Nominal Compound Compositionality: A Multilingual Lexicon and Predictive Model

Silvio Cordeiro[1,2] srcordeiro@inf.ufrgs.br
Carlos Ramisch[2] carlos.ramisch@lif.univ-mrs.fr
Aline Villavicencio[1] avillavicencio@inf.ufrgs.br

[1] Federal University of Rio Grande do Sul (Brazil) [2] Aix Marseille Université, CNRS, LIF UMR 7279 (France)

Summary

- We collect human **compositionality judgments** for 180×3 **nominal compounds** in English, French and Portuguese.
  - Example: insurance company $>$ climate change $> \ldots >$ milk tooth $> \ldots >$ nut case $>$ eager beaver $>$ cloud nine.
- We consider a **predictive model** that uses **word embeddings** to predict the compositionality of these nominal compounds.
- We thoroughly **evaluate** several aspects of the model and overcome **state of the art** results in standard datasets.

1. Predictive model

Hypothesis:
A compound w_1w_2 is compositional $\Leftrightarrow$ $\overrightarrow{w_1w_2}$ is similar to $\overrightarrow{w_1} + \overrightarrow{w_2}$

Model:

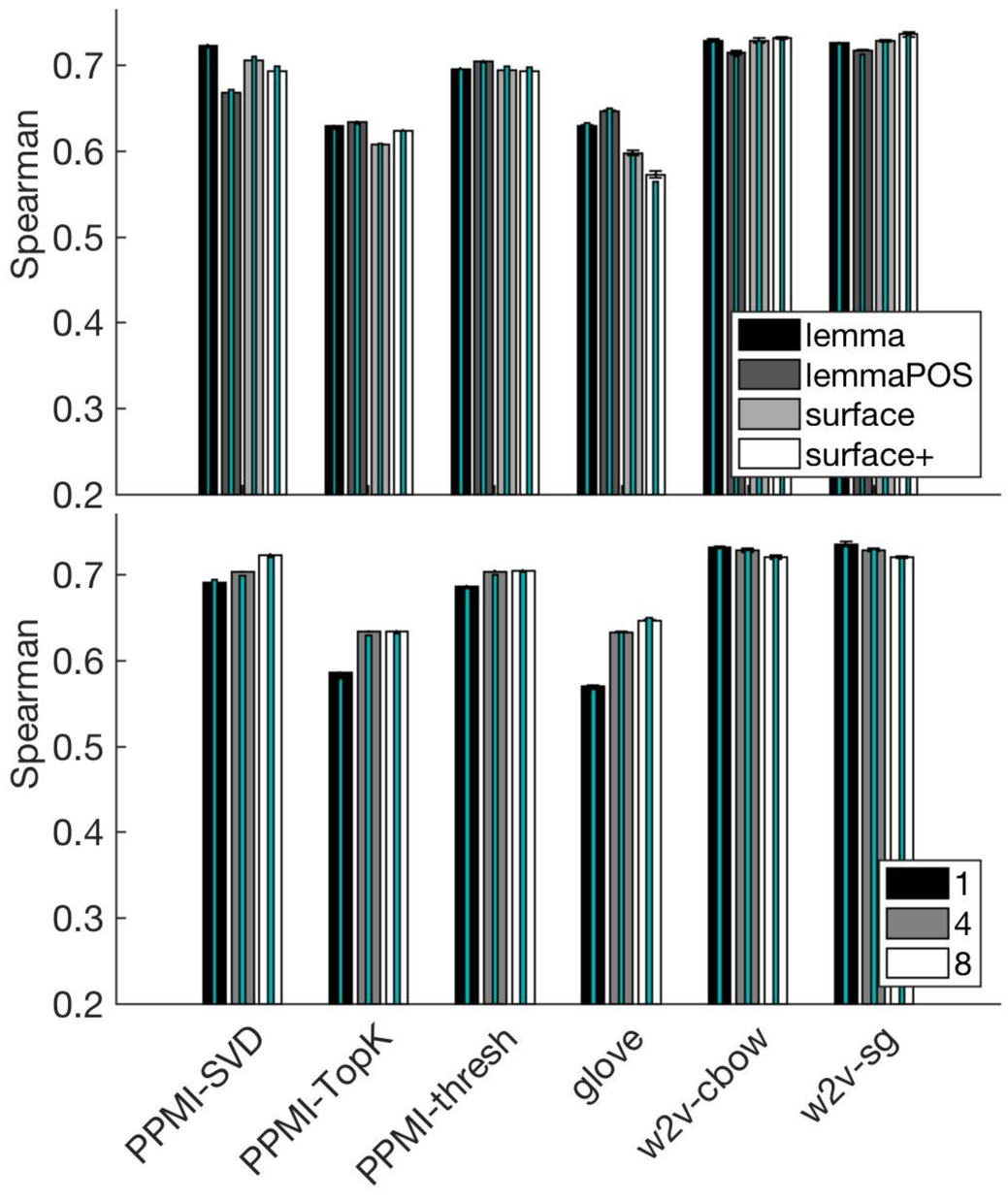
1. Build word embeddings for w_1w_2 and components w_1 and w_2
2. Add vectors $\overrightarrow{w_1}+\overrightarrow{w_2}$, assuming the compound is compositional
3. Predict compositionality score using cosine similarity
4. Calculate Spearman rank correlation between prediction and human judgments

