TakeLab at SemEval-2018 Task12: Argument Reasoning Comprehension with Skip-Thought Vectors

Ana Brassard, Tin Kuculo, Filip Boltužić, Jan Šnajder
Text Analysis and Knowledge Engineering Lab
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, 10000 Zagreb, Croatia
{name.surname}@fer.hr

Abstract

This paper describes our system for the SemEval-2018 Task 12: Argument Reasoning Comprehension Task. We utilize skip-thought vectors, sentence-level distributional vectors inspired by the popular word embeddings and the skip-gram model. We encode preprocessed sentences from the dataset into vectors, then perform a binary supervised classification of the warrant that justifies the use of the reason as support for the claim. We explore a few variations of the model, reaching 54.1% accuracy on the test set, which placed us 16th out of 22 teams participating in the task.

1 Introduction

Reasoning is the process of thinking in a logical way to form a conclusion. Inferring conclusions using commonsense reasoning has become a popular topic in NLP. *Textual entailment* (TE) aims to determine whether a *hypothesis* can be inferred from a *premise* (Dagan et al., 2006). Approaches to solving *TE* have ranged from robust approaches based on shallow lexical and semantic features (Marelli et al., 2014) to formal computational semantics approaches based on translating sentences into logical form (Beltagy et al., 2015). The current state-of-the-art approaches to TE use deep learning for *natural logic inference* to capture human deductive reasoning (Bowman et al., 2015; Rocktäschel et al., 2015).

In online discussions, when arguing for or against a stance, people provide arguments leaving their readers to rely on common sense and non-deductive reasoning to evaluate the validity of their arguments. Human annotators can infer the reasons from claims fairly well (Boltužić and Šnajder, 2014; Hasan and Ng, 2014), even when additional (implicit) premises are required to make reasoning deductive (Boltužić and Šnajder, 2016). Habernal et al. (2018) emphasize the importance of implicit premises in argumentation by introducing the *argument reasoning comprehension* task, where one chooses between two mutually exclusive warrants to make a reason warrant the claim. They demonstrate that human experts can perform this task extremely well (up to 90% accuracy).

In this paper, we describe a system for solving the argument reasoning comprehension task, with which we participated in the SemEval-2018 Task 12. Given the reason R and claim C, debate title, debate description, and two warrants, W_1 and W_2, the task is to choose warrant W that justifies the use of R as support for C. For all warrant pairs (W_1, W_2), it holds that if warrant W_1 is W, then W_2 is $\neg W$, which justifies the use of R as support for $\neg C$ and vice versa.

Our system frames the problem as supervised classification and utilizes skip-thought vectors to represent sentences. Our system (TakeLab) ranked 16th out of 22 systems submitted to the SemEval-2018 Task 12, achieving 54.1% accuracy on the test set and 69.0% on the development set.

2 Related Work

Structuring argumentative discussions using Toulmin's argumentation model (Toulmin, 2003) is an established research area in *argumentation mining*, involving detecting claims (Levy et al., 2014; Lippi and Torroni, 2015; Rinott et al., 2015), detecting claim relations (Cabrio and Villata, 2012; Boltužić and Šnajder, 2014; Stab and Gurevych, 2014), and even reconstructing entire argumentation graphs from text (Stab and Gurevych, 2017). While systems have been proposed that tackle some of these problems, they do not as yet provide mechanisms for commonsense reasoning. The *argument comprehension problem*, warranting reasons for claim, explores the problems when the gap between the claim and reason is too wide for textual entailment.

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 1133–1136
New Orleans, Louisiana, June 5–6, 2018.

3 System Description

Our system works in three steps. First, we preprocess the dataset ending up with the claim, reason, both warrant sentences, and the correct warrant sentence label per instance. Second, we utilize *skip-thought vectors* (Kiros et al., 2015) to encode sentences as vectors. Third, we use encoded sentences as features in a supervised classification setup where we predict the warrant label given the encoded sentences.

