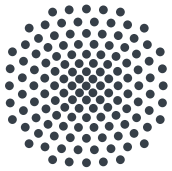


University of Stuttgart
Germany

OP-IMS @ DIACR-Ita: Back to the Roots: SGNS+OP+CD still rocks Semantic Change Detection

December 3, 2020

Jens Kaiser, Dominik Schlechtweg, Sabine Schulte im Walde

Institute for Natural Language Processing, University of Stuttgart, Germany

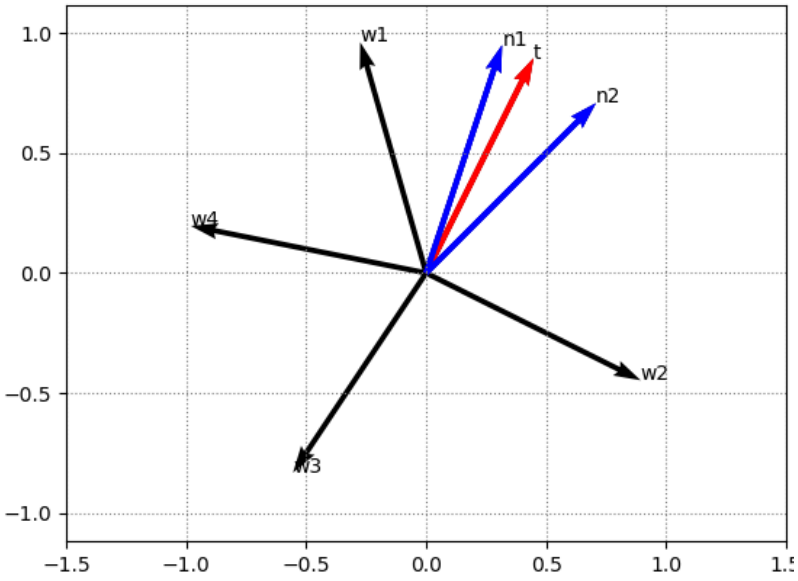
Data and Task

- **given**:
  - set of target words
  - diacronic corpus pair C_1 and C_2
- **task**:
  - decide which words lost or gained sense(s) between C_1 and C_2, and which ones did not (Schlechtweg et al., 2020)
  - dataset contained only words that gained sense(s) (Basile et al., 2020)

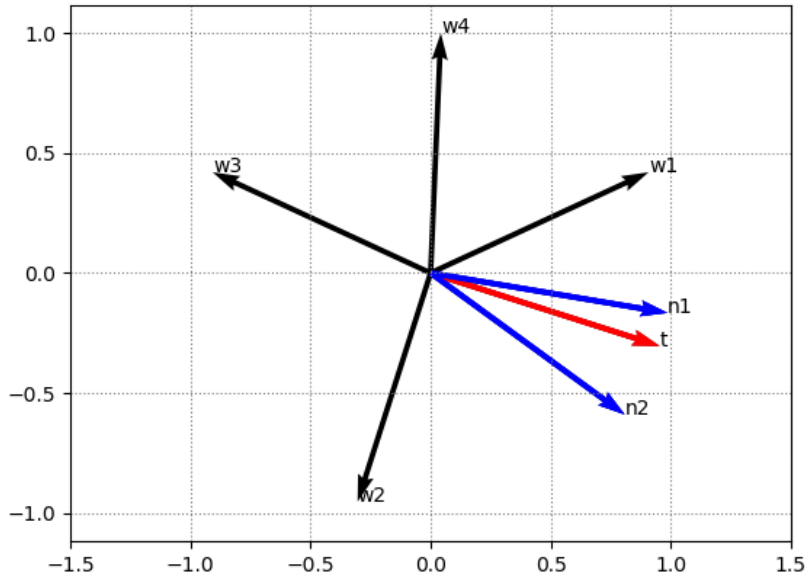
Model

- SGNS+OP+CD (Hamilton, Leskovec, & Jurafsky, 2016)
  1. **Semantic Representation**: Skip-gram with Negative Sampling (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013)
  2. **Alignment**: Orthogonal Procrustes (Schönemann, 1966)
  3. **Change Measure**: Cosine Distance (Salton & McGill, 1983)

