Multitask Learning of Negation and Speculation using Transformers

Aditya Khandelwal
College of Engineering Pune
Pune, India
khandelwalar16.comp
@coep.ac.in

Benita Kathleen Britto
Veermata Jijabai Technological Institute
Mumbai, India
bcbritto_b16
@it.vjti.ac.in

Abstract

Detecting negation and speculation in language has been a task of considerable interest to the biomedical community, as it is a key component of Information Extraction systems from Biomedical documents. Prior work has individually addressed Negation Detection and Speculation Detection, and both have been addressed in the same way, using a 2 stage pipelined approach: Cue Detection followed by Scope Resolution. In this paper, we propose Multitask learning approaches over 2 sets of tasks: Negation Cue Detection & Speculation Cue Detection, and Negation Scope Resolution & Speculation Scope Resolution. We utilise transformer-based architectures like BERT, XLNet and RoBERTa as our core model architecture, and finetune these using the Multitask learning approaches. We show that this Multitask Learning approach outperforms the single task learning approach, and report new state-of-the-art results on Negation and Speculation Scope Resolution on the BioScope Corpus and the SFU Review Corpus.

1 Introduction

Detection of linguistic phenomena like Negation and Speculation are key components of Biomedical Information Retrieval systems, as they significantly alter the meaning of a sentence. While detecting these are also useful in Sentiment Analysis systems, and systems used to determine the veracity of information, their primary use is in biomedical systems. Thus, these tasks have attracted significant interest from researchers over the years, and due to the similarity between these tasks, similar approaches have been used to address them, and parallel corpora containing annotations for both Negation and Speculation have also been created, including:

- BioScope Corpus (Szarvas et al., 2008)
- SFU Review Corpus (Konstantinova et al.)

Prior research has converged to using a 2 stage approach for both Negation Detection and Speculation Detection: Cue Detection and Scope Resolution, and solved each task independently. These subtasks and their relevance to the Biomedical domain can be better understood using the following example:

(1) We found that T cells were [not] present, [perhaps] indicating an immuno deficiency.

Cue Detection involves finding the word(s) that express the linguistic phenomena being detected. In the example given above, *not* is the negation cue, as it expresses the negation in the sentence. Similarly, *perhaps* is the speculation cue.

Scope Resolution involves finding the word(s) that were affected by the cue word of the linguistic phenomena being considered. In the example above, the underlined words outline the scope for each cue. Specifically, for the negation cue *not*, the word *present* was negatively affected by it. Similarly, for the speculation cue *perhaps*, the words *indicating an immuno deficiency* were affected by it, indicating that these words have an associated uncertainty.

The approaches addressing these tasks have varied significantly over the years, with recent work focusing on using transformer-based architectures to perform transfer learning, and have given the best results to date on Scope Resolution. On Cue Detection, they yield the best performance among neural models, but due to the small dataset sizes, the best performance is still given by rule-based heuristic approaches.

Despite the similarity among the subtasks, prior systems have looked at these tasks independently. We believe that a system can improve performance on both tasks by learning from both simultaneously, due to the similarity, which is what Multitask Learning is about.

Multitask Learning involves jointly training the same architecture to perform multiple tasks. It

Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, pages 79–87
November 20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

relies on the concept of using shared knowledge between both tasks, which is what the model is forced to learn to perform well at all tasks, eventually leading to better performance on all tasks. For neural models, this is especially useful, as the lower layers can share the same input representation, thus getting more data to learn better lower level features from the input, and a task specific layer final layer can learn the task specific features.

In this paper, inspired by the success of transformers, we propose a method to perform Multitask Learning of negation and speculation using transformer-based architectures. We explore a few design choices, and analyse the impact of these design choices. We show that our approach provides significant benefits over the normal independently trained version. We also make all our code publicly available[1] . This paper is structured as follows: Section 2 contains a brief Literature Review, Section 3 describes the Methodology in detail, Section 4 talks about the Experimentation Details, Section 5 contains the Results and their Analysis, and Section 6 contains the Conclusion and Future Scope.

2 Literature Review

Over the years, methods addressing these subtasks have ranged from simple whitelists based on frequency (rule-based), to traditional Machine Learning algorithms like SVMs, neural models like BiLSTMS and transformer based models.

Khandelwal and Sawant (2020) provide an extensive literature review of the methods for Negation Cue Detection and Scope Resolution. For Speculation Cue Detection, most methods used were similar to the methods used for Negation Cue Detection. Below, we summarise a few papers that addressed Speculation Cue Detection and Scope Resolution.

2.1 Traditional Machine Learning Methods

Özgür and Radev (2009) used a Support Vector Machine (SVM) with a linear kernel to detect speculation cues. Once the speculation cues were identified, the parts-of-speech tags and the syntactic structure were used for scope resolution.

Morante and Daelemans (2009) used the IGTREE classifier with the help of gain ratio (TiMBL implementation) to classify the cues. For scope resolution, three classifiers were used to classify if a token was the first token in the scope sequence (F-scope), the last (L-scope), or neither. These three classifiers were Memory-based learning (as implemented in TiMBL), SVM and Conditional Random Field (CRF). A fourth classifier, the metalearner, used the output of the three classifiers to predict the scope classes.

Velldal et al. (2010) used a Maximum Entropy Classification approach to find the speculation cue. For scope resolution, they used a rule-based approach. The rules operated on the dependency structure of the parser (MaltParser and XLE).

Kilicoglu and Bergler (2008) used linguistic knowledge to detect speculation cues. This was achieved by using a semi-automatic lexical acquisition strategy as well as by using a dictionary of weighted speculation cues. In a follow-up paper (Kilicoglu and Bergler, 2010), they improved on their previous work with the help of vagueness quantifiers and syntactic dependency relations. Cues were detected with rules that operate on lexical information and syntactic information obtained from the Stanford Lexical Parser. The scopes of the cues were detected with the help of the Stanford Lexical Parser and dependency-based heuristics.

