Stance Classification of Context-Dependent Claims

Roy Bar-Haim[1], **Indrajit Bhattacharya**[2], **Francesco Dinuzzo**[3*]
Amrita Saha[2], and **Noam Slonim**[1]

[1]IBM Research - Haifa, Mount Carmel, Haifa, 31905, Israel
[2]IBM Research - Bangalore, India
[3]IBM Research - Ireland, Damastown Industrial Estate, Dublin 15, Ireland
{roybar,noams}@il.ibm.com, {indrajitb,amrsaha4}@in.ibm.com

Venue : 2017 EACL

발제자 : 이다현

HUMANE Lab

2024-02-27

Goal

- 논란이 되는 주제(Controversial Topic)에 대해 관련된 주장(claims)을 탐지하는 문제에 이어, 주장의 입장 분류(claim stance classification)라는 보완적인 작업을 소개
  - 첫 번째 벤치마크 데이터셋을 제공
  - 주장의 입장 분류 문제를 세 부분으로 분해
    - Open-domain target identification
    : 주제(topic)와 주장(claim) 각각에 대한 target을 식별
    - 각 target에 대한 Sentiment classification

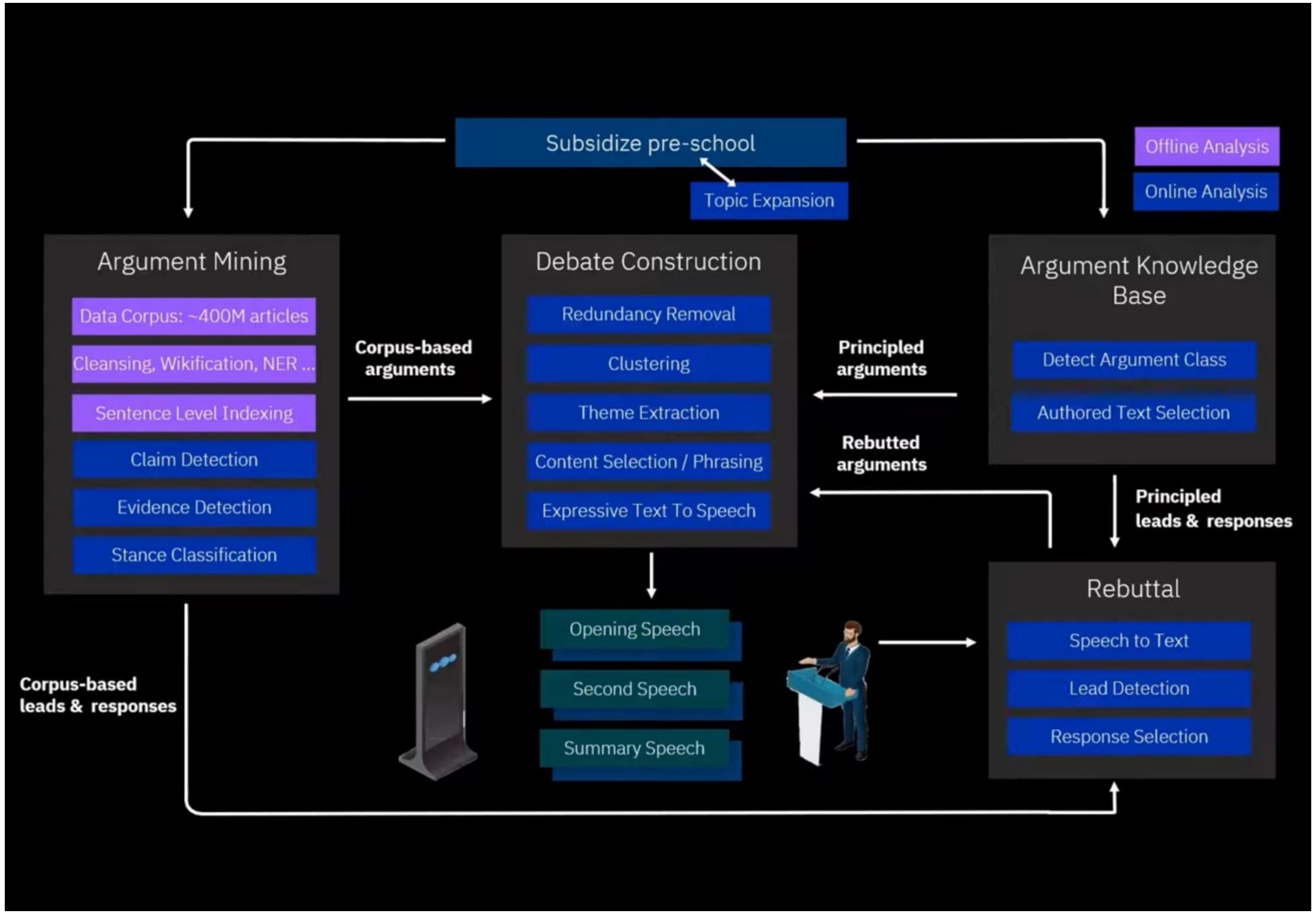
Target: phrase about which they make a positive or a negative statement

