CogALex-VI Shared Task: Transrelation - A Robust Multilingual Language Model for Multilingual Relation Identification

Lennart Wachowiak
University of Vienna / Vienna, Austria
lennartw99@univie.ac.at

Christian Lang
University of Vienna / Vienna, Austria
a0809558@univie.ac.at

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

Dagmar Gromann
University of Vienna / Vienna, Austria
dagmar.gromann@univie.ac.at

Abstract

We describe our submission to the CogALex-VI shared task on the identification of multilingual paradigmatic relations building on XLM-RoBERTa (XLM-R), a robustly optimized and multilingual BERT model. In spite of several experiments with data augmentation, data addition and ensemble methods with a Siamese Triple Net, Translrelation, the XLM-R model with a linear classifier adapted to this specific task, performed best in testing and achieved the best results in the final evaluation of the shared task, even for a previously unseen language.

1 Introduction

Determining whether a semantic relation exists between words and which type of relation it represents is a central challenge in numerous NLP tasks, such as extracting terminological concept systems and paraphrase generation. Adding a multilingual dimension renders this task at the same time more relevant and more challenging. Recent approaches rely on aligned vector spaces for individual languages (Bojanowski et al., 2017) or meta-learning approaches (Yu et al., 2020) for hypernymy detection and a Siamese Triple Net for antonymy-synonymy distinction inherent in word embeddings (Samenko et al., 2020). However, in general a distinction of paradigmatic relations with word embeddings is difficult (im Walde, 2020). In a multilingual scenario, frequently lexical resources are utilized to reinforce the model's transfer learning abilities (Geng et al., 2020). Given relatively small training datasets and a necessity to support a previously unknown language, we decided to rely on a multilingual pretrained language model.

The CogALex-VI shared task focuses on the identification of semantic relations of the types synonymy (e.g. *chap* and *man*), antonymy (e.g. *big* and *small*), hypernymy (e.g. *screech* and *noise*), or random (e.g. *ink* and *closure*) between a given word pair. Random indicates that the word pair is unrelated. The shared task provided two subtasks. For the first subtask, participating teams were allowed to design monolingual systems being provided training and validation data for the languages Mandarin Chinese, German, and English. For the second subtask, participating teams were expected to design a single multilingual system that can correctly classify semantic relations in all three languages as well as a previously unknown surprise language, which turned out to be Italian. Additional resources were permitted with the exclusion of anything related to WordNet (Miller, 1995) or ConceptNet (Liu and Singh, 2004).

Our initial intention was to target the second subtask with a multilingual system relying on the state-of-the-art multilingual model XLM-RoBERTa (XLM-R) (Conneau et al., 2020) adapted to the task at hand utilizing a linear layer and CogALex-VI training datasets, a model we call *Transrelation* that we provided within the Text to Terminological Concept System (Text2TCS)[1] project. To support the model's ability to distinguish relations we experimented with data augmentation, data addition and ensemble methods, joining Transrelation[2] with a model trained on a Siamese Triple Net. Finally, the adapted XML-R model outperformed all other experiments as well as all other submitted models to CogALex-VI on both tasks.

[1] https://text2tcs.univie.ac.at/

[2] Code and datasets are available at https://github.com/Text2TCS/Transrelation

Proceedings of the Workshop on Cognitive Aspects of the Lexicon, pages 59–64
Barcelona, Spain (Online), December 12, 2020

2 Background

2.1 Lexico-Semantic Relations

Lexico-semantic relations, also called semantic and lexical semantic relations, represent the major organizing means for structuring lexical knowledge. A common distinction for such relations is between paradigmatic and syntagmatic relations, where the former represents relations between natural language expressions that could be found in the same position in a sentence and the latter refers to co-occurring elements. Importance of paradigmatic relations might differ by word class (im Walde, 2020), i.e, hypernymy is particularly central for the organization of nouns but less important for organizing verbs. In the CogALex VI shared task all relations are paradigmatic, which are particularly difficult to be distinguished by regular word embedding models and between different word classes (im Walde, 2020).

