Transformer-based Context-aware Sarcasm Detection in Conversation Threads from Social Media

Xiangjue Dong
Computer Science
Emory University
Altanta, GA, USA
xiangjue.dong@emory.edu

Changmao Li
Computer Science
Emory University
Altanta, GA, USA
changmao.li@emory.edu

Jinho D. Choi
Computer Science
Emory University
Altanta, GA, USA
jinho.choi@emory.edu

Abstract

We present a transformer-based sarcasm detection model that accounts for the context from the entire conversation thread for more robust predictions. Our model uses deep transformer layers to perform multi-head attentions among the target utterance and the relevant context in the thread. The context-aware models are evaluated on two datasets from social media, Twitter and Reddit, and show 3.1% and 7.0% improvements over their baselines. Our best models give the F1-scores of 79.0% and 75.0% for the Twitter and Reddit datasets respectively, becoming one of the highest performing systems among 36 participants in this shared task.

1 Introduction

Sarcasm is a form of figurative language that implies a negative sentiment while displaying a positive sentiment on the surface (Joshi et al., 2017). Because of its conflicting nature and subtlety in language, sarcasm detection has been considered one of the most challenging tasks in natural language processing. Furthermore, when sarcasm is used in social media platforms such as Twitter or Reddit to express users' nuanced intents, the language is often full of spelling errors, acronyms, slangs, emojis, and special characters, which adds another level of difficulty in this task.

Despite of its challenges, sarcasm detection has recently gained substantial attention because it can bring the last gist to deep contextual understanding for various applications such as author profiling, harassment detection, and irony detection (Van Hee et al., 2018). Many computational approaches have been proposed to detect sarcasm in conversations (Ghosh et al., 2015; Joshi et al., 2015, 2016). However, most of the previous studies use the utterances in isolation, which makes it hard even for human to detect sarcasm without the contexts. Thus, it's essential to interpret the target utterances along with contextual information comprising textual features from the conversation thread, metadata about the conversation from external sources, or visual context (Bamman and Smith, 2015; Ghosh et al., 2017; Ghosh and Veale, 2017; Ghosh et al., 2018).

This paper presents a transformer-based sarcasm detection model that takes both the target utterance and its context and predicts if the target utterance involves sarcasm. Our model uses a transformer encoder to coherently generate the embedding representation for the target utterance and the context by performing multi-head attentions (Section 4). This approach is evaluated on two types of datasets collected from Twitter and Reddit (Section 3), and depicts significant improvement over the baseline using only the target utterance as input (Section 5). Our error analysis illustrates that the context-aware model can catch subtle nuance that cannot be captured by the target-oriented model (Section 6).

2 Related Work

Just as most other types of figurative languages are, sarcasm is not necessarily complicated to express but requires comprehensive understanding in context as well as commonsense knowledge rather than its literal sense (Van Hee et al., 2018). Various approaches have been presented for this task.

Most earlier works had taken the target utterance without context as input. Both explicit and implicit incongruity features were explored in these works (Joshi et al., 2015). To detect whether certain words in the target utterance involve sarcasm, several approaches based on distributional semantics were proposed (Ghosh et al., 2015). Additionally, word embedding-based features like distance-weighted similarities were also adapted to capture the subtle forms of context incongruity (Joshi et al., 2016). Nonetheless, it is difficult to detect sarcasm by considering only the target utterances in isolation.

Proceedings of the Second Workshop on Figurative Language Processing, pages 276–280

https://doi.org/10.18653/v1/P17

Non-textual features such as the properties of the author, audience and environment were also taken into account (Bamman and Smith, 2015). Both the linguistic and context features were used to distinguish between information-seeking and rhetorical questions in forums and tweets (Oraby et al., 2017). Traditional machine learning methods such as Support Vector Machines were used to model sarcasm detection as a sequential classification task over the target utterance and its surrounding utterances (Wang et al., 2015). Recently, deep learning methods using LSTM were introduced, considering the prior turns (Ghosh et al., 2017) as well as the succeeding turns (Ghosh et al., 2018).

3 Data Description

Given a conversation thread, either from Twitter or Reddit, a target utterance is the turn to be predicted, whether or not it involves sarcasm, and the context is an ordered list of other utterances in the thread. Table 1 shows the examples of conversation threads where the target utterances involve sarcasm.[1]

	Utterance
C1	This feels apt this morning but I don't feel fine ... <URL>
C2	@USER it is what's going round in the heads of many I know ...
T	@USER @USER I remember a few months back we were saying the Americans shouldn't tell us how to vote on brexit

(a) Sarcasm example from Twitter.

