CIRCE at SemEval-2020 Task 1:
Ensembling Context-Free and Context-Dependent Word Representations

Martin Pömsl[*]
Osnabrück University
mpoemsl@uos.de

Roman Lyapin
Cogent Labs Inc.
rlyapin@cogent.co.jp

Abstract

This paper describes the winning contribution to SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection (Subtask 2) handed in by team UG Student Intern. We present an ensemble model that makes predictions based on context-free and context-dependent word representations. The key findings are that (1) context-free word representations are a powerful and robust baseline, (2) a sentence classification objective can be used to obtain useful context-dependent word representations, and (3) combining those representations increases performance on some datasets while decreasing performance on others.

1 Introduction

SemEval-2020 Task 1 poses an evaluation framework for unsupervised Lexical Semantic Change Detection (LSCD). Its two subtasks operate on a set of non-parallel corpus pairs from two different time periods and are evaluated against human annotations for semantic change of a subset of words. Subtask 1 requires binary classification of whether or not the meaning of the words has changed. Subtask 2 requires ranking the words by degree of lexical semantic change and is evaluated in Spearman's rank-order correlation coefficient ρ (Schlechtweg et al., 2020). This paper primarily addresses Subtask 2.

One of the most successful methods for predicting the degree of semantic change of words is comparing context-free (static) semantic vector spaces. Such models seperately induce word vectors for all words in two corpora (e.g. with Word2Vec (Mikolov et al., 2013a)), align the resulting vector spaces and take the distance of the word vectors as a measure of semantic change (Schlechtweg et al., 2019).

With recent advances in language model pretraining, it is now possible to extract context-dependent (contextualized) word representations for each use of a word in the two corpora by using a language model (e.g. BERT (Devlin et al., 2019)) as feature extractor. The distance between these context-dependent word representations can then be taken as a measure of semantic change (Giulianelli, 2019).

In this paper, we present a model based on context-free word representations, a model based on context-dependent word representations and an ensemble model that combines their predictions. We obtain context-free representations by following the methodology of the best model reported by Schlechtweg et al. (2019), aligning SGNS vectors. For context-dependent representations, we finetune BERT with a sentence classification objective - predicting the time period of sentences - and extract internal representations for all words from the finetuned BERT model. We show that this classification finetuning can both produce useful word representations and provide an indicator for how to parameterize the ensemble.

In the evaluation phase of SemEval-2020 Task 1, the context-free model ranked first out of 128 contributions to Subtask 2. The ensemble model performed better in one language, but significantly worse in another, causing it to be ranked fifth among all contributions. The context-dependent model ranked 69th, suggesting that it can sometimes add information to the context-free model, but is largely inadequate on its own. In the results section, we analyze the submission experiments and show that the usefulness of the context-dependent representations is linked to the BERT classification accuracy. Code and predictions for all models are available publicly at https://github.com/mpoemsl/circe.

---

[*] Work done during internship at Cogent Labs Inc.

2   Related Work

2.1   LSCD with Context-Free Word Representations

Approaches that rely on the comparison of context-free word representations in diachronic corpora have a long history in lexical semantic change detection. The word vectors are usually either explicitly derived from co-occurrence statistics or implicitly through the use of neural methods (Tahmasebi et al., 2018). Context-free models often follow a three-step scheme: Representing words in semantic vectors, aligning the resulting vector spaces and comparing relevant word vectors.

Kim et al. (2014) use Skip-Gram, the neural word embedding method introduced by Mikolov et al. (2013a) to represent words over multiple diachronic corpora and compare the representations. A similar method that includes an alignment step was used by Hamilton et al. (2016).

2.2   LSCD with Context-Dependent Word Representations

Context-dependent word representations assign a semantic vector to each word-use within the context of its sentence, rather than to each unique word. One way to get context-dependent word representations is using a pretrained neural language model as a feature extractor (Peters et al., 2018). Such representations have proven useful for a wide range of NLP tasks (Liu et al., 2019) and are becoming increasingly popular for LSCD. Like context-free models, context-dependent models usually follow a three-step scheme: Extracting semantic vectors for each use of the relevant words, clustering the resulting semantic vectors and comparing the mean clusters with a distance metric.

Hu et al. (2019) use context-dependent representations derived from a neural language model as the basis for word sense tracking. Giulianelli (2019) clusters the resulting word-use representations into usage type distributions, which can then be compared.