Velldal (2011) compiled a list of words that were observed to be cues in the training data, under the assumption that speculation cues can be treated as a closed class. He then checked occurrences of these words in the test data via a large-margin SVM classifier to determine whether they were a cue or not.

A CRF based approach was used in (Tang et al., 2010) to identify the hedge cues and their scopes in sentences. A CRF and large margin-based model were trained simultaneously. Their outputs were provided to another CRF model to get the cues of sentences. These cues were passed to another CRF model followed by post processing, to detect scopes of the given cues.

Morante et al. (2010) described a memory based approach (IGTree as implemented in TiMBL) for cue detection. For scope resolution, a memory-based approach was used with the help of syntactic dependencies of a sentence.

Read et al. (2011) described a methodology to resolve the scope of a sentence using an SVM based constituent ranker. The scope of a cue was assumed to be a constituent. Three broad classes of rules were used to extract features from the parse trees. The parse trees containing the cue were fed to the SVM-based ranker to output a ranked order of parse

[1] adityak6798.github.io

trees which was then declared to be the scope of the cue.

Velldal et al. (2012) used an SVM classifier based on manually defined rules for speculation cue detection. For scope resolution, they experimented with three architectures: a rule-based system that used Data Driven Dependency Parsing to generate dependency structures, an SVM Ranker for selecting subtrees in the constituent structures obtained via a Grammar-Driven Phrase Structure Parser and hybrid of both the above systems.

The approach of (Moncecchi et al., 2012) was based on CRF and usage of domain knowledge. The task of cue detection was solved by using the sequential classifier of CRF. The scope of these cues was resolved by using the CRF at the initial stage with a window size of two. Later, domain knowledge was used to incorporate rules in the system, which showed an improvement in performance of the system.

Cruz et al. (2016) used a classifier-based approach for speculation cue detection and scope resolution. The features for the classifier were manually defined. For cue detection, an SVM-based classifier was used to predict the BIO tags. The scope was also identified using an SVM-based classifier to predict the in-scope and out-of-scope tags when the cues and tokens were provided as input to the classifier. A Radial Basis Function (RBF) Kernel was used with Cost Sensitive Learning to handle imbalanced classes.

2.2 Deep Learning Methods

Qian et al. (2016) used a CNN based approach to re-solve the scope of a speculation cue. The CNN framework took as input position and path features.

Fei et al. (2020) used a Recursive Neural Network (RecurNN) followed by a CRF to detect the scope in a sentence which is named as the Recur-CRF model. The dependency tree based RecurNN learnt a high-level representation of words in the given content. The output of the RecurNN was given to the CRF to fully under-stand the contextual information required to predict the scope of a given cue.

Recently, Britto and Khandelwal (2020) extended the approach by Khandelwal and Sawant (2020), who used BERT (Devlin et al., 2018) to address Negation Cue Detection and Scope Resolution. They experimented with using various transformer-based architectures (BERT, XLNet (Yang et al., 2019) and RoBERTa (Liu et al., 2019)), and jointly training on multiple datasets to address speculation cue detection and scope resolution. This approach gave the best results to date on Negation and Speculation Scope Resolution.

2.3 Multitask Learning using Negation Scope Resolution

We also review a couple of papers which have used Multitask Learning with Negation Scope Resolution as one of the many tasks to jointly train the model. It is important to note that these paradigms were explored to improve performance in the auxiliary tasks the model was trained, which were almost always harder than Negation Scope Resolution.

Bhatia et al. (2018) perform joint entity extraction and negation detection for biomedical articles. Initially, they used a hierarchical encoder-decoder model used for Named Entity Recognition (NER), and adapted it for the Multitask setting by sharing the encoder, but using separate decoders for both the tasks. To overcome the overparameterization during low-resource settings, they propose usage of a conditional softmax shared decoder, where instead of using 2 different decoder architectures, they shared the decoder as well, and only had separate classification heads. They also feed the output of the NER head as an additional input to the negation head, which helps improve the performance. They use BiLSTMs for both the encoder and decoder.

Barnes et al. (2019) explore another joint task that has been explored often, namely Sentiment Analysis systems that are jointly trained with Negation Detection systems. They mention that since Sentiment Analysis is a harder task than negation detection, and negation data is used as a task in the pipeline for sentiment analysis, they perform selective sharing of LSTM layers, and use negation as an auxillary task on which the sentiment analysis system is trained. Specifically, they use a separate CRF tagger for negation detection on the outputs of an intermediate layer for the sentiment analysis system, whose final layer is used for sentiment classification. They use a BiLSTM-based network.

3 Methodology

Similar to (Khandelwal and Sawant, 2020) and (Britto and Khandelwal, 2020), we use the transformer model (BERT/XLNet/RoBERTa) with a classification head as our base model. To jointly train the model, we propose the following additions to the model.

3.1 Cue Detection

For Cue Detection, we use 2 separate classification heads for Negation Cue Detection and Speculation Cue Detection respectively. The architecture can be visualized as in Figure 1. We feed an input sentence to the model, and use the output corresponding to the task we are looking to perform. This architecture halves the number of parameters and inference time if we want to perform negation and speculation detection simultaneously.

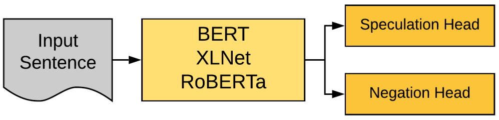


Figure 1: Multitask Cue Detection (Model Overview)

To train this model, we only train on those sentences that have both negation and speculation cue labels. Since we train on the BioScope Corpus, and the SFU Review Corpus, all training samples have labels for both negation and speculation. A single input sentence is fed, and the model is trained on the losses computed for both heads, negation and speculation.

3.2 Scope Resolution

For Scope Resolution, we use the same classification head for both Negation and Speculation Scope Resolution, and use preprocessing techniques to implicitly tell the model which task to perform.