Introduction

- On demand generation of pro and con arguments for a given controversial topic would be of great value in many domains
- Debating support and decision support
- E.g
  - “The sale of violent video games to minors should be banned.”(controversial topic)
  - (Pro)Violence is damaging the kindness of children.(claim)
  - (Pro)Love makes world peaceful and harmony.(claim)

Notable Research

- IBM의 Debater® Project



Motivation

- Sorting extracted claims into pro and con improves the usability of both debating and decision support systems.
  - Topic $\boldsymbol{t}$ and extracted claims $\boldsymbol{c}$ from $\boldsymbol{t}$
  - Semantic model for predicting claim stance(Pro/Con)
    - Three sub-problems
      - Identify Targets
      - Identify sentiment towards target
      - Consistency or Contrast between targets
    - More general topic & contrast scope

Dataset

- Randomly selected 55 topics worded as “This house + …”
- 2394 Claims assessed polarity by 5 annotators

#	Debate Topic (Motion)		Claim	
1	This house believes that **advertising** is harmful. ⊖	⇔	**Marketing** promotes consumerism and waste. ⊖	Pro
2	This house would ban **boxing**. ⊖	⇔	**Boxing** remains the 8th most deadly sport. ⊖	Pro
3	This house would embrace **multiculturalism**. ⊕	⇎	**Unity** is seen as an essential feature of the nation and the nation-state. ⊕	Con
4	This house supports **the one-child policy of the republic of China**. ⊕	⇎	**Children with many siblings** receive fewer resources. ⊖	Pro
5	This house would **build hydroelectric dams**. ⊕	⇔	As an alternative energy source, **a hydroelectric power source** is cheaper than both nuclear and wind power. ⊕	Pro
6	This house believes that it is sometimes right for the government to restrict **freedom of speech**. ⊖	⇔	**Human rights** can be limited or even pushed aside during times of national emergency. ⊖	Pro
7	This house would abolish **the monarchy**. ⊖	⇔	**Hereditary succession** is outdated. ⊖	Pro
8	This house would **unleash the free market** ⊕	⇎	Virtually all developed countries today successfully promoted their national industries through **protectionism**. ⊕	Con
9	This house supports **the one-child policy of the republic of China**. ⊕		If, for any reason, the single child is unable to care for their older adult relatives, the oldest generations would face a lack of resources and necessities.	Con

Table 1: Sample topic and claim annotations. Targets are marked in bold. ⊕/⊖ denote positive/negative sentiment towards the target, and ⇔/⇎ denote consistent/contrastive targets.

Semantic Model

- Topic: This house supports the freedom of speech
- (Pro claim) “in favor of free discussion” or “criticizing censorship”
- (consistent) “freedom of speech” and “free discussion”
- (contrastive) “freedom of speech” and “censorship”

Semantic Model

- (Continuous)Claim stance classification model
  - Claim $\boldsymbol{c}$, Topic $\boldsymbol{t}$
  - Claim-target $\boldsymbol{x_c}$, Topic-target $\boldsymbol{x_t}$
  - Claim-Sentiment $\boldsymbol{s_c}$, Topic-Sentiment $\boldsymbol{s_t}$
  - Contrast relation $\boldsymbol{R(x_c, x_t)}$
  - E.g $\boldsymbol{Stance(c,t) = s_c \times R(x_c, x_t) \times s_t}$
- Real-valued prediction
  - Top K predictions
  - Threshold
  - Class: sign / Confidence: absolute value