SGNS+OP+CD

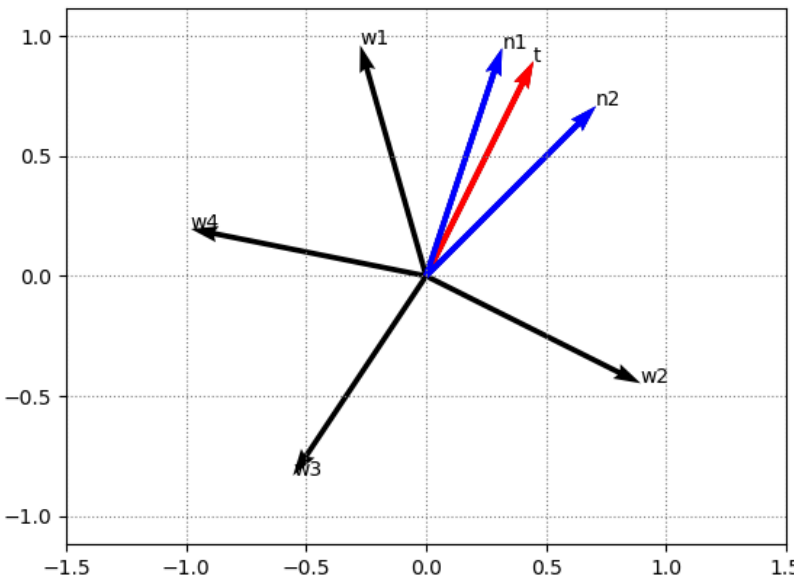


C_1

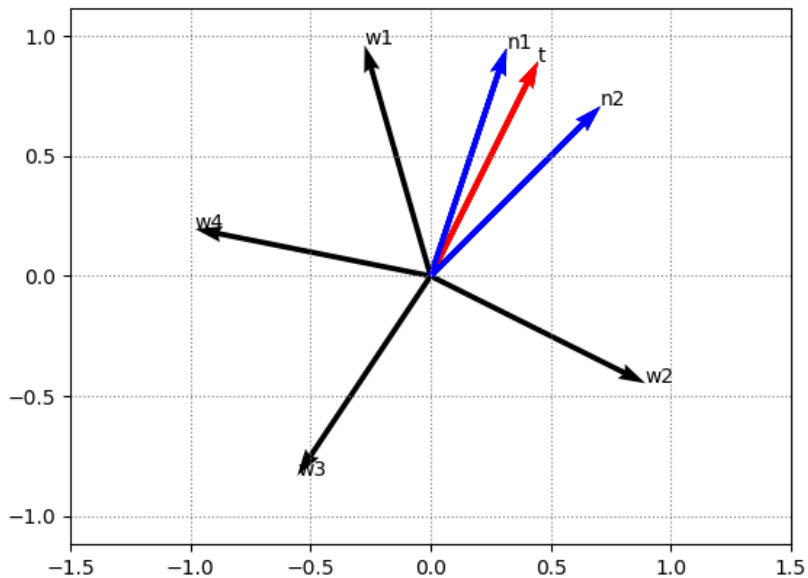


C_2

SGNS+OP+CD



C_1

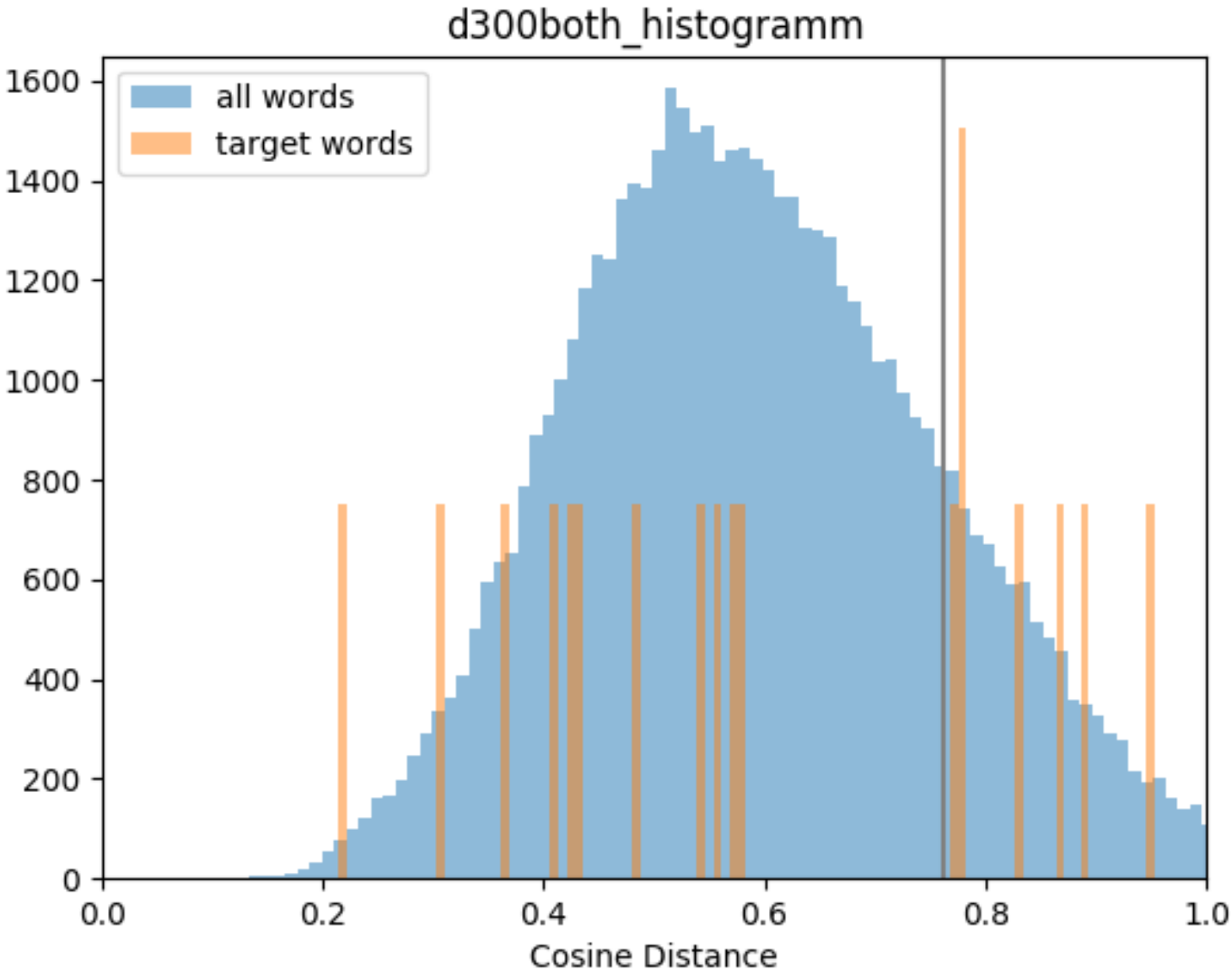


C_2

Why SGNS+OP+CD?

- SemEval 2020: Task 1 (Schlechtweg et al., 2020)
- dominates task (Arefyev & Zhikov, 2020; Pömsl & Lyapin, 2020)
- surprisingly robust (Kaiser et al., 2020)