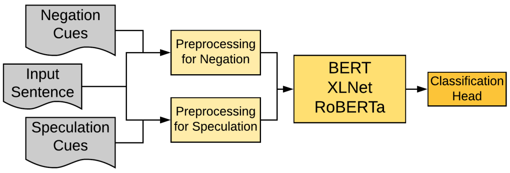


Figure 2: Multitask Scope Resolution (Model Overview)

For Scope Resolution, we need to represent the cue words in the input sentence for which we want to find the scope. This could be done via the Augment Preprocessing method used by (Britto and Khandelwal, 2020). This involves appending a special token before the cue word in the input sentence which represents the type of cue word. The types of cue words considered are:

- Single Word Cue: tok[0]
- Part of a Multiword Cue: tok[1]
- Affix (Suffix / Prefix): tok[2]

Consider the following example:

Input Sentence: *It seems that the treatment is not successful.*
Negation Cues: *not*
Preprocessed Sentence: *It seems that the treatment is tok[0] not successful.*

To jointly train the same model to make predictions, we have to also tell the model which task we expect it to perform. To do this, we propose the following methods which are slight modifications of the Augment preprocessing method:

- Global: We represent the task by appending the name of the task to be performed at the end of the input sentence followed by a [SEP] token. The cue words for both negation and speculation are represented by the same set of special tokens. Specifically,

  Input Sentence: *It seems that the treatment is not successful.*
  Negation Cues: *not*
  Speculation Cues: *seems*
  Input Sentence for Negation: *It seems that the treatment is tok[0] not successful [SEP] Negation.*
  Input Sentence for Speculation: *It tok[0] seems that the treatment is not successful [SEP] Speculation.*

  Thus, the type of cue for both negation and speculation is the same (single word cue), hence we use the same token (tok[0]) to augment the input sentence. The task is represented by appending the task name to the end of the sentence.

- Local: Here, we use the following tokens to represent the different types of negation and speculation cues.
  - Negation-Single Word Cue: tok[0]

- Negation-Part of a Multiword Cue: tok[1]
- Negation-Affix (Suffix / Prefix): tok[2]
- Speculation-Single Word Cue: tok[4]
- Speculation-Part of a Multiword Cue: tok[5]
- Speculation-Affix (Suffix / Prefix): tok[6]

Specifically,

Input Sentence: *It seems that the treatment is not successful.*
Negation Cues: *not*
Speculation Cues: *seems*
Input Sentence for Negation: *It seems that the treatment is tok[0] not successful.*
Input Sentence for Speculation: *It tok[4] seems that the treatment is not successful.*

Here, the tokens used to represent different types of negation cues are different than the tokens used to represent the different types of speculation cues, thus implicitly telling the model which scope it has to find.

4 Experimentation Details

We perform experimentation on the following datasets:

- BioScope Corpus:
  - BioScope Abstracts (BA) SubCorpora
  - BioScope Full Papers (BF) SubCorpora
- SFU Review Corpus (SFU)

We believe that by training on multiple datasets, the overfitting of the models can reduce, as the datasets are fairly small in size (200-2000 samples), despite the different domains of the datasets (BioScope Corpora is from the Biomedical Domain, and SFU Review Corpus contains general online review text). Hence, we also experiment with training the models on multiple datasets, and testing on the individual datasets.

We use a 70-15-15 train-dev-test split. The results are reported as an average of 5 runs for training on a single dataset and an average of 3 runs for training on a combination of multiple datasets. We report the Macro F1 Average (Token-level) score for both Cue Detection and Scope Resolution.

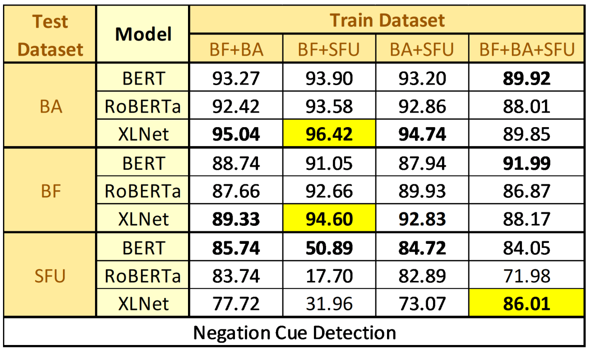
We perform an early stopping (with a patience of 6) on the validation F1 Score. Since we jointly train 2 tasks, we experiment with these 2 ways to perform early stopping:

- Separate: Here, we use 2 early stopping counters: One for Negation and one for Speculation. Specifically, we have separate validation sets for Negation and Speculation, and for each validation set, we run a different Early Stopping Counter. Thus, the final models for Negation and Speculation differ, although they are trained jointly.
- Combined: Here, there is only one Early Stopping used. Training is stopped when the average of the validation F1 scores on the Negation Validation set and the Speculation Validation set do not improve for 6 epochs. Here, we have the same final model for both Negation and Speculation.

We train the models using GPUs available via Google Colaboratory. The code is publicly available.

5 Results and Analysis

To perform a better comparison of independently trained models on multiple datasets, we train BERT, XLNet and RoBERTa on BF+BA, BF+SFU, BA+SFU and BF+BA+SFU for Negation Cue Detection and Negation Scope Resolution. We train the model as per the paper by Britto and Khandelwal (2020), and average the results of 3 runs. The results are shown in Tables 1 and 2. An analysis of the results shown below is done in Section 5.4.

