"*Sharks are not the threat humans are*": Argument Component Segmentation in School Student Essays

Tariq Alhindi and Debanjan Ghosh

Motivation

- The central component of the argument are the claim and premise
  - "an assertion that deserves our attention" (Toulmin, 2003)
  - "a statement that is in dispute and that we are trying to support with reasons" (Govier, 2010)

- Argumentation
  - segmentation, component (claim, premise) and relation (support, attack) detection
  - claim example "Parents should be limiting screen time for their children"

- Argument/Claim Detection applications
  - Assess public opinion on political and social issues to foster public deliberation
  - Refine search and information retrieval
  - other applications: legal documents, fact-checking

Motivation

Educational Applications

Analyze students' writings for essay scoring
(Klebanov et al. 2014, Ghosh et al. 2016)

Generate quantitative and qualitative feedbacks to help students (K-12) writings such as on the Writing Mentor app

Writing Mentor App

Should Artificial Sweeteners be Banned in America?

Diet soda , sugar - free gum, and low - calorie sweeteners are what most people see as a way to sweeten up a day without the calories.

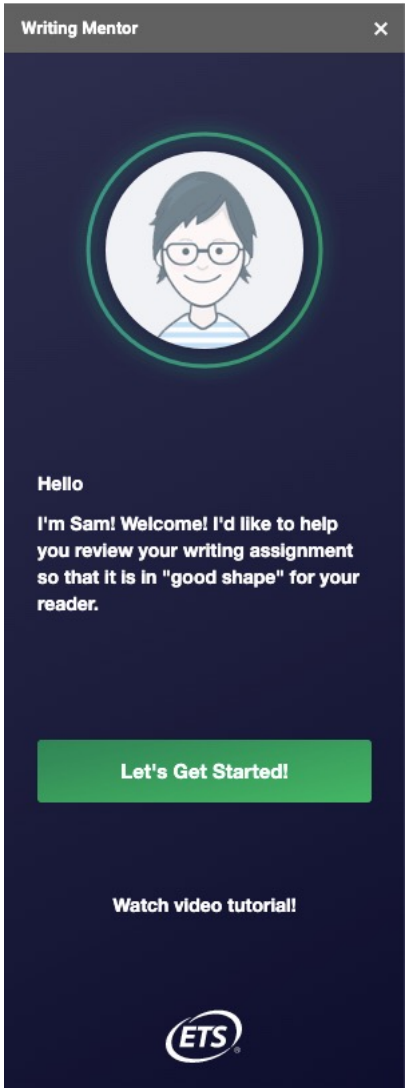
Despite the lack of calories, artificial sweeteners have multiple negative health effects.

Over the past century, science has made it possible to replicate food with fabricated alternatives that simplify weight loss.

Although many thought these new replacements would benefit overall health, there are more negative effects on manufactured food than the food they replaced.

Artificial sweeteners have a huge impact on current day society.

Legends: B-Claim I-Claim B-Premise I-Premise O-Arg



https://mentormywriting.org/

Research questions

Can we detect claim and premise boundaries at the token level from school student essays?

Can we develop robust claim and premise detection models that go beyond lexicon-based matching?

Task

Claim and Premise Segmentation in each sentence using a B-I-O setup

B-Claim/B-Premise	beginning of a claim/premise
I-Claim/I-Premise	inside a claim/premise
O-Arg	outside argumentative text

Although many thought these new replacements would benefit overall health, there

are more negative effects on manufactured food than the food they replaced

Task

Claim and Premise Segmentation in each sentence using a B-I-O setup

B-Claim/B-Premise	beginning of a claim/premise
I-Claim/I-Premise	inside a claim/premise
O-Arg	outside argumentative text

Although	many	thought	these	new	replacements	would	benefit	overall	health	,	there
O-Arg	O-Arg	O-Arg	B-Claim	I-Claim	I-Claim	I-Claim	I-Claim	I-Claim	I-Claim	I-Claim	B-Premise

are	more	negative	effects	on	manufactured	food	than	the	food	they	replaced	.
I-Premise	I-Premise	I-Premise	I-Premise	I-Premise	I-Premise	I-Premise	I-Premise	I-Premise	I-Premise	I-Premise	I-Premise	I-Premise