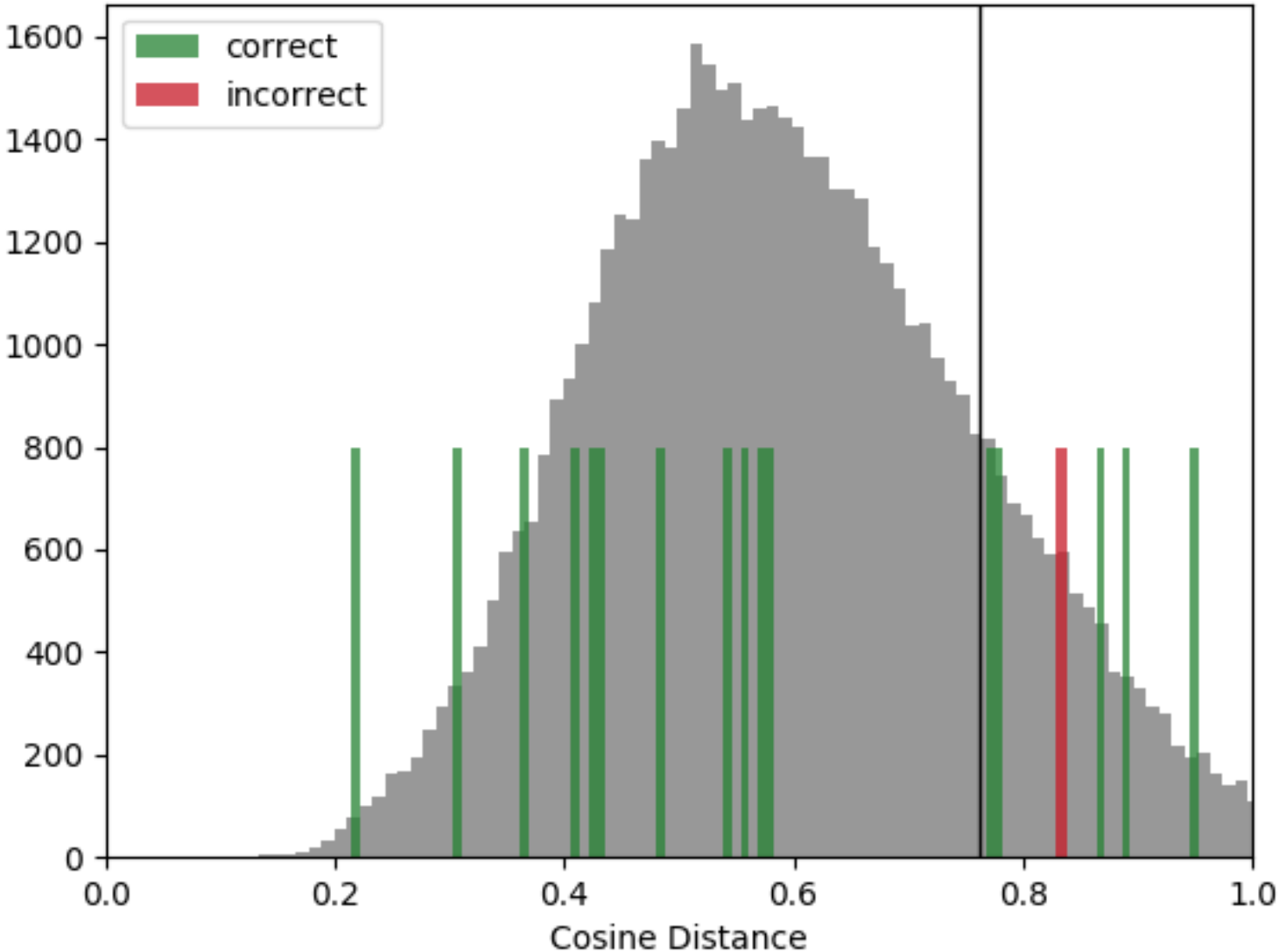
Threshold


d300both_histogramm
all words
target words
1600
1400
1200
1000
800
600
400
200
0
0.0
0.2
0.4
0.6
0.8
1.0
Cosine Distance

Results

entry	dim	threshold		ACC
#2	300	$(\mu+\sigma)$	.76	**.944**
#4	500	$(\mu+\sigma)$	.78	.889
#1	300	(50:50)	.57	.833
#3	500	(50:50)	.64	.833
major. baseline		-		.667
freq. baseline		unk.		.611
colloc. baseline		unk.		.500

Results


correct
incorrect
1600
1400
1200
1000
800
600
400
200
0
0.0
0.2
0.4
0.6
0.8
1.0
Cosine Distance

Conclusion

- **SGNS+OP+CD still rocks Semantic Change Detection**
- near to perfect accuracy of .94
- reproducing results from SemEval 2020: Task 1 (Schlechtweg et al., 2020)
- reproduced by another team (Pražák, Přibáň, & Taylor, 2020)
- off-the-shelf: no annotated data or fine-tuning of parameters
- assumes: graded change is indicative of binary classes

Bibliography

Arefyev, N., & Zhikov, V. (2020). BOS at SemEval-2020 Task 1: Word Sense Induction via Lexical Substitution for Lexical Semantic Change Detection. In *Proceedings of the 14th international workshop on semantic evaluation.* Barcelona, Spain: Association for Computational Linguistics.

Basile, P., Caputo, A., Caselli, T., Cassotti, P., & Varvara, R. (2020). Overview of the EVALITA 2020 Diachronic Lexical Semantics (DIACR-Ita) Task. In V. Basile, D. Croce, M. Di Maro, & L. C. Passaro (Eds.), *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020).* Online: CEUR.org.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1489–1501). Berlin, Germany.

Kaiser, J., Schlechtweg, D., Papay, S., & Schulte im Walde, S. (2020). IMS at SemEval-2020 Task 1: How low can you go? Dimensionality in Lexical Semantic Change Detection. In *Proceedings of the 14th international workshop on semantic evaluation.* Barcelona, Spain: Association for Computational Linguistics. Retrieved from `https://arxiv.org/abs/2008.03164`

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio & Y. LeCun (Eds.), *1st international conference on learning representations, ICLR 2013, scottsdale, arizona, usa, may 2-4, 2013, workshop track proceedings.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS.*

Pömsl, M., & Lyapin, R. (2020). CIRCE at SemEval-2020 Task 1: Ensembling Context-Free and Context-Dependent Word Representations. In *Proceedings of the 14th international workshop on semantic evaluation.* Barcelona, Spain: Association for Computational Linguistics.

Pražák, O., Přibáň, P., & Taylor, S. (2020). UWB @ DIACR-Ita: Lexical Semantic Change Detection with CCA and Orthogonal Transformation. In V. Basile, D. Croce, M. Di Maro, & L. C. Passaro (Eds.), *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020).* Online: CEUR.org.

Salton, G., & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw - Hill Book Company.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). SemEval-2020 task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation.* Barcelona, Spain: Association for Computational Linguistics.

Schönemann, P. H. (1966, Mar 01). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, *31*(1), 1–10.