3   Available Datasets

A complete dataset for lexical semantic change ranking as defined by Schlechtweg et al. (2019) consists of two lemmatized corpora from different time periods (t_1 and t_2) and a corresponding testset. This testset contains gold ranks for a subset of words as annotated by human experts. Predictions are made for all target words on the basis of the two corpora and evaluated against the true ranks in the testset.

In the development experiments, we validate our approach by following Schlechtweg et al. (2019) in evaluating our models on the diachronic testset DURel (Schlechtweg et al., 2018) (de-durel) and the synchronic testset SURel (Hätty et al., 2019) (de-surel) in combination with the corresponding corpora. It should be noted that de-surel presents domain-specific rather than time-specific meaning differences, but since diachronic and synchronic lexical semantic change detection are closely related, it is still a useful development dataset.

In the submission experiments, we compute our submission for the evaluation phase based on the diachronic datasets provided in SemEval-2020 Task 1 (Schlechtweg et al., 2020), which consist of an English (en-semeval), German (de-semeval), Latin (ln-semeval) and Swedish (sw-semeval) dataset. Results are averaged over submission datasets. An overview of all datasets can be found in Table 1.

		Language	Total Sentences	Annotated Words	Time Period t_1	Time Period t_2
Development	de-durel	German	$3.2 \cdot 10^6$	19	1750 - 1799	1850 - 1899
	de-surel	German	$0.9 \cdot 10^6$	21	-	-
Submission	en-semeval	English	$0.6 \cdot 10^6$	37	1810 - 1860	1960 - 2010
	de-semeval	German	$6.1 \cdot 10^6$	48	1800 - 1899	1946 - 1990
	ln-semeval	Latin	$0.6 \cdot 10^6$	40	-200 - 0	0 - 2000
	sw-semeval	Swedish	$8.5 \cdot 10^6$	31	1790 - 1830	1895 - 1903

Table 1: Properties of datasets used in development and submission experiments.

4 System Overview

We present three models for SemEval-2020 Task 1 Subtask 2: A context-free model, a context-dependent model and the ensemble model CIRCE, which stands for Classification-Informed Representation Comparison Ensemble. For Subtask 1, we binarize the CIRCE rank predictions by naively assuming that the upper half of ranks has changed while the lower half has not.

4.1 Context-Free Model

The context-free model is structured analogously to the best performing model reported by Schlechtweg et al. (2019). We adopt the use of word vectors generated by Skip-Gram with Negative Sampling (SGNS) (Mikolov et al., 2013b) as context-free word representations. Similarly, we follow their use of orthogonal Procrustes analysis (Schönemann, 1966) to align the embeddings.

However, we diverge from the best performing model reported by Schlechtweg et al. (2019) in that we employ Euclidean distance rather than cosine distance to compare the aligned representations, since this metric achieved more robust results in development experiments.

4.2 Context-Dependent Model

The context-dependent model follows a similar outline to the one described in Giulianelli (2019), with a few key changes. We adopt the use of context-dependent representations derived from the masked language model BERT (Devlin et al., 2019). We also recognize the need for domain-adaptive finetuning of the pretrained BERT model as described by Han and Eisenstein (2019).

However, instead of the standard language modelling objective, we use a sentence time classification objective. This is motivated by the assumption that a successful time classifier for sentences must learn time-specific word features that are useful for measuring lexical semantic change.

In order to reduce the number of model parameters, we do not cluster the resulting word-use representations, but instead directly compare all representations of one relevant word W_{t_1} at time t_1 and W_{t_2} at time t_2 using a Mean Pairwise Euclidean (MPE) distance metric:

$$d_{MPE}(W_{t_1}, W_{t_2}) = \frac{1}{|W_{t_1}| \cdot |W_{t_2}|} \cdot \sum_{w_{t_1} \in W_{t_1}} \sum_{w_{t_2} \in W_{t_2}} \|w_{t_1} - w_{t_2}\|$$

4.3 CIRCE

We propose the ensemble model CIRCE to combine the predictions of context-free and context-dependent models. The CIRCE rank prediction r_{CIRCE} is generated from the context-free rank r_{CF} and the context-dependent rank r_{CD} through linear combination with a single parameter $\theta \in [0.0, 1.0]$:

$$r_{CIRCE} = \theta \cdot r_{CD} + (1 - \theta) \cdot r_{CF}$$

Development experiments confirmed that there often are values for θ that cause CIRCE to perform better than both the context-free and the context-dependent model. This indicates that both kinds of representations contain unique time-specific features that can be exploited in order to predict the degree of lexical semantic change.