2.2 Relation Identification

Recent approaches trying to identify hypernym relations in a multilingual setting utilize fastText embeddings (Bojanowski et al., 2017) of different languages being aligned into a single vector space (Wang et al., 2019) or train models using different fastText embeddings in a multilingual setting with the help of meta-learning algorithms (Yu et al., 2020). Synonym and antonym differentiation has been a key problem for automatic relation identification and has in the past been tackled with partial success using word alignment over large multilingual corpora with statistical methods to determine distributional similarity (van der Plas and Tiedemann, 2006) or statistical translation to a pivot language for synonymy discovery (Wittmann et al., 2014). Samenko et al. (2020) utilize Siamese Triple Nets (Bromley et al., 1994) to train so-called contrasting maps, vector representations trained on monolingual embeddings that reinforce the distinction between antonyms and synonyms. Approaches that tackle all three relations at once in a multilingual environment frequently rely on active transfer learning and lexical resources (Geng et al., 2020) or prototypical vector representations for each type of relation (im Walde, 2020).

2.3 Language Models

Recent advances in the field of natural language processing are based on deep neural language models, which can be pretrained on large amounts of data in an unsupervised fashion and are fine-tuned afterwards on a specific task making use of the previously learned language representations. One of the most prominent example of such a model is BERT (Devlin et al., 2018) utilizing the now ubiquitous Transformer architecture. Compared to earlier approaches like word2vec (Mikolov et al., 2013) and fastText (Bojanowski et al., 2017) the word embeddings generated by these deep neural language models are context-specific, i.e., a word's embedding changes depending on its surrounding words. Language models do not have to be monolingual, but the pretraining can be extended to multiple languages at the same time, e.g. by making use of a shared subword vocabulary. Prominent examples are multilingual BERT and the more recent XLM-R (Conneau et al., 2020).

3 System Description

3.1 Architecture

Our system makes use of the multilingual language model XLM-R (Conneau et al., 2020). We use the implementation provided by the transformers library (Wolf et al., 2019), which offers the XLM-R model pretrained on 100 different languages using CommonCrawl data. We use the base model size, which uses less parameters than the large version of XLM-R, but performed equally well in our experiments. A linear layer is added on top of the pooled output in order to allow for classification into one of the four possible classes, i.e., three semantic relations or random.

3.2 Datasets

The CogALex VI shared task provided training and validation datasets in English (Santus et al., 2015), German (Scheible and Im Walde, 2014) and Mandarin Chinese (Liu et al., 2019). The test data for the surprise language Italian were taken from Sucameli and Lenci (2017). Word pair counts for the training datasets are provided in Table 1.

Language	ANT	HYP	SYN	RANDOM
English	916	998	842	2554
German	829	841	782	2430
Chinese	361	421	402	1330

Table 1: Word pair counts of training sets

Language	ANT	HYP	SYN	Weighted
English	0.587	0.483	0.473	0.517
German	0.534	0.535	0.427	0.500
Chinese	0.914	0.876	0.849	0.881
Italian	0.447	0.462	0.513	0.477

Table 2: F1-score on test set

3.3 Input and Preprocessing

The input provided to the model consists of a word pair labeled with a relation surrounded by XLM-R specific classification and sequence separation tokens, as well as additional padding tokens, which guarantee that all inputs have the same length. For instance, the input pair *tiger* and *animal* is encoded as ‘<s>’, ‘_tiger’, ’</s>’, ‘</s>’, ‘_animal’, ‘</s>’, excluding the padding tokens.

3.4 Training and Hyperparameters

This model was then trained on the training datasets (see Table 1) in three languages simultaneously. Hyperparameters were fine-tuned manually and via gridsearch on the given validation sets. The best results were achieved with the following hyperparameters: Optimizer: AdamW, Learning rate = 2e-5, Epsilon = 1e-8, Weight Decay = 0, Warm-up steps = 0, Epochs = 7, Batch size = 32.