3.1 Preprocessing

We extract only the warrants, reasons, claims, and labels from the dataset, disregarding optional additional information about the debates. We clean up this data by merging multi-sentence elements in a natural way, i.e., connecting the sentences with the conjunction "and" and modifying the punctuation accordingly. For example, the two-sentence reason:

> *Biking is good for one's health and the environment. It is more expensive to maintain roads than bike lanes.*

becomes a single sentence:

> *Biking is good for one's health and the environment and it is more expensive to maintain roads than bike lanes.*

The motivation behind this is to obtain sentences that convey a single *idea* behind the reason or warrant (claims were always single-sentence). This way, we attempt to extract a *vector per thought*. This results in a consistent set of four sentences per instance.

As the warrant and alternative warrant were extremely similarly worded (68% had two or less different words), we represent warrants W_1 and W_2 as word-level relative complements:

$$W_1' = W_1 \setminus W_2 = \{w_i \in W_1 \mid w_i \notin W_2\}$$
$$W_2' = W_2 \setminus W_1 = \{w_j \in W_2 \mid w_j \notin W_1\}$$

where w_i and w_j denote words. This allowed us to boil down the warrants to their meaningful differences, e.g., just the words *"does"* and *"doesn't"* in instances where the warrants were negated, but otherwise identically worded. We experimented with other combinations, but as they did not lead to performance improvement, we omit them here.

3.2 Skip-Thought Vectors

The skip-thoughts model (Kiros et al., 2015) is a sentence-level abstraction of the skip-gram model (Mikolov et al., 2013). Instead of predicting the surrounding text from a word, it predicts sentences around the target sentence in the text. Kiros et al. (2015) chose to implement an encoder-decoder model, using an RNN encoder with GRU (Chung et al., 2014) activations and an RNN decoder with a conditional GRU. This model is nearly identical to the RNN encoder used by Cho et al. (2014) for machine translation. The encoder-decoder is trained on a large dataset of English books – *BookCorpus* (Zhu et al., 2015), chosen for its abundance of long, context-building sentences. A trained skip-thoughts model can be used as an out-of-the-box encoder-decoder able to convert sentences to skip-thought vectors. The encoder maps words to a sentence vector and the decoder is internally used by the model to generate the surrounding sentences.

The skip-thoughts model was tested on the tasks of semantic relatedness, image-sentence ranking, and paraphrase detection. The last, when combined with basic pairwise statistics, becomes competitive with the state of the art which incorporates much more complicated features and hand-engineering. On the task of semantic relatedness, the model of Kiros et al. (2015) outperformed all previous systems from the SemEval 2014 competition despite its simplicity and the lack of feature engineering. The authors also report good results on a number of classification benchmarks for evaluating sentence representation learning methods.

The model's consistently good results on a variety of tasks motivated us to apply these vectors on our own task, which relies on the interpretation of sentences. We encode the sentences obtained from the preprocessing step into skip-thoughts using Kiros et al. (2015)'s encoder, which gives four feature vectors with 4,800 dimensions with values ranging from -0.2 to 0.2.

3.3 Classification

The final step in our system is classifying instances using an SVM classifier, whose hyperparameters were optimized with a 5-fold cross-validated grid search.[1] We also explored Gaussian Processes, Random Forests, and AdaBoost models, which were all outperformed by the SVM. Input features

[1] We obtained best results on the dev set with: gamma=0.3, C=3.8, degree=2, kernel ='poly'.