Claim target identification

- Open-domain, generic target identification
  - $\boldsymbol{x_t}$ and $\boldsymbol{s_t}$ are given
  - Focusing on finding $\boldsymbol{x_c}$ and $\boldsymbol{s_c}$ with L2-regularized Logistic Regression Classifier

Syntactic and Positional: The dependency relation of x in c; whether x is a direct child of the root in the dependency parse tree for c; the minimum distance of x from the start or the end of the chunk containing it.
Wikipedia: whether x is a Wikipedia title, (e.g. *human rights*)

Table 2: Featured extracted for a target candidate $\boldsymbol{x}$ in a claim $\boldsymbol{c}$

Sentiment: The dependency relation connecting x to any sentiment phrase in the rest of c. The (Hu and Liu, 2004a) sentiment lexicon was used. For example, *Hereditary succession* is the sentiment target of *outdated*, indicated by the subject-predicate relation connecting them (Table 1, row 7).
Topic relatedness: Semantic similarity between x and the topic target , e.g. *Marketing* and *advertising* (Table 1, row 1). We consider morphological similarity, paths in WordNet (Miller, 1995; Fellbaum, 1998), and cosine similarity of word2vec embeddings (Mikolov et al., 2013).

- True target과 같거나 상당히 겹치는 candidate phrases □ positive training example
- 나머지는 negative training example로 사용

Claim sentiment Classification

- Sentiment score

$$\frac{\boldsymbol{p} - \boldsymbol{n}}{\boldsymbol{p} + \boldsymbol{n} + \mathbf{1}}$$

  - Sentiment matching: Hu와 Liu의 sentiment lexicon을 이용
  - Sentiment shifters application: 약 160개의 Sentiment Shifters를 포함하는 어휘집을 수동으로 구성
  - Sentiment weighting and score computation: $\boldsymbol{p}$ and $\boldsymbol{n}$ 각각 claim에서 detect된 positive 가중합과 negative 가중합
  - A weight of $\boldsymbol{d}^{-0.5}$. $\boldsymbol{d}$ 는 sentiment term과 target 사이의 거리를 나타냄

Contrast Classification

- Anchor pair

  <u>*Atheism*</u> *& denying the existence of God*

  - $\boldsymbol{w(u)} \times |\boldsymbol{r(u,v)}| \times \boldsymbol{w(v)}$
  - $\boldsymbol{w(x)} = \frac{\boldsymbol{tf(x)}}{\boldsymbol{df(x)}}$
- Relatedness measure

  based on *co-occurrence* of the anchor pair with consistent and contrastive *cue-phrases.*

  - $\mathbf{P}(\boldsymbol{Lex}+|\boldsymbol{u},\ \boldsymbol{v}) = \frac{\boldsymbol{Freq(u,Lex+,v)}}{\boldsymbol{Freq(u,v)}}$
  - $\mathbf{P}(\boldsymbol{Lex}+|\boldsymbol{u},\ \boldsymbol{v})$ if $\mathbf{P}(\boldsymbol{Lex}+|\boldsymbol{u},\ \boldsymbol{v}) > \mathbf{P}(\boldsymbol{Lex}-|\boldsymbol{u},\ \boldsymbol{v})$ otherwise, $-\mathbf{P}(\boldsymbol{Lex}+|\boldsymbol{u},\ \boldsymbol{v})$
- Contrast score → Random forest classifier
  - $\mathbf{P}(\boldsymbol{x_c}, \boldsymbol{a_c}) \times \mathbf{r}(\boldsymbol{a_c}, \boldsymbol{a_t}) \times \mathbf{P}(\boldsymbol{x_t}, \boldsymbol{a_t})$