Outline of the Talk

Background and Data

Features and Models

Results

Discussions and Future Work

Background and Data

Segmentation then Type Detection and Relation Detection

using 400 argumentative essays (Stab & Gurevych, 2017, Eger et al. 2017))

Sentence Claim Detection

Cross-domain claim detection in 6 datasets (Daxenberger et al. 2017)
Using unlabeled data from the same domain (Chakrabarty et al. 2019)

Characteristics of our Dataset
(why merging segmentation and Type Detection)

Multiple argument components in one sentence

Many unsupported claims

Annotation Efforts

- Annotation of *argumentative* essays with 3 annotators
- Annotation of claims and premises
- Several rounds of calibration to finalize the guideline
  - 10 essays were triple annotated in the pilot task
  - 65 essays are annotated by pair(s); rest are annotated by one annotator
- For claim: modest agreement of 0.71 and for premise: high agreement of 0.9 (Krippendorff's α)

Dataset

Split	Essays	B-Claim	I-Claim	B-Premise	I-Premise	O-Arg	Total
training	100	1,780	21,966	317	3,552	51,478	36,546
dev	10	171	1,823	32	371	4,008	6,405
test	35	662	8,207	92	1,018	14,987	24,955

Outline of the Talk

Background and Data

Features and Models

Results

Discussions and Future Work

Features

- Discrete Features
- Word Embeddings from Transformers (e.g., BERT)

Models

- Conditional Random Field (Sequence Classification)
- Deep Learning Models
  - BiLSTMs
  - Transformers (e.g. BERT)

Discrete Feature Groups

Structural

token position features, punctuation features
position of covering sentence

Syntactic

POS, Lowest common ancestor (LCA), LCA types

Lex-Syntactic

Relations governing the token and its context
N hops deep in retrieving relations
Features: (token, previous, next): token_dependency-relation

Christian Stab, and Iryna Gurevych. "Parsing argumentation structures in persuasive essays." Computational Linguistics (2017)

Feature-based Models

Conditional Random Field (CRF)

Takes a sequence as input (sentence)

Make prediction to each element (token) in the sequence while considering other neighboring elements (tokens in the same sentence)

Deep Learning Models

- **Training a BiLSTM-CRF tagger**

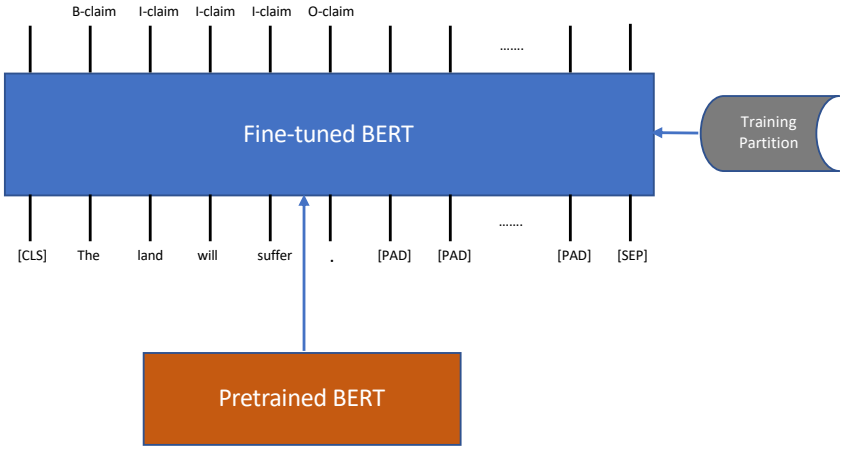
- **Pretrained transformers (e.g., BERT)** as feature extractors
  - language models trained on a lot of data (wikipedia, google books)
  - using embeddings generated by these models as features

- **Pretrained transformers as classifiers**
  - fine-tuning on the task (training partition)
  - fine-tuning on a related unlabeled corpus
  - fine-tuning with a multi-task objective

