CogALex-VI Shared Task: Bidirectional Transformer based Identification of Semantic Relations

Saurav Karmakar
Insight Centre for Data Analytics
Data Science Institute
National University of Ireland Galway
Galway, Ireland
`saurav.karmakar@insight-centre.org`

John McCrae
Insight Centre for Data Analytics
Data Science Institute
National University of Ireland Galway
Galway, Ireland
`john@mccr.ae`

Abstract

This paper[1] presents a bidirectional transformer based approach for recognising semantic relationships between a pair of words as proposed by CogALex VI[2] shared task in 2020. The system presented here works by employing BERT embeddings of the words and passing the same over tuned neural network to produce a learning model for the pair of words and their relationships. Afterwards the very same model is used for the relationship between unknown words from the test set. CogALex VI[2] provided Subtask 1 as the identification of relationship of three specific categories amongst English pair of words and the presented system opts to work on that. The resulted relationships of the unknown words are analysed here which shows a balanced performance in overall characteristics with some scope for improvement.

1 Introduction

Predicting the relationship between two words in terms of semantics has become a quintessential problem to be solved in the present day NLP world and reflect great impacts on the theoretical psycholinguistic modeling of the mental lexicon as well. The field of NLP finds many useful applications through tackling this direction, such as thesaurus generation (Grefenstette, 1994), ontology learning (Zouaq and Nkambou, 2008), paraphrase generation and identification (Madnani and Dorr, 2010), question answering and recognizing textual entailment (Dagan et al., 2013), as well as drawing inferences (Martínez-Gómez et al., 2016). Many NLP applications make use of handcrafted resources such as WordNet (Fellbaum, 1998). As a matter of fact, WordNet came from a similar direction with a substantial manual effort. Creating such resources is expensive and time consuming; thus efforts of this sort do not cover the variety of languages equally. Practically, coverage of a wide range of languages through such manual initiatives also far from completion. Many organizations and institutes, who are interested in creating knowledge bases on the field of their practice have attempted to classify such word pairs to shape taxonomies (Pereira et al., 2019).

Lately distributional or corpus based approaches came into popularity for investigating the semantic linkage between words; this approach utilizes the usage and appearance of the words in the corpus. These methods have been able to reflect potential in pattern-recognition-based exploration for word to word semantic mapping through distributional parameters. Exploring and connecting semantic relationships are quite difficult and variety of approaches have been tried yet.

Cognitive Aspects of the Lexicon VI (CogALex VI)[2] has arranged a shared task in 2020; it was looking to explore different efforts to figure out paradigmatic semantic relations, specifically synonymy, antonymy and hypernymy. In the field of NLP these type of relations are notoriously difficult to be distinguished between word pairs given a distribution.

To tackle this problem, we employ a deep learning framework to develop training models and test their performances through semantic link prediction between unknown word set. In this paper, we demonstrate our bidirectional transformer based approach to classify whether a given pair of words are semantically

---

[2]https://sites.google.com/view/cogalex-2020/home/shared-task

connected with one of the three relationships just mentioned above or some other random ones: as asked by this shared task. In the following Section 2 we describe the related work, Section 3 provides the system description; following that Section 4 describes the experimental set up, while Section 5 describes the results and Section 6 lays out the conclusion and future direction.

2 Related Work

2.1 Identifications of semantic relations

Recognizing the semantic meaning of words in terms of connecting them with semantic relationships has become a key direction to grow knowledge base and many further practices in NLP. This connects to a wide range of applications, such as textual entailment, text summarization, sentiment analysis, ontology learning, and so on. Following this, several supervised and unsupervised approaches have been initiated and for reference the works of (Lenci and Benotto, 2012) and (Shwartz et al., 2016). (Mohammad et al., 2013) and (Santus et al., 2014) on antonymy are of relevance here. One key commonality amongst these were that these approaches targeted one semantic relationship discovery at once amongst the words rather multiple. There were pattern based multiclass classification task carried out by (Turney, 2008) on similarity, antonymy and analogy, and by (Pantel and Pennacchiotti, 2006) on generic pattern recognition and filtering. These approaches resulted in higher precision and lower recall compared to distributional-semantic-model-based methods due to their sole dependency on patterns. The later mentioned method has been explored in an unsupervised manner (Weeds and Weir, 2003); (Lenci and Benotto, 2012); (Santus et al., 2015) and didn't measure up in terms of efficacy. Thereafter supervised methods have been adopted (Kruszewski et al., 2015); (Roller and Erk, 2016); (Nguyen et al., 2016); (Shwartz et al., 2016) in the very same direction aiming for classifying the multiclass relationships better. Count-based vectors have been substituting the prediction based ones in some recent approaches, which apparently performed better in some task, such as similarity estimation (Baroni et al., 2014), even though (Levy et al., 2015) demonstrated that these improvements were most likely due to the optimization of hyper-parameters that were instead left unoptimized in count based models. (Shwartz et al., 2016) had an approach combining patterns and distributional information reflected promising parameters in hypernymy recognition.