	Utterance
C1	Promotional images for some guy's Facebook page
C2	I wouldn't let that robot near me
T	Sounds like you don't like science, you theist sheep

(b) Sarcasm example from Reddit.

Table 1: Examples of the conversation threads where the target utterances involve sarcasm. C_i: i'th utterance in the context, T: the target utterance.

The Twitter data is collected by using the hashtags #sarcasm and #sarcastic. The Reddit data is a subset of the Self-Annotated Reddit Corpus that consists of 1.3 million sarcastic and non-sarcastic posts (Khodak et al., 2017). Every target utterance is annotated with one of the two labels, SARCASM and NOT_SARCASM. Table 2 shows the statistics of the two datasets provided by this shared task.

[1]Note that the target utterance can appear at any position of the context although its exact position is not provided in this year's shared task data.

Notice the huge variances in the utterance lengths for both the Twitter and the Reddit datasets. For the Reddit dataset, the average lengths of conversations as well as utterances are significantly larger in the test set than the training set that potentially makes the model development more challenging.

	NC	AU	AT
TRN	5,000	4.9 (±3.2)	140.4 (±112.8)
TST	1,800	4.2 (±1.9)	128.5 (±78.8)

(a) Twitter dataset statistics.

	NC	AU	AT
TRN	4,400	3.5 (±0.8)	45.8 (±17.3)
TST	1,800	5.3 (±2.0)	93.6 (±57.8)

(b) Reddit dataset statistics.

Table 2: Statistics of the two datasets provided by the shared task. TRN: training set, TST: test set, NC: # of conversations, AU: Avg # of utterances per conversation (including the target utterances) and its stdev, AT: Avg # of tokens per utterance and its stdev.

4 Approach

Two types of transformer-based sarcasm detection models are used for our experiments:

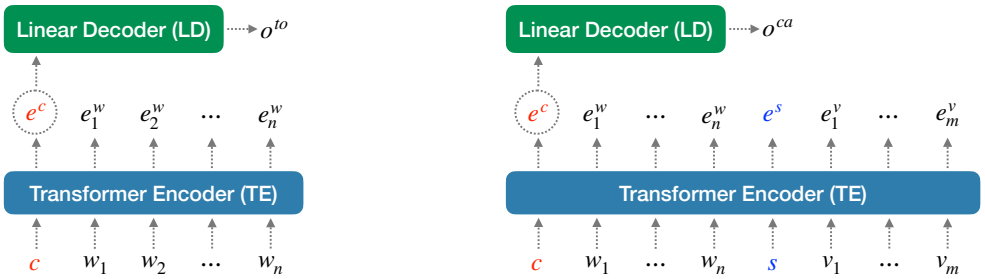
a) The target-oriented model takes only the target utterance as input (Section 4.1).

b) The context-aware model takes both the target utterance and the context utterances as input (Section 4.2).

These two models are coupled with the latest transformer encoders e.g., BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2020), and ALBERT (Lan et al., 2019), and compared to evaluate how much impact the context makes to predict whether or not the target utterance involves sarcasm.

4.1 Target-oriented Model

Figure 1a shows the overview of the target-oriented model. Let $W = \{w_1, \ldots, w_n\}$ be the input target utterance, where w_i is the i'th token in W and n is the max-number of tokens in any target utterance. W is first prepended by the special token c representing the entire target utterance, which creates the input sequence $I^{to} = \{c\} \oplus W$. I^{to} is then fed into the transformer encoder, which generates the sequence of embeddings $\{e^c\} \oplus E^w$, where $E^w = \{e^w_1, \ldots, e^w_n\}$ is the embedding list for W and (e^c, e^w_i) are the embeddings of (c, w_i) respectively. Finally, e^c is fed into the linear decoder to generate the output vector o^{to} that makes the binary decision of whether or not W involves sarcasm.



(a) Target-oriented model (Section 4.1)

(b) Context-aware model (Section 4.2)

Figure 1: The overview of our transformer-based target-oriented and context-aware models.

4.2 Context-aware Model

Figure 1b shows the overview of the context-aware model. Let L_i be the i'th utterance in the context. Then, $V = L_1 \oplus \cdots \oplus L_k = \{v_1, \ldots, v_m\}$ is the concatenated list of tokens in all context utterances, where k is the number of utterances in the context, v_1 is the first token in L_1 and v_m is the last token in L_k. The input sequence I^{to} from Section 4.1 is appended by the special token s representing the separator between the target utterance and the context, and also V, which creates the input sequence $I^{ca} = I^{to} \oplus \{s\} \oplus V$. Then, I^{ca} gets fed into the transformer encoder, which generates a sequence of embeddings $\{e^c\} \oplus E^w \oplus \{e^s\} \oplus E^v$, where $E^v = \{e_1^v, \ldots, e_m^v\}$ is the embedding list for V, and (e^s, e_i^v) are the embeddings of (s, v_i) respectively. Finally, e^c is fed into the linear decoder to generate the output vector o^{ca} that makes the same binary decision to detect sarcasm.