4 Results and Analysis

Table 2 shows the results of our model on the four provided test sets. The computed score is a weighted F1-score excluding unrelated words labeled with RANDOM. The strongest performance can be observed in Chinese with a weighted F1-score of 0.881. English and German are far behind with scores of 0.517 and 0.500 respectively. Interestingly, the model performs nearly as well on the Italian test set with a score of 0.477, although the model had not been trained on this language, thus showing the remarkable zero-shot-learning abilities of XLM-R.

Fig. 1 shows the normalized confusion matrix based on the joined results on all four test sets. Besides confusing meaningful relations with RANDOM, which can be explained by the fact that RANDOM is the majority class, the highest confusion exists between hypernyms and synonyms. For Chinese, for instnace, 19 HYP/SYN labeled test examples were confused. From these examples, in 11 pairs some characters in one sequence are present in the other, such as 海水- 水(sea water - water) (label: HYP) and 船- 船舶(ship/boat - ship) (label: SYN). This also occurred in four SYN/ANT labeled examples, e.g. 無線- 有線(wireless - wired) (gold: ANT). For the remainder of wrongly classified SYN/ANT examples, our model frequently selected RANDOM, e.g. 私人- 公立(private individual - public) (gold: ANT).

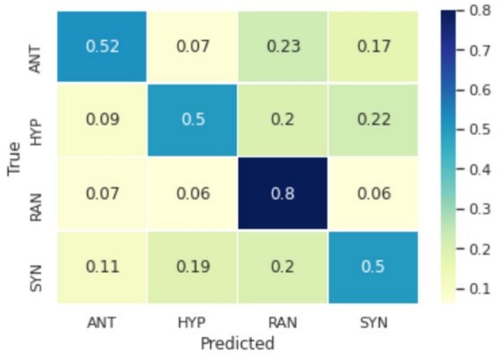

Figure 1: Normalized Confusion Matrix

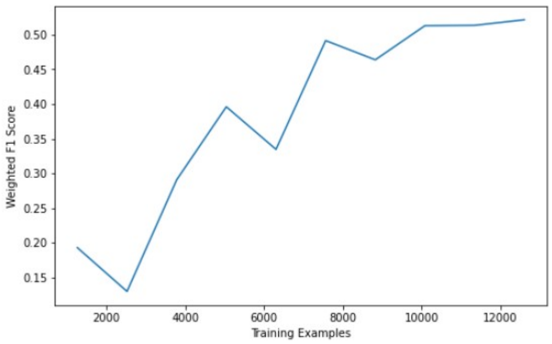

Figure 2: Learning Curve

The learning curve shown in Fig. 2 plots the achieved weighted F1 score in relation to the number of samples in the training set. For each training set size we trained four models and reported the highest observed score. The model greatly benefits from additional training samples when the training set size is below 8,000. However, the usefulness of adding more data diminishes quickly as the learning curve seems to plateau towards the end. This was confirmed when we tried to add additional training data to data provided by CogALex-VI observing the WordNet/ConceptNet exclusion.

5 Discussion

In additional experiments we trained a Siamese Triplet Net (Bromley et al., 1994) to learn meta-embeddings that contrast synonyms and antonyms, which we also tried for hypernym and synonym distinction. However, an ensemble method combining this model and XLM-R performed worse than XLM-R on its own. Due to our model's strong performance in Chinese we also experimented with data augmentation by machine translating the training and validation sets from Chinese to the other languages. The model's performance on these translated datasets was, however, considerably worse than solely on the original untranslated datasets. Additionally, performance of both models trained for individual languages or consecutively one language after another lagged considerably behind our final model.