2.2 Shared Task regarding Semantic Relations Identification

Several shared tasks has been emerged from the NLP related conferences in this decade and the following covers a brief survey on such tasks. Seven "encyclopedic" semantic relations between nouns (cause-effect, instrument-agency, product-producer, origin-entity, content-container, theme-tool, part-whole) were asked for exploration in the SemEval-2007 shared task 4 (Girju et al., 2007). The participants were allowed to use WordNet synsets on the sentences in which the noun pairs could be observed for this task. There were fifteen participants and the best one achieved 76.3% average accuracy. Entity-destination, component-whole, member-collection and message-topic relations were added for exploration along with the first five semantic relations of SemEval-2007 (Girju et al., 2007) shared task 4 in the SemEval-2010 shared task 8 (Hendrickx et al., 2009). Given a sentence and two tagged nominals, the task was to predict the relation between those nominals and its direction towards which these nominals were pointing to the relationships. Twenty-eight participants explored this with the freedom of using semantic, syntactic and morphological resources and the best system produced 82% accuracy. SemEval-2015 (Bordea et al., 2015) and SemEval-2016 (Bordea et al., 2016) were the initiative to find participation and exploration on taxonomy generation through a specific lexical semantic relation identification of hypernymy (and its inverse, hyponymy). A list of domain terms were provided as the test data and formation of taxonomy (a list of pairs: [term, hypernym]) is asked with possible addition of intermediate terms when needed. The participating systems experimented using dictionary definitions, Wikipedia, knowledge bases, lexical patterns and vector space models. Related to this SemEval-2016 Task 14 (Jurgens and Pilehvar, 2016) asked participants to enrich WordNet taxonomy by augmenting new words to the existing synsets (thus combining detection of hypernyms with word sense disambiguation). The last CogALex shared task (CogALex V)[3] is different in terms of the relationship explorations

---

[3]https://sites.google.com/site/cogalex2016/

from the earlier mentioned shared tasks. CogALex V[3] asked for the detection of synonymy, antonymy, hypernymy, part-whole meronymy, and random or "semantically unrelated" relationships between word pairs. Unlike the above tasks, the CogALex-V[3] shared task forbade the use of any thesauri, knowledge bases, or semantic networks (particularly WordNet and ConceptNet), forcing the participating systems to rely on the merit of of the corpus data and their developing system. This CogALex shared task (CogALex VI)[2] has asked for finding synonymy, antonymy and hypernymy relationships amongst word pairs along with undetectable relations as random ones. This one also blocked the usage of of any thesauri, knowledge bases, or semantic networks (particularly WordNet and ConceptNet) so that the system merit should reflect its capability without augmented help. This shared task brought a variant of English only relationship mapping in the Subtask 1 and as well multilingual word mapping as in the Subtask 2.

3 System Description

The system first reads the training data and preprocess it to a structure consumable further in the process flow. We use BERT (Devlin et al., 2018) embeddings to represent each words. BERT adopts the transformer architecture to learn embeddings for words. Since each term can have one or more words, we use an LSTM layer to track the context of the terms. Each of the two terms is first sent into a BERT embedding layer providing the embeddings of the two terms. Afterword each embedding is individually sent into an LSTM layer. The output of the two LSTM layers are concatenated and sent to a convolutional layer. The output of the convolutional layer is then flattened and sent to dense layer and further into a softmax output layer of the model as final phase of training. System produced model trained on training data is used for predicting on test data.

Cross entropy loss has been chosen as the loss function for this solution and a softmax output layer is chosen since multiple semantic relationship classes have to be learnt and predicted.

Figure 1 shows the system architecture deployed for our experimentation.

3.1 Experimental Setup

This section specifies the system specific parameters chosen for experimentation in the model settings section and then the description of the data follows.

3.1.1 Model Settings:

In the deployed BERT (Devlin et al., 2018) model for this experiment, each word is represented by an embedding space of 768 dimensions. The embeddings of the two input terms are concatenated along the row to produce an output of dimension 2 times 768. Further on the concatenated output 2Dconvolution is applied with 9 filters, kernel size of 2 times 2, strides as 1 along each direction and the activation function is chosen as "relu". The output of the convolution layer is flattened and then passed through a dense layer with 256 nodes and then sent to the final output layer.The learning rate for the model is kept at 0.005 after tuning and the batch size for training is kept at 256. The above parameters were chosen after tuning on this experiment keeping in mind of overfitting and underfitting.

3.1.2 Data Description:

For the Subtask 1, an English training data set as well test data set has been provided (Santus et al, 2015), where each data set came in tab separated text file with each row having two words and corresponding relationships in abbreviations of "SYN" (synonymy), "ANT" (antonymy), "HYP" (hypernymy) and "RANDOM". Table 1 shows the distribution of relationship counts in the English dataset.