2. Evaluated parameters

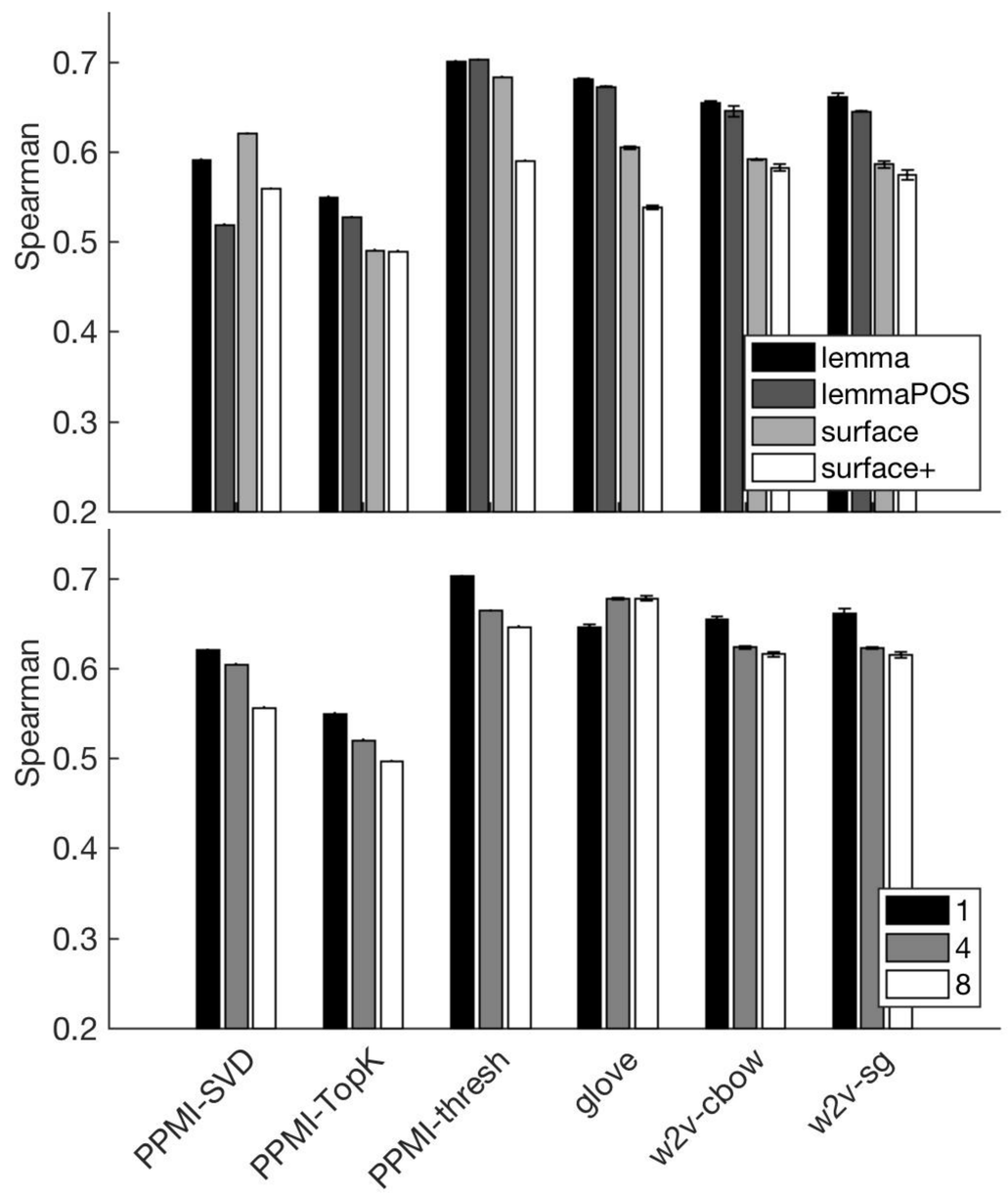
- Languages: English, French
- Models: PPMI matrix, GloVe, word2vec cbow & skipgram
- Preprocessing: surface$+^{\text{stopw}}$, surface, lemmaPOS, lemma
- Context window sizes: 1, 4, 8 words to the left/right
- Dimensions: 250, 500, 750

3. Best correlations for English

- **strict** evaluation (remove missing compounds)
- loose evaluation (use fallback = average of predictions)



4. Best correlations for French



5. State of the art (English only)

- Dataset from Reddy et al [2011]:

Model & Parameters	Spearman ρ
Reddy et al [2011]	.71
Salehi et al [2015]	.80
Best *w2v* (*sg, WF=surface, D=750, W=1*)	**.82** / .80
Best *PPMI* (*thresh, WF=surface, D=750, W=8*)	**.80** / .80

- Dataset from Farahmand et al [2015]:

Model & Parameters	Best-F_1
Yazdani et al [2015]	.49
Best *w2v* (*sg, WF=lemma, D=500, W=1*)	**.51** / .47
Best *PPMI* (*svd, WF=lemma, D=750, W=4*)	**.52** / .45

6. Conclusions

- Lemmas better than surface forms for French
- Small windows better for French, not relevant for English
- More dimensions is consistently better
- Classical PPMI models are comparable to word2vec
- Two papers in ACL 2016: Ramisch et al & Cordeiro et al

Financial support from the projects: AIM-WEST (FAPERGS-INRIA 1706-2551/13-7), CAMELEON (CAPES-COFECUB 707-11), Samsung Eletronica da Amazonia Ltda. Simplificação Textual de Expressões Complexas, Law number 8.248/91, CNPq 482520/2012-4, CNPq 312114/2015-0, PARSEME (COST IC1207), PARSEME-FR (ANR-14-CERA-0001), ORFEO (ANR-12-CORP-0005).