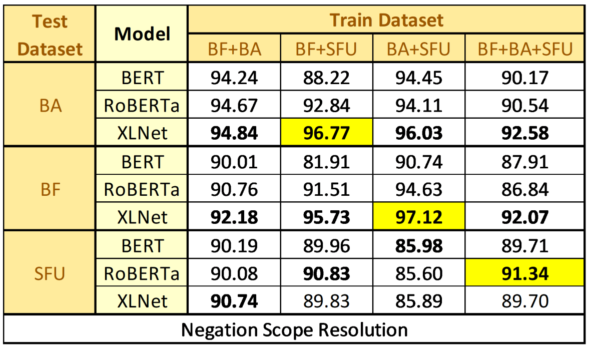


Test Dataset	Model	Train Dataset			
		BF+BA	BF+SFU	BA+SFU	BF+BA+SFU
BA	BERT	93.27	93.90	93.20	**89.92**
	RoBERTa	92.42	93.58	92.86	88.01
	XLNet	**95.04**	**96.42**	**94.74**	89.85
BF	BERT	88.74	91.05	87.94	**91.99**
	RoBERTa	87.66	92.66	89.93	86.87
	XLNet	**89.33**	**94.60**	**92.83**	88.17
SFU	BERT	**85.74**	**50.89**	**84.72**	84.05
	RoBERTa	83.74	17.70	82.89	71.98
	XLNet	77.72	31.96	73.07	**86.01**
Negation Cue Detection					

Table 1: Negation Cue Detection (Trained on Multiple Datasets)

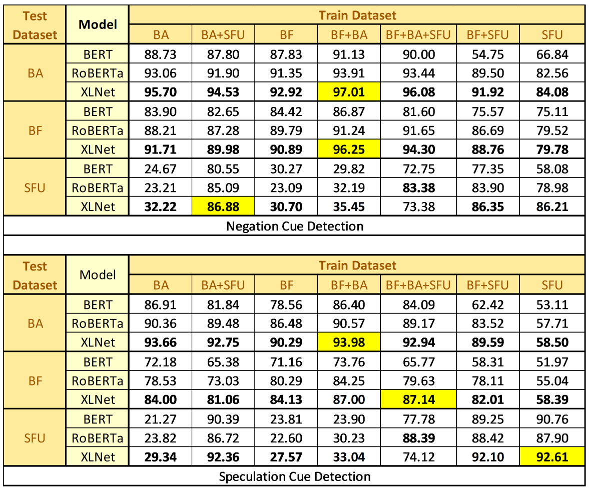
5.1 Cue Detection

The results for Negation and Speculation Cue Detection are shown in Table 3 (trained using the Com-



Test Dataset	Model	Train Dataset			
		BF+BA	BF+SFU	BA+SFU	BF+BA+SFU
BA	BERT	94.24	88.22	94.45	90.17
	RoBERTa	94.67	92.84	94.11	90.54
	XLNet	**94.84**	**96.77**	**96.03**	**92.58**
BF	BERT	90.01	81.91	90.74	87.91
	RoBERTa	90.76	91.51	94.63	86.84
	XLNet	**92.18**	**95.73**	**97.12**	**92.07**
SFU	BERT	90.19	89.96	**85.98**	89.71
	RoBERTa	90.08	**90.83**	85.60	**91.34**
	XLNet	**90.74**	89.83	85.89	89.70
Negation Scope Resolution					

Table 2: Negation Scope Resolution (Trained on Multiple Datasets)

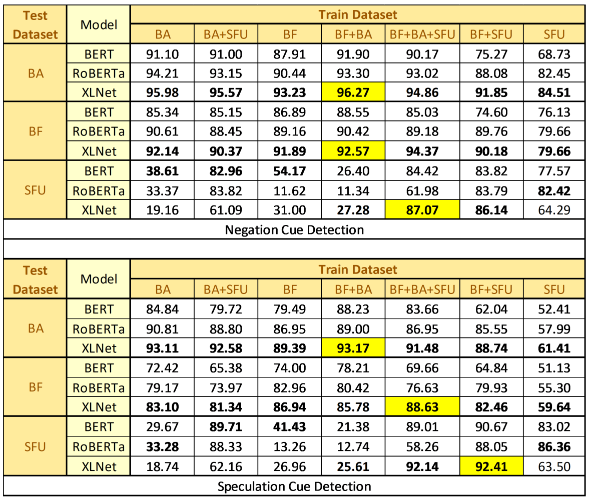


Test Dataset	Model	Train Dataset						
		BA	BA+SFU	BF	BF+BA	BF+BA+SFU	BF+SFU	SFU
BA	BERT	88.73	87.80	87.83	91.13	90.00	54.75	66.84
	RoBERTa	93.06	91.90	91.35	93.91	93.44	89.50	82.56
	XLNet	**95.70**	**94.53**	**92.92**	**97.01**	**96.08**	**91.92**	**84.08**
BF	BERT	83.90	82.65	84.42	86.87	81.60	75.57	75.11
	RoBERTa	88.21	87.28	89.79	91.24	91.65	86.69	79.52
	XLNet	**91.71**	**89.98**	**90.89**	**96.25**	**94.30**	**88.76**	**79.78**
SFU	BERT	24.67	80.55	30.27	29.82	72.75	77.35	58.08
	RoBERTa	23.21	85.09	23.09	32.19	**83.38**	83.90	78.98
	XLNet	**32.22**	**86.88**	**30.70**	**35.45**	73.38	**86.35**	**86.21**
Negation Cue Detection								

Test Dataset	Model	Train Dataset						
		BA	BA+SFU	BF	BF+BA	BF+BA+SFU	BF+SFU	SFU
BA	BERT	86.91	81.84	78.56	86.40	84.09	62.42	53.11
	RoBERTa	90.36	89.48	86.48	90.57	89.17	83.52	57.71
	XLNet	**93.66**	**92.75**	**90.29**	**93.98**	**92.94**	**89.59**	**58.50**
BF	BERT	72.18	65.38	71.16	73.76	65.77	58.31	51.97
	RoBERTa	78.53	73.03	80.29	84.25	79.63	78.11	55.04
	XLNet	**84.00**	**81.06**	**84.13**	**87.00**	**87.14**	**82.01**	**58.39**
SFU	BERT	21.27	90.39	23.81	23.90	77.78	89.25	90.76
	RoBERTa	23.82	86.72	22.60	30.23	**88.39**	88.42	87.90
	XLNet	**29.34**	**92.36**	**27.57**	**33.04**	74.12	**92.10**	**92.61**
Speculation Cue Detection								

Table 3: Results for Cue Detection (Combined Early Stopping)

bined Early Stopping method), and Table 4 (trained using the Separate Early Stopping method). We compare the Combined and Early Stopping methods in Section 5.4.