3.2 Results and Analysis

The participants of CogALex VI[2] were provided with a Python script for the evaluation. The system produced relationship labeled output file from the test data file and was tested with gold standard test file with respect to their precision, recall and F1 score. All these metrics were tested individually and as well as whole and Table 2 depicts them all.

Looking at these result from Table 2 we can see the system did an overall balanced job for synonymy category in terms of precision, recall and F1 score. Noticeably for the other two categories of relation-
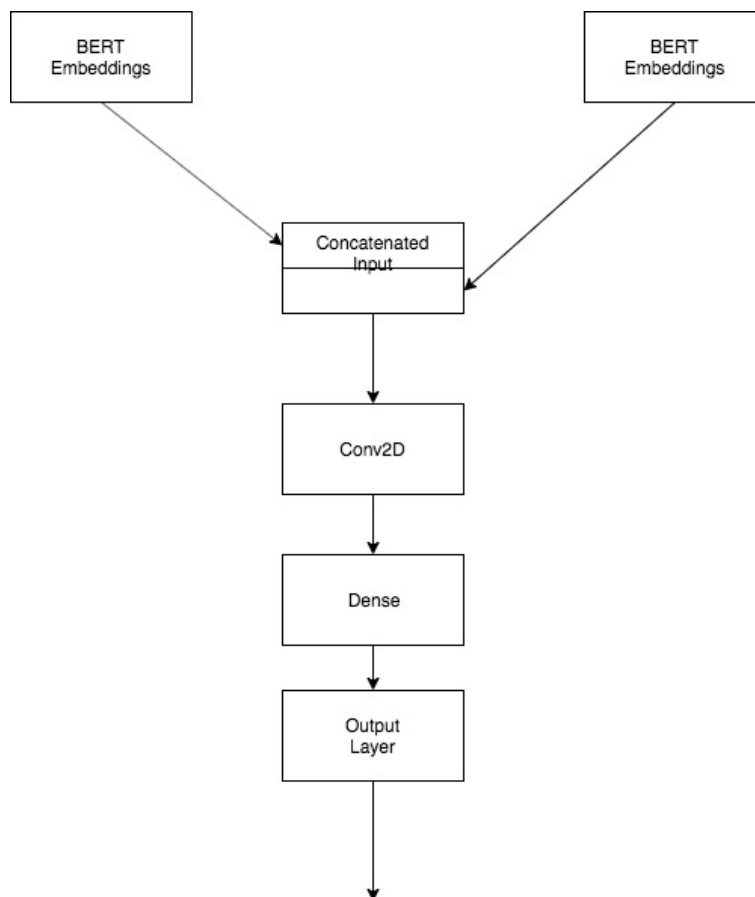
Figure 1: System (Model Training) Architecture

Relationship	Train(Count)	Test-Gold(Count)
SYN	842	266
ANT	916	306
HYP	898	279
RANDOM	2554	887
Total	5210	1738

Table 1: Relationship Distribution of English Data

ships (antonymy and hypernymy) the system had higher precision but lower recall while Table 1 reflects that the support for these two types were higher than synonymy in both the training and gold test set. These numbers reflect that the system does not detect the antonymy and hypernymy as greatly as compared to the synonymy ones. That being said, it has been noticed that for antonymy and hypernymy it doesn't mis-classify as much through false positives. As the system is based on bidirectional transformer model BERT (Devlin et al., 2018), so maybe with BERT embeddings it finds higher support for synonymous words compared to the other two types but whenever it founds any such it grasps well.

The system output shows highest recall for synonyms while the lowest for the hypernyms, whereas the highest precision shows up for antonyms and lowest for synonyms.

We looked into the system-produced relationships with the gold-test-set data for tallying and analysing some error spectrum. For example the implemented system correctly related the word "fiscal" with "commercial" as hypernym and "non financial" as antonym, but it produced synonymy relationship with the word "financial" in place of hypernymy: I believe such intricate examples are pointers for further deep diving. Another example to mention is the relationship between the words "elephant" and "goliath": our

Relationship	Precision	Recall	F1 Score
SYN	0.472	0.417	0.443
ANT	0.654	0.402	0.498
HYP	0.548	0.244	0.337
Overall	0.563	0.355	0.428

Table 2: Precision, Recall and F1 Score for Subtask 1

implemented approach couldn't find any specific kind of relationship and marked as random while in actual it is synonymy. In this last example the word "goliath" is quite of rare use and as our system was primarily depended on the BERT embeddings, therefore the support might have been very less or null in terms of the embedding: this calls for learning using further balanced and rich resources for reference.