Contrast Classification

- Two corpus: Query log(advertising and marketing), Wikipedia(Military vs diplomatic)
- Targets "atheism" and "denying the existence of God"
- Anchor pair candidate $(atheism, God), (atheism, existence)$ ….
- $R(atheism, God)\ w(atheism)\ and\ w(God)$
  - $\boldsymbol{w(x)} = \frac{\boldsymbol{tf(x)}}{\boldsymbol{df(x)}}$
- Relatedness measure
  - $\mathbf{P}(\boldsymbol{Lex} + |\boldsymbol{atheism},\ \boldsymbol{God})$
- Contrast score → consistency
  - $\mathbf{P}(\boldsymbol{atheism}, \boldsymbol{atheism}) \times \mathbf{r}(\boldsymbol{atheism}, \boldsymbol{God}) \times \mathbf{P}(\boldsymbol{denying\ the\ existence\ of\ God}, \boldsymbol{God})$

Evaluation

- A training set, comprising 25 topics (1,039 claims)
- A test set, comprising 30 topics (1, 355 claims)
- The training set was used to train the target identification classifier and the contrast classifier in system
- Predicting the test data
- Accuracy and coverage
  - Two baseline
  - Our model
  - Combine our model with baseline

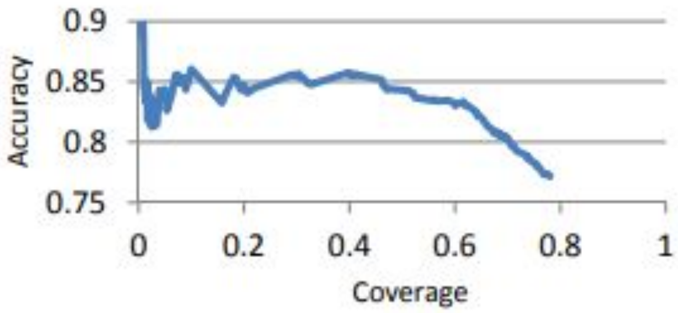
$$coverage(\alpha) = \frac{predicted(\alpha)}{claims}$$

$$accuracy(\alpha) = \frac{correct(\alpha)}{predicted(\alpha)}$$

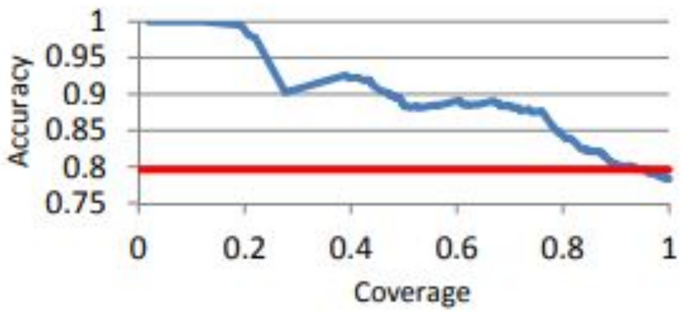
Result

Configuration	Accuracy@Coverage 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Baselines										
Unigrams SVM	0.688	0.688	0.659	0.612	0.587	0.563	0.560	0.554	0.554	0.547
Unigrams+Sentiment SVM	0.717	0.717	0.717	0.709	0.693	0.691	0.687	0.668	0.655	0.632
Our System										
Sentiment Score	0.752	0.720	0.720	0.720	0.720	0.720	0.636	0.636	0.636	0.636
+Targeted Sentiment	0.770	0.770	0.770	0.749	0.734	0.734	**0.706**	0.632	0.632	0.632
+Contrast Detection	**0.849**	**0.847**	**0.836**	**0.793**	**0.767**	**0.740**	0.704	0.632	0.632	0.632
Our System+Unigrams SVM	0.784	0.758	0.749	0.743	0.730	0.711	0.682	**0.671**	**0.658**	**0.645**

Table 3: Stance classification results. Majority baseline accuracy: 51.9%



(a) Sentiment (majority baseline: 56.2%)



(b) Contrast (majority baseline: 79.6%)

Figure 1: Performance of Sub-Components

Conclusion

- Open-domain claim stance classification
- 복잡한 task를 더 간단하게 쪼갠 모델 구현
- 이를 잘 설명하는 subtask 제안
- 이러한 subtask를 수행하기 위한 주석이 달린 dataset 구축

Open Question

- 수동 주석 처리나 알고리즘, 지식 그래프를 사용하는 방식에는 한계가 있을 것 같다. LLM이나 Retrieval LM을 활용해 성과를 낼 수 없을까?