Fine-tuning Transformers

Training a Bert-based classifier

1. Fine-tuning on the training partition
   model choice and hyperparameter tuning (cased vs uncased, batch size, number of epochs, etc.), weighted loss per class according to class frequencies



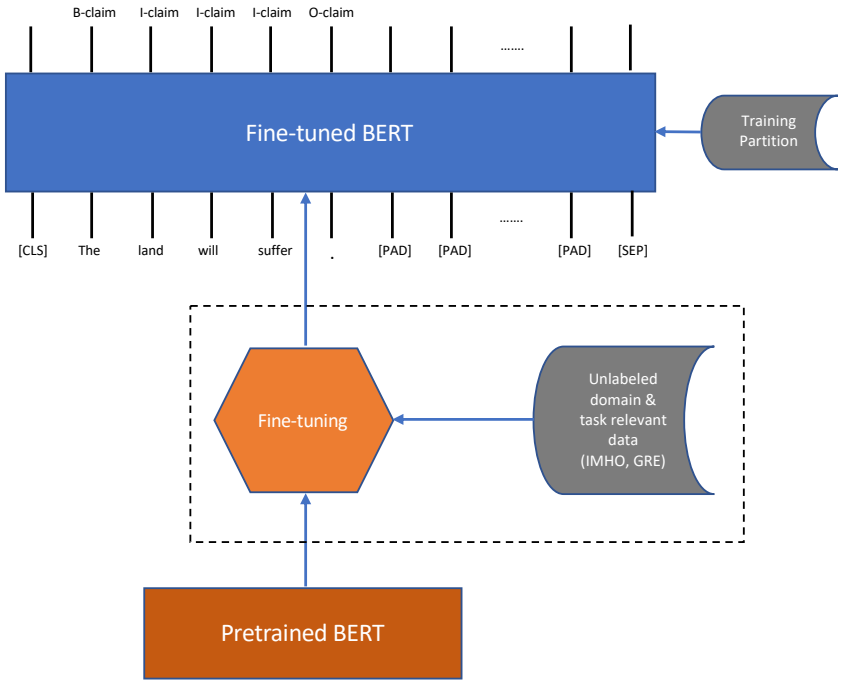
Fine-tuning Transformers

Training a Bert-based classifier

1. Fine-tuning on the training partition
   model choice and hyperparameter tuning (cased vs uncased, batch size, number of epochs, etc.), weighted loss per class according to class frequencies

2. Adaptive pretraining on argumentative corpora:
   IMHO: trained on argumentative subreddit
   (Chakrabarty et al. 2019)

   GRE essays: trained on student essays
   (Ghosh et al. 2020)



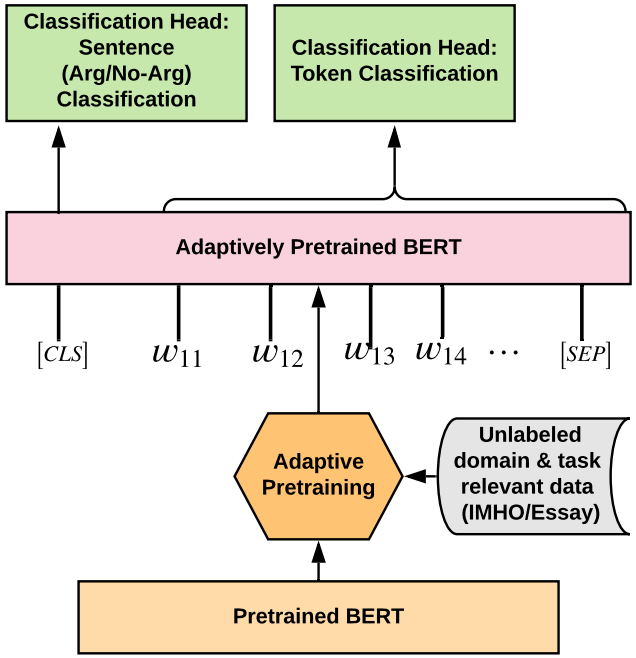
Fine-tuning Transformers

Training a Bert-based classifier

3. Fine-tuning with adaptive pretraining on unlabeled data from a relevant domain followed by fine-tuning on the labeled dataset with the multitask variation