4 Conclusion

The current BERT based system demonstrates here a prominent approach for recognizing semantical classes between word pairs. This approach has been applied to the Shared task proposed by CogALex VI[2] on their Subtask 1 for English data to identify synonymy, antonymy, hypernymy or random semantic relationship amongst provided words. These relationships are paradigmatic ones and quite hard for discovering them together. The approach employed here through bidirectional transformer based deep learning network BERT handles this task quite well and produce reasonable precision and recall for each category.

Looking at the difference between the evaluation parameters observed amongst the relationship classes on testing it seems like a thorough investigation should be guided further for better outcomes. To improve the outcomes, introducing variability in deep learning models as well as tuning inside of the model could well have potential. Some specific example based error analysis showed there is ample scope of improvement in the identification of closely related word relationships as well relationship prediction for rarely used words. The exploration of the multiclass semantic relationship mappings could be further interesting if the classes for the experimentation are closely connected ones and many in numbers. Different distributional training data should be exploited further to train model for evaluating the efficacy of such models on predicting such semantic relationships.

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland. Association for Computational Linguistics.

Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. 2015. SemEval-2015 Task 17: Taxonomy Extraction Evaluation (TExEval). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 902–910, Denver, Colorado. Association for Computational Linguistics.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. SemEval-2016 Task 13: Taxonomy Extraction Evaluation (TExEval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1081–1091, San Diego, California. Association for Computational Linguistics.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing Textual Entailment: Models and Applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220, July.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. Language, speech, and communication. MIT Press, Cambridge, Mass.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. SemEval-2007 task 04: classification of semantic relations between nominals. In *Proceedings of the 4th International*

Workshop on Semantic Evaluations - SemEval '07, pages 13–18, Prague, Czech Republic. Association for Computational Linguistics.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Springer US, Boston, MA.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. SemEval-2010 task 8: multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions - DEW '09*, page 94, Boulder, Colorado. Association for Computational Linguistics.

David Jurgens and Mohammad Taher Pilehvar. 2016. SemEval-2016 Task 14: Semantic Taxonomy Enrichment. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1092–1102, San Diego, California. Association for Computational Linguistics.

German Kruszewski, Denis Paperno, and Marco Baroni. 2015. Deriving Boolean structures from distributional vectors. *Transactions of the Association for Computational Linguistics*, 3:375–388, December.

Alessandro Lenci and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 75–79.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do Supervised Distributional Methods Really Learn Lexical Inference Relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976, Denver, Colorado. Association for Computational Linguistics.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods. *Computational Linguistics*, 36(3):341–387, September.

Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. 2016. ccg2lambda: A Compositional Semantics System. In *Proceedings of ACL-2016 System Demonstrations*, pages 85–90, Berlin, Germany. Association for Computational Linguistics.

Saif M. Mohammad, Bonnie J. Dorr, Graeme Hirst, and Peter D. Turney. 2013. Computing lexical contrast. volume 39, page 555–590.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2016. Integrating Distributional Lexical Contrast into Word Embeddings for Antonym-Synonym Distinction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 454–459, Berlin, Germany. Association for Computational Linguistics.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL '06*, pages 113–120, Sydney, Australia. Association for Computational Linguistics.

Bianca Pereira, Cecile Robin, Tobias Daudert, John P. McCrae, Pranab Mohanty, and Paul Buitelaar. 2019. Taxonomy extraction for customer service knowledge base construction. In Maribel Acosta, Philippe Cudré-Mauroux, Maria Maleshkova, Tassilo Pellegrini, Harald Sack, and York Sure-Vetter, editors, *Semantic Systems. The Power of AI and Knowledge Graphs*, pages 175–190, Cham. Springer International Publishing.

Stephen Roller and Katrin Erk. 2016. Relations such as Hypernymy: Identifying and Exploiting Hearst Patterns in Distributional Vectors for Lexical Entailment. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2163–2172, Austin, Texas. Association for Computational Linguistics.

Enrico Santus, Qin Lu, Alessandro Lenci, and Chu-Ren Huang. 2014. Unsupervised antonym-synonym discrimination in vector space. In *Atti della Conferenza di Linguistica Computazionale Italiana*.

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, Beijing, China. Association for Computational Linguistics.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving Hypernymy Detection with an Integrated Path-based and Distributional Method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2389–2398, Berlin, Germany. Association for Computational Linguistics.

Peter D. Turney. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 905–912, Stroudsburg, PA, USA. Association for Computational Linguistics.

Julie Weeds and David Weir. 2003. A general framework for distributional similarity. In *Proceedings of the 2003 conference on Empirical methods in natural language processing -*, volume 10, pages 81–88, Not Known. Association for Computational Linguistics.

A. Zouaq and R. Nkambou. 2008. Building Domain Ontologies from Text for Educational Purposes. *IEEE Transactions on Learning Technologies*, 1(1):49–62, January.