Given the vast differences in model performance on the different languages, we briefly analyzed the data quality. In the confusion matrix in Fig. 1 it becomes evident that our model tended to confuse hypernyms and synonyms a well as random and antonyms. A brief check on the German data where the model performed worse showed that some word pairs labeled as hypernyms might be understood as synonyms by human classifiers, e.g. *fett* (fat) - *dick* (plump), *unruhig* (anxious/restless) - *erregt* (excited/aroused), and *radikal* (radical) - *drastisch* (radical/extreme) could instead be labeled as synonyms. Additional training data not related to WordNet or ConceptNet we experimented with (e.g. Kober et al. (2020)) had similar issues and data addition did not improve performance of both the tested models. So on the one hand we attribute this confusion problem of our model to word pairs that might easily be confused by human users. On the other hand, the number of training examples was rather low and data augmentation/addition with high-quality data might have considerably improved performance.

Depending on the fact that the semantics of these examples change with context, we believe that providing words in context could be one way to alleviate this misclassification problem. One curious example underlining this issue was the result we got for the surprise language Italian not seen during training, where *farfalla* (butterfly) and *coccinella* (ladybug) are labeled as antonyms, while our system labeled the pair as a synonym. Since both can be used to lovingly refer to a young female person in Italian, the result of our system could be regarded as correct if the words are understood in this sense. Further such examples can be found in great number in the training, validation and test datasets. Curiously, performance on Mandarin Chinese did not seem to be impacted as heavily by this problem, which might be due to the fact that the training datasets were compiled from a different source of different quality.

6 Conclusion

In this paper, we present our system Transrelation for the CogALex VI shared task on multilingual relation identification called Transrelation. We experimented with data addition, data augmentation and ensemble methods joining pretrained transformer-based models with a Siamese Triple Net. The final system is based on the multilingual pretrained language model XLM-R, which turned out to be the winning system and delivered a strong performance on all four languages, including one previously unknown and unseen additional language.

In the future, it would be interesting to apply ideas from curriculum learning (Bengio et al., 2009) or meta-learning, as already done for simpler models in the case of hypernymy detection (Yu et al., 2020) to improve the learning process of our model. This would especially apply to similar scenarios of few available training datasets. Furthermore, it would be interesting to evaluate the model's performance on different lexico-semantic relations as well as languages from different language families, e.g. Slavic.

Acknowledgements

This work has been supported by the project Text2TCS funded by the European Language Grid H2020 (grant number 825627).

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1994. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. pages 8440–8451, July.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

ZhiQiang Geng, GuoFei Chen, YongMing Han, Gang Lu, and Fang Li. 2020. Semantic relation extraction using sequential and tree-structured lstm with attention. *Information Sciences*, 509:183–192.

Sabine Schulte im Walde. 2020. Distinguishing between paradigmatic semantic relations across word classes: human ratings and distributional similarity. *Journal of Language Modelling*, 8(1):53–101.

Thomas Kober, Julie Weeds, Lorenzo Bertolini, and David Weir. 2020. Data augmentation for hypernymy detection. *ArXiv e-prints*.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Igor Samenko, Alexey Tikhonov, and Ivan P. Yamshchikov. 2020. Synonyms and Antonyms: Embedded Conflict. *arXiv:2004.12835v1 [cs]*.

Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. 2015. 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.

Silke Scheible and Sabine Schulte Im Walde. 2014. 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.

Irene Sucameli and Alessandro Lenci. 2017. Parad-it: Eliciting italian paradigmatic relations with crowdsourcing. *CLiC-it 2017 11-12 December 2017, Rome*, page 310.

Lonneke van der Plas and Jörg Tiedemann. 2006. Finding synonyms using automatic word alignment and measures of distributional similarity. (July):866–873.

Chengyu Wang, Yan Fan, Xiaofeng He, and Aoying Zhou. 2019. A family of fuzzy orthogonal projection models for monolingual and cross-lingual hypernymy prediction. In *The World Wide Web Conference*, pages 1965–1976.

Moritz Wittmann, Marion Weller, and Sabine Schulte Im Walde. 2014. Automatic extraction of synonyms for German particle verbs from parallel data with distributional similarity as a re-ranking feature. *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, (1998):1430–1437.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

Changlong Yu, Jialong Han, Haisong Zhang, and Wilfred Ng. 2020. Hypernymy detection for low-resource languages via meta learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3656.