dynamic weighting of task specific losses

$$L = \sum_t \frac{1}{2\sigma_t^2} L_t + \log \sigma_t^2$$



Outline of the Talk

Background and Data

Features and Models

Results

Discussions and Future Work

Results

- Discrete features (different sets) with CRF

	CRF						
Features	B-Claim	I-Claim	B-Premise	I-Premise	O-Arg	Acc.	F1
lexSyn	.395	.530	.114	.176	.768	.673	.397
Discrete*	.269	.504	0	.013	.695	.595	.296
Embeddings	.401	.560	.048	.139	.769	.676	.384
Embeddings+lexSyn	.482	.610	.134	.180	.764	.682	.434
Embeddings+Discrete*	.434	.593	.055	.152	.762	.676	.399

- BiLSTM standalone model
- +CRF based identification

	BiLSTM						
Setup	B-Claim	I-Claim	B-Premise	I-Premise	O-Arg	Acc.	F1
BiLSTM	.556	.680	.239	.438	.797	.735	.542
BiLSTM-CRF	.558	.676	.199	.378	.789	.727	.520

Results

- $BERT_{bl}$: BERT based baseline
- $BERT_{bl\_IMHO}$: fine-tuned on IMHO
- $BERT_{bl\_essay:}$ fine-tuned on essays

- $BERT_{MT:}$ like above in a MT setting (token + sentence classification)

	BERT						
Setup	B-Claim	I-Claim	B-Premise	I-Premise	O-Arg	Acc.	F1
$BERT_{bl}$	.563	.674	.274	.425	.795	.728	.546
$BERT_{bl_{IMHO}}$	.571	.681	.304	.410	.795	.730	.540
$BERT_{bl_{essay}}$	.564	.676	.261	.406	.792	**.747**	.561
$BERT_{mt}$	.567	.685	.242	.439	.805	.741	.548
$BERT_{mt_{IMHO}}$	.562	.684	.221	.413	.794	.731	.534
$BERT_{mt_{essay}}$	.580	.702	.254	.427	.810	**.752**	**.574**

Takeaways

- Best model is $BERT_{mt\_essay}$
  - Gains in all five token-classes, up to 7.5% accuracy
- 5-class: Merged Segmentation and Type Detection
- 3 class: Argument Segmentation (B-Arg, I-Arg, O-Arg)
  - only gives 2 points improvements
  - Much lower than other datasets
- Low accuracy for Premise detection

Outline of the Talk

Background and Data

Features and Models

Results

Discussions and Future Work

Challenges

Allowed Transitions

- back-to-back B-claim, B-premise
- I-claim starts after O-Arg

post-processing based on the possible transitions

B→ I

I →B I → I I → O

O→B O→O

Challenges

non-arguments classified as arguments (frequent)

- Mixing arguments with opinions

  e.g. *that actually makes me feel good afterward*

Challenges

missing multiple-claims from a sentence

Some coral can recover from this though for most it is the final straw.

B I I I I I O B I I I I I

Some coral can recover from this though for most it is the final straw.

B I I I I I O O I I I I I

Challenges

investigating run-on sentences

Humans in today ‘s world do not care about the consequences , only the money they may gain.

The oceans are also another dire need in today’s environment, each day becoming more filled with trash and plastics.

Effect of the Multitask Learning

- identify claims and premises that are missed by the single task model(s), such as:

  many more negative effects that come with social media — *claim*

- clever handling of the back-propagation helps the multitask model to reduce false positives to be more precise.

  internet's social networks help teens find communities — *non-arg*

  take $1.3 billion o↵ $11.3 billion the NCAA makes and give it to players — *opinion*

Conclusion and Future Work

- Our findings show that a multitask BERT performs the best with an absolute gain of 7.5% accuracy over the discrete features.

- We can also generate personalized and relevant feedback for the students (e.g., which are the supported/unsupported claims in the essay?) that is useful in the paradigm of automated writing assistance.

- domain identification (college essays vs. school essays)
  genre prediction in essays (argumentative vs. narrative),

Thank You