CogComp@Penn is looking
for
excellent post-docs

Benchmarking Zero-shot Text Classification:

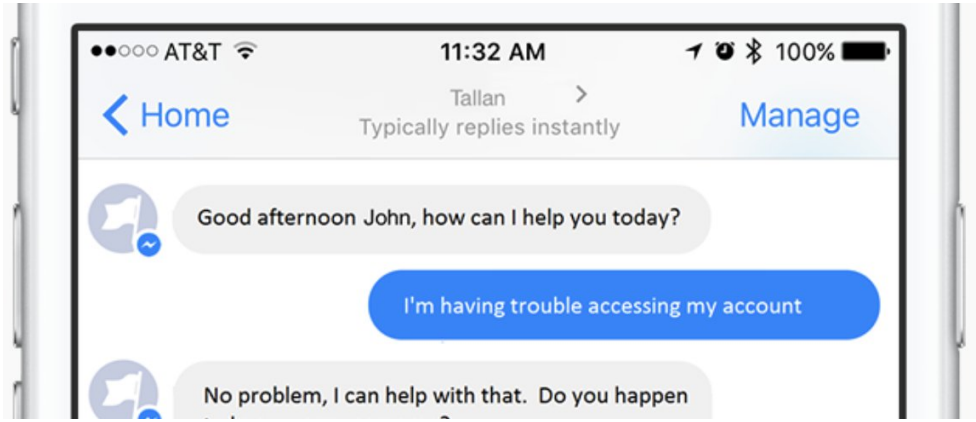
Datasets, Evaluation and Entailment Approach

Wenpeng Yin, Jamaal Hay, **Dan Roth**

University of Pennsylvania

Text Classification

- A lot of the work in NLP can still be viewed as **text classification**
  - ☐ Categorization into topics
  - ☐ Identify intention of text
  - ☐ Identify abusing text
  - ☐ Framing of news stories
  - ☐ Classify fact/opinion
  - ☐ ……



Frame Class
2nd Amendment
Gun control/regulation
Politics
Mental health
School/Public space safety
Race/Ethnicity
Public opinion
Society/Culture
Economic consequences

- While you understand the labels, models are not given information about the labels
- NLP simply views these tasks as multi-class classification

Text Categorization

Second seed Rafael Nadal continues his quest for an unprecedented 12th title at this event against Grigor Dimitrov of Bulgaria. The Spaniard leads their FedEx ATP Head2Head 11-1 and has won their past four matches, in addition to all four of their previous battles on clay. Their most recent [illegible] semi-finals, which saw

It's about Sport
It's about Tennis

- Traditional text categorization requires training a classifier over a set of labeled documents (1,2,….k)
- Someone needs to label the data (costly)
- All your model knows is to classify into these given labels

Total costs for on-the-j[illegible] an average of 5% in 2018, surpassing $14,000 a year per employee, according to a National Business Group on Health survey of large employers. Specialty drugs continue to be the top driver of increasing costs. Companies will pick up nearly 70% of the tab, but employees must still bear about 30%, or roughly $4,400, on average.

It's about Money
It's about Health Care

You can classify these documents **without task specific annotation**, since you have an "understanding" of the labels

Categorization without Labeled Data [AAAI'08, AAAI'14, IJCAI'16]

- Given:
  - A single document (or: a collection of documents)
  - A taxonomy of categories into which we want to classify the documents
- Dataless/Zero-Shot procedure:
  - Let $f(l_i)$ be the semantic representation of the labels (label descriptions)
  - Let $f(\mathbf{d})$ be the semantic representation of a document
  - Select the most appropriate category:

$$l_i^* = \text{argmin}_i \text{ dist } (f(l_i) - f(\mathbf{d}))$$

- Key Question:
  - How to generate good Semantic Representations?
- Originally:
  - Wikipedia-based (ESA) sparse representations
    - Other representations possible.
- Works great for Topic categorization.
  - Competitive with supervised approaches (w/o too much data)
  - Even cross-lingually [Song et al'16]