We employ a simple heuristic to calculate θ in an unsupervised setting such as SemEval-2020 Task 1. Development experiments suggested that the usefulness of context-dependent representations as measured by the optimal value for $\theta_{optimal} = \text{argmax}_{\theta \in [0.0, 1.0]} \, \rho(r_{CIRCE}, r_{true})$ roughly correlates with the time classification accuracy of the BERT model after finetuning. Consequently, we predict θ at test time with a simple linear mapping from the set of realistic BERT classification accuracies $acc_{BERT} \in [0.5, 1.0]$ to the set of valid CIRCE weights $\theta \in [0.0, 1.0]$:

$$\theta_{CIRCE} = 2 \cdot (acc_{BERT} - 0.5)$$

5 Experimental Setup

5.1 Preprocessing

In the context-free model, we follow Schlechtweg et al. (2019) in removing words with frequencies below a threshold. We use $\frac{|S|}{5 \cdot 10^4}$ for this threshold, where $|S|$ is the number of sentences in the corpus. In order to preserve information, we skip this step for datasets with fewer than 10^6 sentences in total.

In the context-dependent model, we create a balanced binary time classification dataset from the two diachronic corpora. We employ a train-test split of $0.8 / 0.2$ for classification finetuning and evaluation. Furthermore, we create a version of the classification dataset in which words that occur only in one corpus are replaced by the `[MASK]`-token. The intention behind this preprocessing step is to avoid the learning of rule-based features that are not useful for LSCD such as the memorization of unique words.

5.2 Implementations

We use the Word2Vec[1] SGNS (Mikolov et al., 2013b) implementation to create word vectors and VecMap[2] (Artetxe et al., 2018) to align them. We use the Transformers[3] library by HuggingFace (Wolf et al., 2019) to finetune BERT and to extract context-dependent word representations.

5.3 Parameters

In the context-free model, we create SGNS vectors of dimensionality 300 with window size 10 and negative sample 1. Following Schlechtweg et al. (2019), we length-normalize and mean-center the representations when applying orthogonal Procrustes analysis.

In the context-dependent model, we use the pretrained BERT model `bert-base-german-cased` in development experiments and `bert-base-multilingual-cased` in the SemEval submission experiments. We finetune the model with a sequence classification head for one epoch at a learning rate of $4 \cdot 10^{-5}$ with a warm-up step ratio of 0.05. We extract a context-dependent representation vector of size 768 for each word-use by feeding the whole sentence and taking the mean over the corresponding tokens in the activations of the last hidden layer of BERT.

6 Results

6.1 LSCD Performance

		Context-Free	CIRCE	Context-Dependent
Development	de-durel	0.7263	<u>0.8018</u>	0.4157
	de-surel	0.5802	<u>0.7251</u>	0.7089
Submission	en-semeval	<u>0.4221</u>	0.2465	0.0816
	de-semeval	<u>0.7253</u>	<u>0.7253</u>	0.2075
	ln-semeval	0.4124	<u>0.4637</u>	0.4439
	sw-semeval	<u>0.5467</u>	<u>0.5467</u>	0.0416

Table 2: Evaluation phase results for Subtask 2 in development and submission experiments in Spearman's ρ. For each dataset all results with the maximum value are underlined.

As Table 2 shows, the context-free model reliably scores well on all datasets for Subtask 2, while the context-dependent model only punctually matches its performance. In cases where both models achieve good results (e.g. de-surel, ln-semeval), the ensemble model CIRCE is able to exceed both predictions.

In the evaluation phase results for Subtask 2, the context-free submission ranks first with a mean correlation of 0.527 over all submission datasets, while the CIRCE submission ranks fifth at a mean correlation of 0.495. The context-dependent submission ranks 69th at a mean correlation of 0.194.

---

[1]https://github.com/danielfrg/word2vec
[2]https://github.com/artetxem/vecmap
[3]https://github.com/huggingface/transformers

In the evaluation phase results for Subtask 1, the binarized CIRCE predictions ranks fifth as well with a mean accuracy of 0.639. On en-semeval, it achieved an accuracy of 0.568, on de-semeval 0.728, on ln-semeval 0.550 and on sw-semeval 0.710.