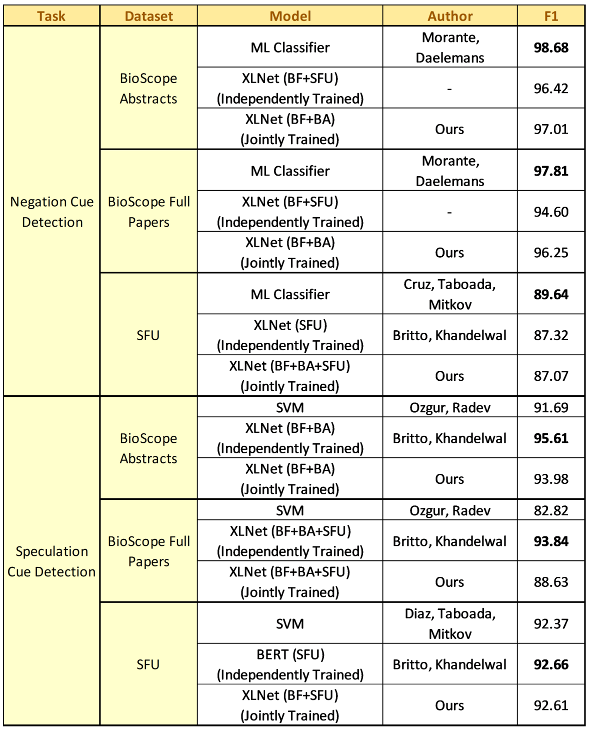


Test Dataset	Model	Train Dataset						
		BA	BA+SFU	BF	BF+BA	BF+BA+SFU	BF+SFU	SFU
BA	BERT	91.10	91.00	87.91	91.90	90.17	75.27	68.73
	RoBERTa	94.21	93.15	90.44	93.30	93.02	88.08	82.45
	XLNet	**95.98**	**95.57**	**93.23**	**96.27**	**94.86**	**91.85**	**84.51**
BF	BERT	85.34	85.15	86.89	88.55	85.03	74.60	76.13
	RoBERTa	90.61	88.45	89.16	90.42	89.18	89.76	79.66
	XLNet	**92.14**	**90.37**	**91.89**	**92.57**	**94.37**	**90.18**	**79.66**
SFU	BERT	**38.61**	**82.96**	**54.17**	26.40	84.42	83.82	77.57
	RoBERTa	33.37	83.82	11.62	11.34	61.98	83.79	**82.42**
	XLNet	19.16	61.09	31.00	**27.28**	**87.07**	**86.14**	64.29
Negation Cue Detection								

Test Dataset	Model	Train Dataset						
		BA	BA+SFU	BF	BF+BA	BF+BA+SFU	BF+SFU	SFU
BA	BERT	84.84	79.72	79.49	88.23	83.66	62.04	52.41
	RoBERTa	90.81	88.80	86.95	89.00	86.95	85.55	57.99
	XLNet	**93.11**	**92.58**	**89.39**	**93.17**	**91.48**	**88.74**	**61.41**
BF	BERT	72.42	65.38	74.00	78.21	69.66	64.84	51.13
	RoBERTa	79.17	73.97	82.96	80.42	76.63	79.93	55.30
	XLNet	**83.10**	**81.34**	**86.94**	**85.78**	**88.63**	**82.46**	**59.64**
SFU	BERT	29.67	**89.71**	**41.43**	21.38	89.01	90.67	83.02
	RoBERTa	**33.28**	88.33	13.26	12.74	58.26	88.05	**86.36**
	XLNet	18.74	62.16	26.96	**25.61**	**92.14**	**92.41**	63.50
Speculation Cue Detection								

Table 4: Results for Cue Detection (Separate Early Stopping)

A comparison of the best models trained jointly on Negation and Speculation compared with the independently trained model variants and the state-of-the-art results is shown in Table 5.



Task	Dataset	Model	Author	F1
Negation Cue Detection	BioScope Abstracts	ML Classifier	Morante, Daelemans	**98.68**
		XLNet (BF+SFU) (Independently Trained)	-	96.42
		XLNet (BF+BA) (Jointly Trained)	Ours	97.01
	BioScope Full Papers	ML Classifier	Morante, Daelemans	**97.81**
		XLNet (BF+SFU) (Independently Trained)	-	94.60
		XLNet (BF+BA) (Jointly Trained)	Ours	96.25
	SFU	ML Classifier	Cruz, Taboada, Mitkov	**89.64**
		XLNet (SFU) (Independently Trained)	Britto, Khandelwal	87.32
		XLNet (BF+BA+SFU) (Jointly Trained)	Ours	87.07
Speculation Cue Detection	BioScope Abstracts	SVM	Ozgur, Radev	91.69
		XLNet (BF+BA) (Independently Trained)	Britto, Khandelwal	**95.61**
		XLNet (BF+BA) (Jointly Trained)	Ours	93.98
	BioScope Full Papers	SVM	Ozgur, Radev	82.82
		XLNet (BF+BA+SFU) (Independently Trained)	Britto, Khandelwal	**93.84**
		XLNet (BF+BA+SFU) (Jointly Trained)	Ours	88.63
	SFU	SVM	Diaz, Taboada, Mitkov	92.37
		BERT (SFU) (Independently Trained)	Britto, Khandelwal	**92.66**
		XLNet (BF+SFU) (Jointly Trained)	Ours	92.61

Table 5: Comparison of Cue Detection Results with State-of-the-Art Results

5.2 Negation Scope Resolution

The results for Negation Scope Resolution are shown in Table 6 (trained using the Combined Early Stopping method) and Table 7 (trained using the Separate Early Stopping method). We compare the Combined and Early Stopping methods in Section 5.4.

A comparison of the best models trained jointly on Negation and Speculation compared with the state-of-the-art results for Negation Scope Resolution is shown in Table 8. Our joint training approach outperforms the existing state-of-the-art models (independently trained transformer based architectures) on all datasets that we experiment with.

5.3 Speculation Scope Resolution

The results for Speculation Scope Resolution are shown in Table 9 (trained using the Combined Early Stopping method) and Table 10 (trained using the Separate Early Stopping method). We compare the Combined and Early Stopping methods in Section 5.4.

Test Dataset	Model	Train Dataset						
		BA	BA+SFU	BF	BF+BA	BF+BA+SFU	BF+SFU	SFU
BA	BERT	94.33	95.48	92.25	95.67	94.91	91.02	84.40
	RoBERTa	95.08	93.80	92.08	93.41	93.77	89.82	83.17
	XLNet	**96.68**	**96.21**	**94.42**	**96.19**	**97.06**	**91.51**	**84.11**
BF	BERT	91.41	90.36	91.10	**97.40**	93.00	88.57	79.94
	RoBERTa	92.74	89.73	90.72	96.53	92.43	89.74	78.28
	XLNet	**95.43**	**93.11**	**93.67**	96.92	**95.17**	**93.03**	**80.18**
SFU	BERT	85.47	90.62	**85.18**	**85.77**	92.07	91.45	91.34
	RoBERTa	85.05	**92.37**	84.20	84.58	**93.19**	90.19	91.31
	XLNet	**86.16**	91.37	84.11	85.63	91.19	**92.69**	**91.41**

Negation Scope Resolution: Global

Test Dataset	Model	Train Dataset						
		BA	BA+SFU	BF	BF+BA	BF+BA+SFU	BF+SFU	SFU
BA	BERT	94.46	94.13	91.52	94.18	95.29	90.86	83.30
	RoBERTa	94.21	93.78	91.59	94.64	93.43	90.84	83.27
	XLNet	**95.89**	**95.89**	**94.82**	**96.30**	**96.01**	**93.59**	**83.94**
BF	BERT	92.61	90.38	92.85	94.44	94.02	89.32	79.72
	RoBERTa	92.64	91.28	91.28	96.28	94.28	89.63	78.90
	XLNet	**94.62**	**93.03**	**94.81**	**96.78**	**96.09**	**93.11**	**80.07**
SFU	BERT	**85.91**	**91.57**	**85.31**	**85.98**	91.15	91.39	90.84
	RoBERTa	84.85	91.22	83.85	84.68	91.03	90.27	**91.71**
	XLNet	85.21	91.37	83.99	85.17	**91.49**	**91.98**	91.51

Negation Scope Resolution: Local

Table 6: Results for Negation Scope Resolution (Combined Early Stopping)

Test Dataset	Model	Train Dataset						
		BA	BA+SFU	BF	BF+BA	BF+BA+SFU	BF+SFU	SFU
BA	BERT	94.27	92.79	92.00	94.50	93.55	90.83	83.75
	RoBERTa	94.55	93.23	91.45	94.46	94.77	89.56	83.05
	XLNet	**96.15**	**96.62**	**94.52**	**96.97**	**95.92**	**94.10**	**83.97**
BF	BERT	91.93	89.18	90.66	94.46	93.39	86.61	79.65
	RoBERTa	92.25	90.04	91.89	95.93	92.02	90.16	78.83
	XLNet	**94.47**	**92.91**	**95.40**	**96.55**	**96.62**	**91.65**	**80.16**
SFU	BERT	85.28	**91.90**	**84.93**	**85.84**	91.19	**91.90**	**91.93**
	RoBERTa	85.01	91.19	83.49	85.42	**91.29**	90.12	91.19
	XLNet	**85.38**	90.45	84.69	85.40	90.38	91.70	90.98

Negation Scope Resolution: Global

Test Dataset	Model	Train Dataset						
		BA	BA+SFU	BF	BF+BA	BF+BA+SFU	BF+SFU	SFU
BA	BERT	94.73	94.17	91.46	94.74	94.19	91.11	**83.49**
	RoBERTa	94.21	93.89	91.26	95.61	94.59	90.27	83.27
	XLNet	**96.22**	**96.80**	**94.30**	**96.08**	**96.33**	**93.14**	83.00
BF	BERT	92.55	90.85	**93.52**	**96.16**	91.62	89.97	79.55
	RoBERTa	92.64	90.19	92.41	94.77	93.43	89.05	78.90
	XLNet	**94.44**	**91.82**	93.50	94.89	**96.29**	**91.59**	**80.35**
SFU	BERT	**86.06**	91.53	**85.40**	**85.88**	91.22	91.89	**92.39**
	RoBERTa	84.85	91.73	83.83	85.12	91.44	91.81	91.71
	XLNet	85.06	**91.81**	82.69	84.43	**92.37**	**92.42**	91.61

Negation Scope Resolution: Local

Table 7: Results for Negation Scope Resolution (Separate Early Stopping)

Task	Dataset	Model	Author	F1
Negation Scope Resolution	BioScope Abstracts	BiLSTM-Joint	Fancellu, Lopez, Webber	92.11
		XLNet (BF+SFU) (Independently Trained)	-	96.77
		XLNet (BA+SFU) (Local) (Jointly Trained)	Ours	96.80
		XLNet (BF+BA+SFU) (Global) (Jointly Trained)	Ours	**97.06**
	BioScope Full Papers	ML MetaLearner	Morante, Daelemans	84.71
		XLNet (BA+SFU) (Independently Trained)	-	97.12
		BERT (BF+BA) (Local) (Jointly Trained)	Ours	96.78
		BERT (BF+BA) (Global) (Jointly Trained)	Ours	**97.40**
	SFU	BiLSTM	Fancellu, Lopez, Webber	89.93
		RoBERTA (BF+SFU) (Independently Trained)	-	91.34
		XLNet (BF+SFU) (Local) (Jointly Trained)	Ours	92.42
		RoBERTa (BF+BA+SFU) (Global) (Jointly Trained)	Ours	**93.19**

Table 8: Comparison of Negation Scope Resolution Results with State-of-the-Art Results

