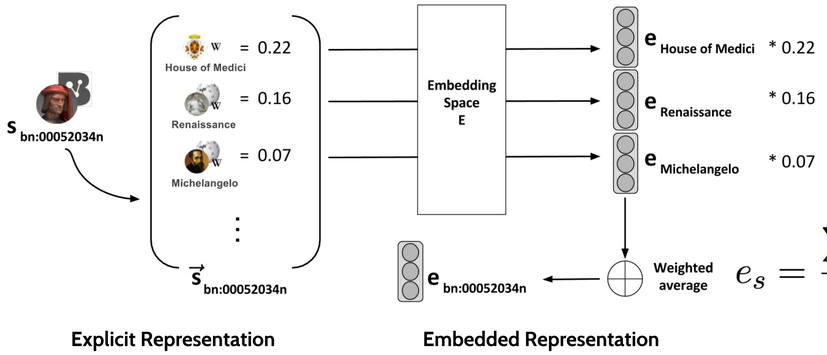
$$\frac{\sum_{q \in O} \left(rank(q, \vec{s_1}) + rank(q, \vec{s_2})\right)^{-1}}{\sum_{i=1}^{|O|} (2i)^{-1}}$$

Multilingual Word Similarity:



RG 65 (ρ) [2]

SEW-EMBED: From Explicit to Embedded Representations



$$e_s = \frac{\sum_{p \in \vec{s}} \omega_p \, e_p}{\sum_{p \in \vec{s}} \omega_p}$$

Explicit Representation — Embedded Representation

If **E** is a *word* embedding space we consider the **Wikipage title** as lexicalization of the corresponding dimension

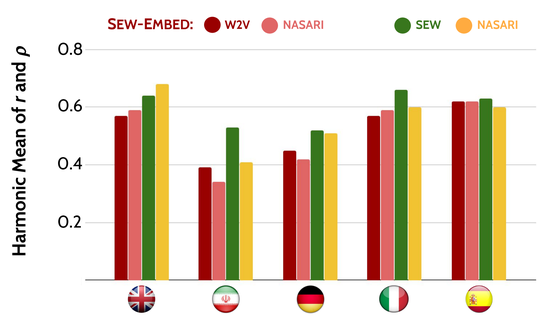
If **E** is a *sense* embedding space we map the Wikipage to the corresponding embedding using **Babelnet's inter-resource mappings**

Embedding spaces:

1. Pre-trained **Google News W2V** [4] embeddings
   ↓
   SEW-EMBED$_{W2V}$
2. Embedded concept vectors of **NASARI** [1]
   ↓
   SEW-EMBED$_{NASARI}$

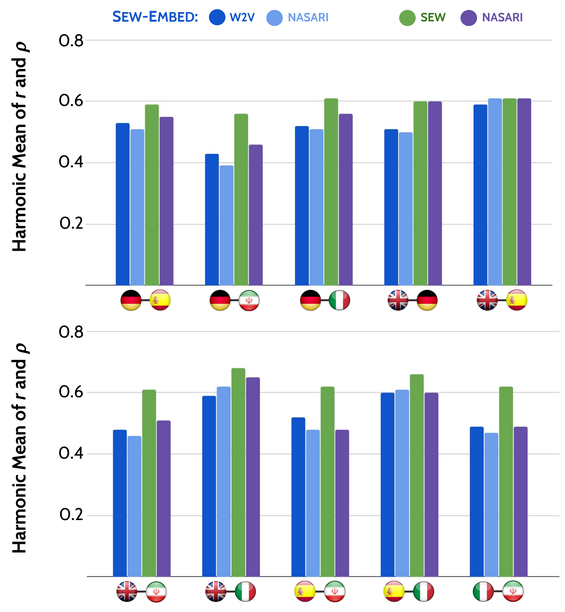
SemEval-2017 Experiments

Subtask 1: Multilingual Word Similarity



Global score: **0.552** (4th position)

Subtask 2: Cross-Lingual Word Similarity



Global score: **0.558** (3rd position)

References

[1] J. Camacho Collados, M. T. Pilehvar, and R. Navigli. 2016. *Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities.* AIJ, 240:36–64.

[2] J. Camacho Collados, M. T. Taher Pilehvar, and R. Navigli. 2015. *A Framework for the Construction of Monolingual and Cross-lingual Word Similarity Datasets.* ACL. pp. 1–7.

[3] P. Lafon. 1980. *Sur la variabilité de la fréquence des formes dans un corpus.* Mots 1(1):127–165.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. *Efficient estimation of word representations in vector space.* ICLR Workshop.

[5] R. Navigli and S. P. Ponzetto. 2012. *BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network.* AIJ, 193:217–250.

[6] A. Raganato, C. Delli Bovi, and R. Navigli. 2016. *Automatic Construction and Evaluation of a Large Semantically Enriched Wikipedia.* IJCAI, pp. 2894–2900.