5 Experiments

5.1 Data Split

For all our experiments, a mixture of the Twitter and the Reddit datasets is used. The Twitter training set provided by the shared task consists of 5,000 tweets, where the labels are equally balanced between `SARCASM` and `NOT_SARCASM` (Table 2). We find, however, 4.82% of them are duplicates, which are removed before data splitting. As a result, 4,759 tweets are used for our experiments. Labels in the Reddit training set are also equally balanced and no duplicate is found in this dataset.

	Twitter		Reddit	
	TRN	**DEV**	**TRN**	**DEV**
SARCASM	2,020	239	1,973	227
NOT_SARCASM	2,263	237	1,987	213

Table 3: Statistics of the data split used for our experiments, where 10% of each dataset is randomly selected to create the development set.

5.2 Models

Three types of transformers are used for our experiments, that are BERT-Large (Devlin et al., 2019), RoBERTa-Large (Liu et al., 2020), and ALBERT-xxLarge (Lan et al., 2019), to compare the performance among the current state-of-the-art encoders. Every model is run three times and their average scores as well as standard deviations are reported. All models are trained on the combined Twitter + Reddit training set and evaluated on the combined development set (Table 3).

5.3 Experimental Setup

After an extensive hyper-parameter search, we set the learning rate to 3e-5, the number of epochs to 30, and use different seed values, 21, 42, 63, for the three runs. Additionally, based on the statistics of each dataset, we set the maximum sequence length to 128 for the target-oriented models while it is set to 256 for the context-aware models by considering the different lengths of the input sequences required by those approaches.

5.4 Results

The baseline scores are provided by the organizers, that are 60.0% for Reddit and 67.0% for Twitter using the single layer LSTM attention model (Ghosh et al., 2018). Table 4 shows the results achieved by our target-oriented (Section 4.1) and the context-aware (Section 4.2) models on the combined development set. The RoBERTa-Large model gives the highest F1-scores for both the target-oriented and context-aware models. The context-aware model using RoBERTa-Large show an improvement of 1.1% over its counterpart baseline so that this model is used for our final submission to the shared task. Note that it may be possible to achieve higher performance by fine-tuning hyperparameters for the Twitter and Reddit datasets separately, which we will explore in the future.

	P	R	F1
B-L	77.3 (±0.6)	79.9 (±0.8)	78.6 (±0.1)
R-L	73.4 (±0.6)	88.5 (±1.4)	**80.2** (±0.5)
A-XXL	76.1 (±1.4)	83.3 (±2.3)	79.5 (±0.2)

(a) Results from the target-oriented models (Section 4.1).

	P	R	F1
B-L	76.3 (±1.0)	82.7 (±1.6)	79.4 (±0.5)
R-L	77.3 (±3.8)	86.1 (±4.0)	**81.3** (±0.2)
A-XXL	76.5 (±3.3)	82.7 (±3.1)	79.4 (±2.2)

(b) Results from the context-aware models (Section 4.2).

Table 4: Results on the combined Twitter+Reddit development set. B-L: BERT-Large, R-L: RoBERTa-Large, A-XXL: ALBERT-xxLarge.

Table 5 shows the results by the RoBERTa-Large models on the test sets. The scores are retrieved by submitting the system outputs to the shared task's CodaLab page.[2] The context-aware models significantly outperform the target-oriented models on the test sets, showing improvements of 3.1% and 7.0% on the F1 scores for the Twitter and the Reddit datasets, respectively. The improvement on Reddit is particularly substantial due to the much greater lengths of the conversation threads and utterances in the test set compared to the ones in the training set (Table 2). As the final results, we achieve 79.0% and 75.0% for the Twitter and Reddit datasets respectively that mark the 2nd places for both datasets at the time of the submission.

	P	R	F1
Twitter	75.5 (±0.7)	76.4 (±0.6)	75.2 (±0.8)
Reddit	67.9 (±0.5)	69.2 (±0.7)	67.4 (±0.5)

(a) Results from the target-oriented RoBERTa-Large models.

	P	R	F1
Twitter	78.4 (±0.6)	78.9 (±0.3)	**78.3 (±0.7)**
Reddit	74.5 (±0.6)	74.9 (±0.5)	**74.4 (±0.7)**

(b) Results from the context-aware RoBERTa-Large models.

Table 5: Results on the test sets from CodaLab.