Test Dataset	Model	Train Dataset						
		BA	BA+SFU	BF	BF+BA	BF+BA+SFU	BF+SFU	SFU
BA	BERT	97.14	96.80	95.13	97.59	97.00	94.66	**82.99**
	RoBERTa	96.81	96.28	95.76	97.16	96.63	95.00	78.22
	XLNet	**97.90**	**97.68**	**96.67**	**97.90**	**97.87**	**95.69**	81.43
BF	BERT	93.16	90.87	93.78	96.22	95.07	90.49	77.67
	RoBERTa	93.50	90.90	92.31	95.39	93.20	**92.29**	75.55
	XLNet	**95.58**	**93.55**	**94.34**	**96.36**	**95.36**	91.19	**77.71**
SFU	BERT	**77.93**	89.99	**77.20**	**77.78**	**91.35**	90.41	89.85
	RoBERTa	75.94	**90.81**	74.78	75.76	90.86	**90.53**	89.68
	XLNet	77.82	90.31	74.54	76.47	89.98	89.89	**90.41**

Speculation Scope Resolution: Global

Test Dataset	Model	Train Dataset						
		BA	BA+SFU	BF	BF+BA	BF+BA+SFU	BF+SFU	SFU
BA	BERT	97.09	96.48	95.20	97.36	96.48	94.68	80.68
	RoBERTa	96.67	96.44	95.01	96.94	96.68	94.68	78.64
	XLNet	**97.86**	**97.67**	**96.36**	**98.28**	**97.71**	**96.06**	**82.71**
BF	BERT	93.43	91.41	91.99	93.74	**93.52**	89.58	76.36
	RoBERTa	93.02	91.48	92.22	95.69	90.72	**91.52**	74.92
	XLNet	**94.74**	**93.87**	**94.53**	**96.04**	93.24	91.40	**78.17**
SFU	BERT	**78.10**	89.46	**77.75**	**78.33**	**90.63**	**90.44**	89.82
	RoBERTa	76.21	89.65	74.51	76.09	89.14	89.93	89.76
	XLNet	77.45	**90.16**	73.14	76.26	89.75	90.39	**91.17**

Speculation Scope Resolution: Local

Table 9: Results for Speculation Scope Resolution (Combined Early Stopping)

Test Dataset	Model	Train Dataset						
		BA	BA+SFU	BF	BF+BA	BF+BA+SFU	BF+SFU	SFU
BA	BERT	97.32	96.72	95.14	97.00	96.67	94.88	81.32
	RoBERTa	96.58	97.02	95.54	96.51	96.68	93.83	76.40
	XLNet	**97.83**	**97.64**	**96.53**	**97.80**	**98.00**	**96.61**	**83.39**
BF	BERT	93.53	90.51	91.57	94.15	94.24	**92.81**	76.88
	RoBERTa	92.87	91.47	92.20	94.68	92.84	90.12	73.71
	XLNet	**95.06**	**93.32**	**94.60**	**95.83**	**95.28**	92.18	**78.64**
SFU	BERT	**78.59**	**91.08**	**77.95**	**77.88**	**90.81**	90.47	90.56
	RoBERTa	75.83	89.05	75.07	75.11	89.77	90.09	**90.64**
	XLNet	76.64	90.22	74.01	76.99	90.52	**90.85**	89.66

Speculation Scope Resolution: Global

Test Dataset	Model	Train Dataset						
		BA	BA+SFU	BF	BF+BA	BF+BA+SFU	BF+SFU	SFU
BA	BERT	96.91	97.05	95.00	97.22	97.21	94.86	80.56
	RoBERTa	96.67	96.97	95.32	97.02	96.45	95.02	78.64
	XLNet	**98.09**	**97.72**	**96.38**	**97.43**	**97.73**	**96.14**	**80.61**
BF	BERT	93.47	92.13	93.21	94.48	92.58	90.86	76.24
	RoBERTa	93.02	91.91	90.61	94.58	94.34	91.42	74.92
	XLNet	**94.39**	**93.50**	**94.61**	**94.95**	**96.39**	**92.15**	**76.49**
SFU	BERT	**77.86**	89.81	**76.92**	**78.37**	89.95	89.89	**90.11**
	RoBERTa	76.21	89.60	75.55	75.28	89.42	**90.27**	89.76
	XLNet	75.97	**90.52**	73.40	73.97	**90.26**	89.57	90.10

Speculation Scope Resolution: Local

Table 10: Results for Speculation Scope Resolution (Separate Early Stopping)

A comparison of the best models trained jointly on Negation and Speculation compared with the state-of-the-art results for Speculation Scope Resolution is shown in Table 11. Our joint training approach outperforms the existing state-of-the-art results on BioScope Abstracts and SFU Review Corpus.

5.4 Analysis

Our proposed joint training scheme clearly yields substantial improvements over the independent task-specific training approach, as we outperform the independently trained models consistently, and report new state-of-the-art results, as is illustrated in Tables 5, 8 and 11.

Task	Dataset	Model	Author	F1
Speculation Scope Resolution	BioScope Abstracts	Recursive Neural Network	Ren, Fei, Peng	93.60
		XLNet (BA) (Independently Trained)	Britto, Khandelwal	97.87
		XLNet (BF+BA) (Local) (Jointly Trained)	Ours	**98.28**
		XLNet (BF+BA+SFU) (Global) (Jointly Trained)	Ours	98.00
	BioScope Full Papers	CNN	Qian et al.	86.69
		XLNet (BF+BA) (Independently Trained)	Britto, Khandelwal	**96.91**
		XLNet (BF+BA+SFU) (Local) (Jointly Trained)	Ours	96.39
		XLNet (BF+BA) (Global) (Jointly Trained)	Ours	96.36
	SFU	SVM	Diaz, Taboada, Mitkov	78.88
		BERT (BF+SFU) (Independently Trained)	Britto, Khandelwal	91.00
		XLNet (SFU) (Local) (Jointly Trained)	Ours	91.17
		XLNet (BF+BA+SFU) (Global) (Jointly Trained)	Ours	**91.35**

Table 11: Comparison of Speculation Scope Resolution Results with State-of-the-Art Results

- XLNet consistently outperforms RoBERTa and BERT on the BioScope Corpora. For the SFU Review Corpus, we see a mixed bag of results, but BERT and XLNet tend to outperform RoBERTa. The impact of the similarity between the pretraining corpora and the dataset for which the model is finetuned could account for these observations.