6.2 Time Classification Performance

	Time Clf. Accuracy	CIRCE Weight θ		LSCD Spearman's ρ	
		θ_{CIRCE}	$\theta_{optimal}$	θ_{CIRCE}	$\theta_{optimal}$
de-durel	0.59	0.18	0.34	0.8018	0.8772
de-surel	0.95	0.90	0.92	0.7251	0.7264
en-semeval	0.82	0.64	0.08	0.2465	0.4274
de-semeval	0.50	0.00	0.08	0.7253	0.7345
ln-semeval	0.73	0.46	0.74	0.4637	0.4986
sw-semeval	0.50	0.00	0.16	0.5467	0.5726

Table 3: Time classification accuracies (independent of θ) as well as evaluation phase (θ_{CIRCE}) and post-evaluation phase ($\theta_{optimal}$) CIRCE weights and Subtask 2 performances in Spearman's ρ.

As Table 3 shows, time classification accuracy of BERT after finetuning varies widely across different datasets. Most remarkably, BERT entirely fails to optimize the time classification objective on de-semeval and sw-semeval, while it achieves exceptionally good results on de-surel.

In general, the classification accuracy seems to be a good predictor for the optimal weight - the submission weight θ_{CIRCE} calculated with the linear mapping is often within ± 0.20 of the optimal weight $\theta_{optimal}$, and consequently the submission correlations are often close to the optimal correlations.

One notable exception is en-semeval. The high classification accuracy leads to a predicted θ_{CIRCE} of 0.64, while $\theta_{optimal}$ is located at 0.08. As a result, CIRCE falls short of its potential in the results for Subtask 2 and ranks below the context-free model despite gains on ln-semeval, where CIRCE improves by 0.05 upon the context-free and by 0.02 upon the context-dependent model.

6.3 Error Analysis

The failure to achieve a significant time classification accuracy for de-semeval and sw-semeval is curious, but might simply be due to the properties of the datasets. Predicting the time period of a given sentence can be a difficult task even for humans, depending on several factors such as the occurrence of unique words (which are masked in the classification dataset) and the distinctiveness of grammatical structures.

The mismatch of predicted submission weight θ_{CIRCE} and optimal weight $\theta_{optimal}$ in the case of en-semeval is a more severe error, since it challenges the notion that classification accuracy is a good indicator for representation usefulness. Without extensive experiments, it is difficult to determine the cause of this outlier. However, one contributing factor might be the masking preprocessing step. While the distribution of labels in the BERT classification dataset is balanced, the distribution of [MASK]-tokens is not, since there is often one time period that has more words unique to it than the other. This makes it possible for BERT to learn [MASK]-specific rather than time-specific features during finetuning, which would cause the representations to be useful for time classification but not for LSCD.

In line with this, repeating the submission experiments without masking in the post-evaluation phase causes the classification accuracy on en-semeval to drop and the predicted submission weight to come within ± 0.20 of the optimal weight for all datasets. This modification boosts the overall mean correlation of CIRCE on the submission datasets to 0.545, which exceeds the performance of all systems for Subtask 2 in the evaluation phase. However, without further experiments on other datasets, the results of this modified model cannot be considered conclusive.

7 Conclusion

We presented a context-free model, a context-dependent model and an ensemble model for SemEval-2020 Task 1 Subtask 2. We showed that while the context-free model outperformed all other systems during the evaluation phase, ensembling its predictions with those of the context-dependent model leads to increased performance on the development datasets (de-surel, de-durel) and one submission dataset (ln-semeval) at the cost of decreased performance on another submission dataset (en-semeval).

The submission experiments have made it clear that, although its performance in the development experiments was competitive, the context-dependent model on its own is not a reliable predictor of lexical semantic change. However, it is worth pointing out that the context-dependent representations in all cases contain at least some information that is useful for LSCD, since the optimal CIRCE weight is greater than zero in all experiments. Still, the marginal improvements accomplished with this information do not justify the significant computational effort of finetuning BERT for one epoch.

In further research, it would be interesting to validate the methods described in this paper on additional datasets. In particular, it could be worthwhile to empirically explore the link between classification accuracy and representation usefulness. If the relation were to hold up for other domains and models with a lower computational complexity than BERT, representations obtained through self-supervised classification training could be used on a whole range of other unsupervised tasks.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mario Giulianelli. 2019. Lexical semantic change analysis with contextualised word representations. Master's thesis, University of Amsterdam, Amsterdam, Netherlands.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 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)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China. Association for Computational Linguistics.

Anna Hätty, Dominik Schlechtweg, and Sabine Schulte im Walde. 2019. SURel: A gold standard for incorporating meaning shifts into term extraction. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 1–8, Minneapolis, Minnesota. Association for Computational Linguistics.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 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.

Peter Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to lexical semantic change.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing.