CogALex 2.0: Impact of Data Quality on Lexical-Semantic Relation Prediction

Christian Lang
University of Vienna
christian.lang@univie.ac.at

Lennart Wachowiak
King's College London
lennart.wachowiak@kcl.ac.uk

Barbara Heinisch
University of Vienna
barbara.heinisch@univie.ac.at

Dagmar Gromann
University of Vienna
dagmar.gromann@gmail.com

Abstract

Predicting lexical-semantic relations between word pairs has successfully been accomplished by pre-trained neural language models. An XLM-RoBERTa-based approach, for instance, achieved the best performance differentiating between hypernymy, synonymy, antonymy, and random relations in four languages in the CogALex-VI 2020 shared task. However, the results also revealed strong performance divergences between languages and confusions of specific relations, especially hypernymy and synonymy. Upon inspection, a difference in data quality across languages and relations could be observed. Thus, we provide a manually improved dataset for lexical-semantic relation prediction and evaluate its impact across three pre-trained neural language models.

1 Introduction

Data-driven machine and deep learning models continue to make strident advances on a variety of tasks, however, their efficacy relies on the availability of abundant and accurate data. On tasks where a paucity of data can be observed, such as predicting lexical-semantic relations, data quality moves to center stage. The importance of data quality has been well-known for decades (see e.g. [1]), but has been largely neglected in favor of focusing on algorithms and architectures [2]. Nevertheless, poor data quality negatively impacts model performance, likely leading to unreliable predictions, especially if available training data is scarce. We evaluate the impact of data quality improvements on model performance in relation prediction. To this end, we propose an improved multilingual dataset for lexical-semantic relation prediction CogALex 2.0 and evaluate the performance impact on four pretrained language models, one monolingual and three multilingual [1].

One of the few multilingual datasets for this task was proposed within the CogALex-VI 2020 shared task[2] [3]. It represents three paradigmatic relations — hypernymy, synonymy, antonymy — and a random relation between word pairs in English, German, Chinese train/val/test sets and an Italian test set. The best performing model in the shared task, Transrelation [4], relies on fine-tuning XLM-RoBERTa (XLM-R) [5] and shows strong weighted F1 divergences across languages and relations. With Chinese, the smallest language set, performance was substantially better than with German and English data. With German, the lowest performance and highest relation confusion was observed.

A manual inspection of the dataset revealed reasons for this confusion. Several pairs labeled with hypernymy would likely be classified as synonyms, by neural and human classifiers, e.g. (fett, HYP,

[1] Dataset and code are available here: https://github.com/Text2TCS/CogALex-2.0

[2] Licensed under Creative Commons Attribution 4.0 International Licence adopted for CogALex 2.0.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Table 1: Label distribution per language of CogALex 1.0 and 2.0

	DE			**EN**			**ZH**		
	Train	Val	Test	Train	Val	Test	Train	Val	Test
HYP	841	294	286	898	292	279	421	145	129
SYN	782	272	265	842	259	266	402	129	122
ANT	829	275	281	916	308	306	361	136	142
RANDOM	2430	786	796	2554	877	887	1330	428	445
Total	**4882**	**1627**	**1628**	**5210**	**1736**	**1738**	**2514**	**838**	**838**

dick) (en: fat, HYP, corpulent). Apart from misleading relations, we also observed decomposed compounds marked as synonyms/hypernyms, e.g. (historic, SYN, landmark), duplicate pairs in training and test set, etc. (full list in Section 2.2). These quality issues and differences across languages can partially be attributed to the fact that the dataset was accumulated from existing monolingual datasets, each with its individual data collection method. Issues observed for this dataset and task are common to semi-automatically created datasets.

We propose a quality-improved version of the original dataset and evaluate its impact on not only Transrelation but in total three models with different training configurations and sizes, i.e., base, large, multilingual, monolingual, cased, and uncased. To assess the generalizability of the data quality improvements, we evaluate on the original Chinese train/val/test sets and Italian test set. We found that dataset quality improvements positively impact model performance with up to 6.2%.

2 Dataset

2.1 CogALex-VI

CogALex-VI represents a multilingual dataset created from four monolingual datasets [3]. The English dataset is derived from EVALution 1.0 [6], which was created by automatically filtering ConceptNet and WordNet and checking the results via crowdsourcing. The German dataset [7] was also obtained by crowdsourcing, where word pairs where balanced for semantic category, polysemy, corpus frequency, and word class, i.e., adjectives, nouns, and verbs. For Chinese, taken from [8], a combination of filtering Chinese WordNet and eliciting relation targets from Chinese participants for six relations was utilized. Hypernymy and hypoynymy were explicitly considered and raters were not presented with word pairs but commissioned to provide related words themselves. Building on these three languages and datasets, the shared task provided labeled training and validation data and initially unlabeled test data. In addition, a surprise Italian test set was provided after the challenge deadline. The distribution of relations across languages is depicted in Table 1.

Curiously, many words appeared more than once in each dataset and across splits (train, val, test). For instance, in German the word “abschneiden” was the most frequent with 63 occurrences, followed by “komplex” with 59, and “verbauen” with 49. The most frequent English words were “fight” with 71, “corrupt” with 67, and “knowledgeable” with 55 occurrences.

2.2 Annotation Procedure

The improved CogALex 2.0 has been created by four domain experts, each evaluating one fourth of the English and German original datasets adhering to the following annotation guidelines:

1. fix upper- and lower-case errors, e.g., “mesz” was changed to “MESZ” (CEST)
2. change at least one word if both are identical
3. change HYP word order to be hyponym–hypernym, e.g., (apple, HYP, fruit)
4. change word(s) or label:
    - if the annotation is RANDOM but the relation is SYN/HYP/ANT
    - if compound term is depicted as relation, e.g., (postal, HYP, region)
    - for incorrect relations, e.g., (elephant, SYN, animal) changed to HYP
    - if both are not of the same word class

Table 2: Overall changes

	DE			EN		
	Train	Val	Test	Train	Val	Test
Capitalization	3214	1021	1004	16	3	3
Replaced word	912	365	299	914	442	405
Replaced relation	485	124	38	52	48	28
Underscores	0	0	0	254	69	71
Total	**4611**	**1510**	**1341**	**1236**	**562**	**507**

Table 3: Changes per label (no capitalization)

	DE			EN		
	Train	Val	Test	Train	Val	Test
HYP	385	150	92	319	149	121
SYN	255	93	77	291	103	115
ANT	202	79	67	310	117	93
RANDOM	164	45	13	206	69	60
Total	**1,006**	**367**	**249**	**1,126**	**438**	**389**
% Changes	**20.6%**	**22.6%**	**15.3%**	**21.6%**	**25.2%**	**22.4%**
with capitalization	**47.9%**	**50.3%**	**45.2%**	**21.6%**	**25.2%**	**22.4%**

Some HYP examples relied on co-hyponyms (branch, HYP, leaf), which we also adapted to more prototypical hypernymy examples. Antonymy relations contained particularly many curious examples, e.g., (lobster, ANT, cow) and (cod, ANT, steak), which we adapted to more common antonyms.

In terms of word choice, the original dataset contained very uncommon, biased, and/or vulgar expressions, including some non-existing words, which were replaced by examples compatible with the remainder of the dataset. For instance, "bumsen" (to bang or thud) was repeatedly used in the German proportion of the data, unparalleled by the English choice of expressions on the level of profanity. Since this replacement procedure could potentially aggravate the duplication of examples, we deduplicated the final dataset, considering also duplicates with a changed directionality in the train/val/test sets.

2.3 Dataset Statistics

Table 1 shows the distribution of the labels per language and individual split of the original dataset. For comparability, CogALex 2.0 follows the same distribution of relations as the original dataset.

The fact that Chinese obtained the best results in Translrelation together with the described differences in dataset creation, i.e., human annotators and relation elicitation instead of crowdsourcing, led to the assumption that it is the dataset of the highest quality from the selection. Thus, it remained unchanged in this new dataset as a testbed for the data quality improvements and its generalizability evaluation.

In total 9,767 sequences were corrected for English and German. In Table 2, the statistics per change operation are presented, where "Underscores" refers to replacing underscores with white spaces in multi-word sequences to align the German representation with that of the other languages. Table 3 represents modified triples according to relation type, which clearly shows the high need of adaptations in German with more than 45% of changes with respect to the original for each dataset. In English more than 20% of each split had to be adapted.

3 Benchmark

3.1 Model

Transrelation made use of the multilingual XLM-R language model [5] trained on 100 languages in its base configuration [9] and serves as our baseline. A linear layer added on top of the pooled output allows for the classification into one of the four possible classes. We included other BERT-based models to compare the effectiveness of XLM-R on CogALex 2.0 as well as to evidence the effectiveness of data improvement irrespective of the pre-trained model or its parameter size.

Table 4: F1 scores for each combination of datasets, languages and models; baselines for each model are represented in bold; abbreviations: b = base, l = large, m = multilingual, c = cased, u = uncased.

Model	Train	Test$_{de}$	Test$_{en}$	Test$_{zh}$	Test$_{it}$	Test$_{de}$ 2.0	Test$_{en}$ 2.0
XLM-R$_{bm}$	CogALex 1.0	**59.8**	**65.3**	**90.6**	**55.8**	+4.8	+3.8
	DE 2.0	-0.1	-0.9	+0.3	0.0	+4.3	+3.8
	EN 2.0	-0.6	-0.8	+0.8	-0.3	+3.2	+4.6
	CogALex 2.0	0.0	-1.0	+1.6	-1.5	+5.1	+4.7
XLM-R$_{lm}$	CogALex 1.0	**71.4**	**73.5**	**92.0**	**63.8**	+5.1	+5.6
	CogALex 2.0	-0.5	-0.7	+0.7	-0.9	+5.1	+6.2
BERT$_{bmc}$	CogALex 1.0	**48.7**	**57.0**	**88.6**	**48.0**	+3.1	+2.6
	CogALex 2.0	+0.6	+1.0	-0.3	-3.7	+4.8	+4.9
BERT$_{bmu}$	CogALex 1.0	**53.6**	**58.5**	**88.3**	**49.0**	+1.3	+3.2
	CogALex 2.0	-0.8	+1.3	-0.3	+0.9	+1.8	+5.5
DistilBERT$_{bmc}$	CogALex 1.0	**45.0**	**52.9**	**85.1**	**40.0**	+2.2	+2.9
	CogALex 2.0	+0.3	+0.1	-0.4	-0.8	+3.4	+2.9
RoBERTa$_{b}$	CogALex 1.0	**44.8**	**62.8**	**78.5**	**37.3**	+0.5	+2.8
	CogALex 2.0	+0.8	+0.5	-1.7	-3.4	+1.5	+5.1

3.1.1 Results

We computed the scores for each model type by training 5 individual models and averaging the scores. Results of our comparison are presented in Table 4 as increases or decreases to the baseline. CogALex 1.0 refers to the original dataset. DE 2.0 denotes improved German and EN 2.0 improved English training data, while CogALex 2.0 refers to both. Test$_{de}$ 2.0 and Test$_{en}$ 2.0 refer to improved test data.

As is to be expected, improved training data showed little impact when evaluated on the problematic original test data. Interestingly, XLM-R results for Chinese increase with improved training data. Other languages show higher gains with the improved test data. In fact, even when trained on CogALex 1.0, the models' performance improved considerably with just the corrected test data. The decrease of performance on original test data indicates that new prediction patterns are being learned.

In terms of model comparison, little difference for cased and uncased models can be observed, only for Italian the latter improves performance. However, parameter size considerably impacts XLM-R performance with up to 11.6% gain in the large configuration. This indicates that the task might be complex and models benefit from a higher number of features for making accurate predictions. Also BERT consistently outperforms DistilBERT. We tested on the monolingual English RoBERTa, which achieved surprisingly high scores across all languages and confirmed our intuition of positive data quality impact even on transferring learned language information to other languages.

4 Discussion and Conclusion

The proposed adaptation measures for data quality improvement had a positive impact on performance results across models. That XLM-R's performance even increases for the unchanged Chinese data indicates that the data adaptations did not merely introduce a novel pattern that is easier to learn. We provide evidence that for this task it is effective to improve all datasets: train, validation, and test.

The created CogALex 2.0 dataset is still missing some phenomena and use-cases of synonyms, hypernyms, and antonyms occurring in natural language. Firstly, the current dataset does not consider multi-word sequences, e.g., "pre-trained multilingual neural language models". Secondly, morphological variations are missing, e.g., "flower" and "flowers". Thirdly, directionality of hypernymy is currently assumed to be hyponym–hypernym. For an input with the switched order it is difficult to interpret what a trained model would predict. Since in a realistic deployment scenario the order of the input words is not known, the dataset needs two labels that reflect both directions.

For future work we plan to consider the problematic issues derived from the original dataset addressed above as well as extending the number of languages and model comparisons. We also intend to perform further experiments on the impact of data quality on transfer learning to other languages and particularly domain-specific corpora containing conventionally numerous multi-word terms.

References

[1] Richard Y Wang and Diane M Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996.

[2] Lisa Ehrlinger, Verena Haunschmid, Davide Palazzini, and Christian Lettner. A daql to monitor data quality in machine learning applications. In *International Conference on Database and Expert Systems Applications*, pages 227–237. Springer, 2019.

[3] Rong Xiang, Emmanuele Chersoni, Luca Iacoponi, and Enrico Santus. The CogALex shared task on monolingual and multilingual identification of semantic relations. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 46–53, Online, December 2020. Association for Computational Linguistics.

[4] Lennart Wachowiak, Christian Lang, Barbara Heinisch, and Dagmar Gromann. CogALex-VI shared task: Transrelation - a robust multilingual language model for multilingual relation identification. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 59–64, Online, December 2020. Association for Computational Linguistics.

[5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.

[6] Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. Evalution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 64–69, 2015.

[7] Silke Scheible and Sabine Schulte Im Walde. A database of paradigmatic semantic relation pairs for german nouns, verbs, and adjectives. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 111–119, 2014.

[8] Hongchao Liu, Emmanuele Chersoni, Natalia Klyueva, Enrico Santus, and Chu-Ren Huang. Semantic relata for the evaluation of distributional models in mandarin chinese. *IEEE access*, 7:145705–145713, 2019.

[9] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.