Model	Task			
	Negation Scope Resolution (Separate)	Speculation Scope Resolution (Separate)	Negation Scope Resolution (Combined)	Speculation Scope Resolution (Combined)
BERT	0.57	-0.12	-0.26	-0.50
RoBERTa	0.24	0.07	0.33	-0.35
XLNet	-0.28	-0.03	-0.53	-0.08

Table 12: Difference between Local and Global Preprocessing methods for Scope Resolution

- For Scope Resolution, the Global preprocessing method tends to outperform the Local preprocessing method. This trend is visible in Table 12, which contains the difference between results using the local preprocessing method and the global preprocessing method (i.e. Local - Global), averaged across all train-test dataset combinations, shown for each model-task combination. The majority differences (8 out of 12, or 66%) are negative, showing that global preprocessing method outperforms the local preprocessing method.

- The Combined Early Stopping training method outperform the Separate Early Stopping training method. This trend is visible in Table 13, which contains the difference between the combined early stopping method and the separate early stopping method, (i.e. Combined - Separate), averaged across all train-test dataset combinations, shown for each model-task combination. The majority differences are positive, showing that Combined outperforms Separate. We reason that the combined early stopping method avoids overfitting to the validation set, due to more examples being considered in the validation set.

Model	Task			
	Negation Cue Detection	Speculation Cue Detection	Negation Scope Resolution	Speculation Scope Resolution
BERT	-5.48	-1.99	0.19	0.24
RoBERTa	1.89	2.36	0.02	0.38
XLNet	2.65	2.73	0.19	-0.20

Table 13: Difference between Combined and Separate Early Stopping training methodologies

6 Conclusion

In this paper, we explored the realm of Multitask training to jointly train the same model to perform both negation and speculation detection. We experimented with transformer-based architectures (BERT, XLNet and RoBERTa), and proposed schemes to jointly train the cue detection model for both negation and speculation, and the scope resolution model for both negation and speculation. Our approach yielded improvements over the independently trained versions of the same architectures, and we reported new state-of-the-art results for both negation and speculation scope resolution on the BioScope Corpus and the SFU Review Corpus. We also evaluated the different design choices that were involved, and observed that the Combined Early Stopping variant gave the best overall performance.

The future scope of this work would be to look at using this scheme to jointly train a model for more such tasks, like NER and Sentiment Analysis, along with Negation and Speculation Detection.

References

Jeremy Barnes, Erik Velldal, and Lilja Øvrelid. 2019. Improving sentiment analysis with multi-task learning of negation. *CoRR*, abs/1906.07610.

Parminder Bhatia, Busra Celikkaya, and Mohammed Khalilia. 2018. End-to-end joint entity extraction and negation detection for clinical text. *CoRR*, abs/1812.05270.

Benita Kathleen Britto and Aditya Khandelwal. 2020. Resolving the scope of speculation and negation using transformer-based architectures.

Noa P Cruz, Maite Taboada, and Ruslan Mitkov. 2016. A machine-learning approach to negation and speculation detection for sentiment analysis. *Journal of the Association for Information Science and Technology*, 67(9):2118–2136.

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*.

Hao Fei, Yafeng Ren, and Donghong Ji. 2020. Negation and speculation scope detection using recursive neural conditional random fields. *Neurocomputing*, 374:22–29.

Aditya Khandelwal and Suraj Sawant. 2020. NegBERT: A transfer learning approach for negation detection and scope resolution. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5739–5748, Marseille, France. European Language Resources Association.

Halil Kilicoglu and Sabine Bergler. 2008. Recognizing speculative language in biomedical research articles: A linguistically motivated perspective. *BMC bioinformatics*, 9 Suppl 11:S10.

Halil Kilicoglu and Sabine Bergler. 2010. A high-precision approach to detecting hedges and their scopes. pages 70–77.

Natalia Konstantinova, Sheila C. M. De Sousa, Noa P. Cruz, Manuel J. Maña, and Ruslan Mitkov. A review corpus annotated for negation, speculation and their scope.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Guillermo Moncecchi, Jean-Luc Minel, and Dina Wonsever. 2012. Improving speculative language detection using linguistic knowledge. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 37–46.

Roser Morante, Vincent Asch, and Walter Daelemans. 2010. Memory-based resolution of in-sentence scopes of hedge cues. pages 40–47.

Roser Morante and Walter Daelemans. 2009. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '09, page 28–36, USA. Association for Computational Linguistics.

Arzucan Özgür and Dragomir R. Radev. 2009. Detecting speculations and their scopes in scientific text. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, page 1398–1407, USA. Association for Computational Linguistics.

Zhong Qian, Peifeng Li, Qiaoming Zhu, Guodong Zhou, Zhunchen Luo, and Wei Luo. 2016. Speculation and negation scope detection via convolutional neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 815–825.

Jonathon Read, Erik Velldal, Stephan Oepen, and Lilja Øvrelid. 2011. Resolving speculation and negation scope in biomedical articles with a syntactic constituent ranker.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The bioscope corpus: Annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, page 38–45, USA. Association for Computational Linguistics.

Buzhou Tang, Xiaolong Wang, Xuan Wang, Bo Yuan, and Shixi Fan. 2010. A cascade method for detecting hedges and their scope in natural language text. pages 13–17.

Erik Velldal. 2011. Predicting speculation: A simple disambiguation approach to hedge detection in biomedical literature. *Journal of biomedical semantics*, 2 Suppl 5:S7.

Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2010. Resolving speculation: Maxent cue classification and dependency-based scope rules. pages 48–55.

Erik Velldal, Lilja Øvrelid, Jonathon Read, and Stephan Oepen. 2012. Speculation and negation: Rules, rankers, and the role of syntax. *Computational Linguistics*, 38:369–410.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.