Classifier	Features	Dev accuracy score	Test accuracy score
AdaBoost	W_0', W_1', R, C	0.614	0.520
Random forest	W_0', W_1', R, C	0.623	0.498
Gaussian process	W_0', W_1', R, C	0.642	0.547
SVM	W_0, W_1, R, C	0.630	0.538
	W_0', W_1', R, C	0.642	0.536
	$W_0 - W_1, R$	0.661	**0.570**
	$W_0' - W_1', R, C$	0.665	0.561
	$W_0' - W_1', R$	**0.687**	0.552*

Table 1: Accuracy scores of model variants. All models have skip-thought vectors as features, where R stands for reason, C for claim, W_1 and W_2 for warrants, and, $W_1^{'}$ and $W_2^{'}$ for warrant word differences vectors. (* The official results are lower (0.541) due to an error while preparing the output)

are created by concatenating skip-thought vectors obtained in the previous step. We experimented with different variants of features by applying arithmetic operations on the vectors before concatenating them, i.e., calculating the difference between warrants. It should be noted that this difference is calculated as an element-wise subtraction of the vectors, as opposed to the word set difference in the preprocessing step. Furthermore, we experimented with two variations of skip-thought vectors – one with the original warrants intact (W_0, W_1) and one with the warrant subset differences (W_0', W_1').

4 Evaluation

4.1 Dataset Analysis

The dataset consists of 1210 training instances, 317 validation instances, and 445 test instances. Each instance is a tuple $(W_1, W_2, R, C, debateTitle, debateInfo, y)$, with y as the label of the correct warrant (0 for W_1 or 1 for W_2). Among the 1210 training instances, there are 111 different debate titles and 169 different claims, indicating the diversity of the training set. Furthermore, we found that 47.75% of the debate titles had unanimous claims (all for or all against) and 56.69% of the claims were affirmative, but only 21.62% had a balanced number of claims for both sides of the debate (a difference of 10% or less). The debate title *Do We Still Need Libraries?* was the most common debate title, and it had unanimously affirmative claims. Around 35% of the instances contained warrants worded differently, as opposed to being directly negated (by adding *not*). All of this presented a challenge in training the system, since the dataset is small, highly variable, and involves multiple domains.

4.2 Results

The official evaluation measure for this task was the accuracy of the classified instances. In the development phase, the system showed promising results – 0.690 accuracy on the validation set, after training with only the training set. Table 1 shows performances of the model variants we explored. Interestingly, the best results were obtained using the least amount of data – the difference between the modified warrants and only the reason, completely disregarding the claim. On the test set, however, the results were much lower, the official result being 0.541. The final result surprised us, since the system showed good results using various "plain" classifiers without fine-tuning the hyper-parameters (around 0.60). We hypothesize that this was due to overfitting, which was difficult to avoid completely of the small size of the dataset.

5 Conclusion

The argument reasoning comprehension task, recognizing the warrant between a claim and a supporting reason, is a challenging but important task for understanding human reasoning in argumentation. We aim to solve the task by converting sentences into skip-thought vectors and classifying justifying warrants given claims and reasons using an SVM model. This approach showed some promising results in the development stage (69% accuracy), but did not succeed to adequately generalize in order to provide competitive results in the test stage (54% accuracy). Besides using a larger sample for training, this system could be improved by applying transfer learning from other similar tasks, such as paraphrase detection or textual entailment.

Acknowledgment

This research has been partly supported by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS).

References

Islam Beltagy, Stephen Roller, Pengxiang Cheng, Katrin Erk, and Raymond J Mooney. 2015. Representing meaning with a combination of logical form and vectors. *CoRR, abs/1505.06816*.

Filip Boltužić and Jan Šnajder. 2016. Fill the gap! Analyzing implicit premises between claims from online debates. In *Proceedings of the 3rd Workshop on Argument Mining*, pages 124–133.

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Elena Cabrio and Serena Villata. 2012. Natural language arguments: A combined approach. In *Proceedings of the 20th European Conference on Artificial Intelligence*, pages 205–210. IOS Press.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. SemEval-2018 Task 12: The Argument Reasoning Comprehension Task. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, page (to appear), New Orleans, LA, USA. Association for Computational Linguistics.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. *CoRR*, abs/1506.06726.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *IJCAI*, volume 15, pages 185–191.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. SemEval-2014 Task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 1–8.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence-an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Stephen E. Toulmin. 2003. *The uses of argument*. Cambridge University Press.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *CoRR*, abs/1506.06724.