6 Analysis

For a better understanding in our final model, errors from the following three situations are analyzed (TO: target-oriented, CA: context-aware):

- TwCc: TO is wrong and CA is correct.
- TcCw: TO is correct and CA is wrong.
- TwCw: Both TO and CA are wrong.

[2] https://competitions.codalab.org/competitions/22247

Table 6 shows examples for every error situation. For TwCc, TO predicts it to be NOT_SARCASM. In this example, it is difficult to tell if the target utterance involves sarcasm without having the context. For TcCw, CA predicts it to be NOT_SARCASM. It appears that the target utterance is long enough to provide enough features for TO to make the correct prediction, whereas considering the extra context may increase noise for CA to make the incorrect decision. For TwCw, both TO and CA predict it to be NOT_SARCASM. This example seems to require deeper reasoning to make the correct prediction.

	Utterance
C_1	who has ever cared about y * utube r * wind .
C_2	@USER Back when YouTube was beginning it was a cool giveback to the community to do a super polished high production value video with YT talent . Not the same now . The better move for them would be to do like 5-6 of them in several categories to give that shine .
T	@USER @USER I look forward to the eventual annual Tubies Awards livestream .

(a) Example when TO is wrong and CA is correct.

	Utterance
C_1	I am asking the chairs of the House and Senate committees to investigate top secret intelligence shared with NBC prior to me seeing it.
C_2	@USER Good for you, sweetie! But using the legislative branch of the US Government to fix your media grudges seems a bit much.
T	@USER @USER @USER you look triggered after someone criticizes me, are conservatives skeptic of ppl in power?

(b) Example when TO is correct and CA is wrong.

	Utterance
C_1	If I could start my #Brand over, this is what I would emulate my #Site to look like .. And I might, once my anual contract with #WordPress is up . Even tho I don't think is very; I can't help but to find ... <URL> <URL>
C_2	@USER There is no design on it except for links ?
T	@USER It's the of what #Works in this current #Mindset of #MassConsumption; wannabe fast due to caused by, and being just another and. is the light, bringing color back to this sad world of and.

(c) Example when both TO and CA are wrong.

Table 6: Examples of the three error situations. C_i: i'th utterance in the context, T: the target utterance.

7 Conclusion

This paper explores the benefit of considering relevant contexts for the task of sarcasm detection. Three types of state-of-the-art transformer encoders are adapted to establish the strong baseline for the target-oriented models, which are compared to the context-aware models that show significant improvements for both Twitter and Reddit datasets and become one of the highest performing models in this shared task.

All our resources are publicly available at Emory NLP's open source repository: https://github.com/emorynlp/figlang-shared-task-2020

Acknowledgments

We gratefully acknowledge the support of the AWS Machine Learning Research Awards (MLRA). Any contents in this material are those of the authors and do not necessarily reflect the views of AWS.

References

David Bamman and Noah Smith. 2015. Contextualized Sarcasm Detection on Twitter. In *International AAAI Conference on Web and Social Media*, pages 574–577.

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Aniruddha Ghosh and Tony Veale. 2017. Magnets for Sarcasm: Making Sarcasm Detection Timely, Contextual and Very Personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 482–491, Copenhagen, Denmark. Association for Computational Linguistics.

Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. 2017. The Role of Conversation Context for Sarcasm Detection in Online Interactions. *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 186–196.

Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. 2018. Sarcasm Analysis using Conversation Context. *Comput. Linguist.*, 44(4):755–792.

Debanjan Ghosh, Weiwei Guo, and Smaranda Muresan. 2015. Sarcastic or Not: Word Embeddings to Predict the Literal or Sarcastic Meaning of Words. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1003–1012, Lisbon, Portugal. Association for Computational Linguistics.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic Sarcasm Detection: A Survey. *ACM Computing Surveys*, 50(5):1–22.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing Context Incongruity for Sarcasm Detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, Beijing, China. Association for Computational Linguistics.

Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman. 2016. Are Word Embedding-based Features Useful for Sarcasm Detection? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1006–1011, Austin, Texas. Association for Computational Linguistics.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2017. A Large Self-Annotated Corpus for Sarcasm. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, abs/1704.05579.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv*, 11942(1909).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In *Proceedings of the International Conference on Learning Representations*.

Shereen Oraby, Vrindavan Harrison, Amita Misra, Ellen Riloff, and Marilyn Walker. 2017. Are you serious?: Rhetorical Questions and Sarcasm in Social Media Dialog. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 310–319, Saarbrücken, Germany. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 Task 3: Irony Detection in English Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Zelin Wang, Zhijian Wu, Ruimin Wang, and Yafeng Ren. 2015. Twitter Sarcasm Detection Exploiting a Context-Based Model. In *WISE*.