Types of Text Classification

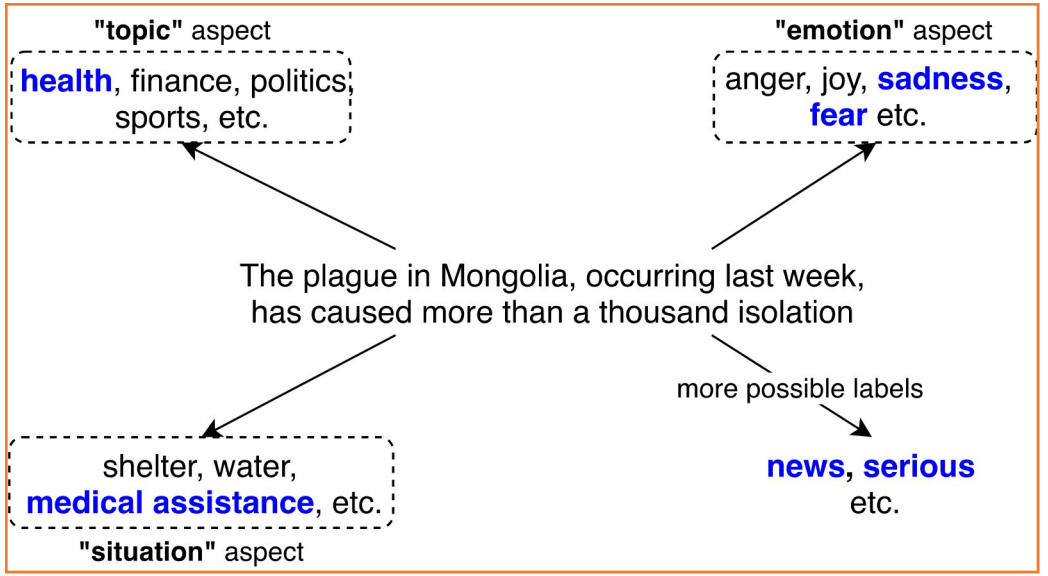
- But there is more to Text classification than topic categorization
- Type (1): The document **is about** [Tennis; Religion; Banking….]
  - Vocabulary will be related to Tennis,….
- Type (2): The document **is** [happy; harassing; abusive]
  - Vocabulary **will not mention** harassment….
- This distinction goes unnoticed when we simply train a multiclass classifier
  - If you give it enough examples, it will get it.
  - But that means, that you need to give it a lot of examples….

Can we do better?

Challenges to Zero-Shot Text Classification

- Label-Aware classification
  - Our models need to “understand” the labels.
  - While the community spends a lot of time representing input data, so far, we have not spent time representing the task.
  - As shown earlier, easy for Topics; challenging for other types of text classifications.

- Methodology: What is 0-shot-TC?
  - **Definition:** Classify into a given set of labels Y, **without observing any Y-labeled documents.**
  - Datasets & evaluation methodologies vary.



Our Focus

- **Tasks**: data supporting 3 TC tasks:
  - "topic categorization", "emotion detection" and "situation detection"

- **Technical Approach**: we propose a textual entailment approach which
  1) Enforces "label understanding"
  2) Does not rely on task-specific training
  3) Deals with multiple TC tasks/datasets uniformly.

Benchmark the dataset

Benchmark the Dataset

- Topic Detection
  - A large-scale **Yahoo** from Zhang et al. (2015)
  - **10 classes**: “Society & Culture”, “Science & Mathematics”, “Health”, “Education & Reference”, “Computer & Internet”, “Sports”, “Business & Finance”, “Entertainment & Music”, “Family & Relationships”, “Politics & Government”
  - We split into **train/dev/test for zero-shot**
    - (originally, it has only train/test for fully supervised topic classification)
  - Evaluation: **accuracy**

Benchmark the Dataset

- Emotion Detection
  - "**UnifyEmotion**" from Bostan and Klinger (2018)
  - It was created by unifying **multiple public emotion datasets** from **multiple domains**.
  - **9+1 classes**: "sadness", "joy", "anger", "disgust", "fear", "surprise", "shame", "guilt", "love", "none"
  - Not balanced
  - Evaluation: **weighted F1**
  - Example:
    - "more than one hundred people died in the earthquake" → "sadness"

Benchmark the Dataset

- Situation Frame Detection
  - An **event-type classification** task.
  - A situation frame could be a "**need**" situation
    - A need for water or medical aid, or
  - An "**issue**" situation
    - Crime, or violence.
  - Released in Mayhew et al. (2019)
  - **8 "need" types**: "food", "infrastructure", "medical assistance", "search", "shelter", "utility", "water", "evacuation"
  - **3 "issue" types**: "regime change", "terrorism" , "crime violence"
  - **multi-label classification**, unbalanced
  - Evaluation: **weighted F1**
  - Example:
    - "Most houses and buildings have been destroyed in the earthquake" → "shelter", "water", "food"

Dataset Summary

	Topic detection	Emotion detection	Situation detection
#class	10	9+1(None)	11+1(None)
distribution	balanced	imbalanced	imbalanced
classification	Single-label	Single-label	Multi-label
evaluation	Acc.	Weighted F1	Weighted F1
Has “None” type	N	Y	Y

Benchmark the evaluation

Benchmark the Evaluation

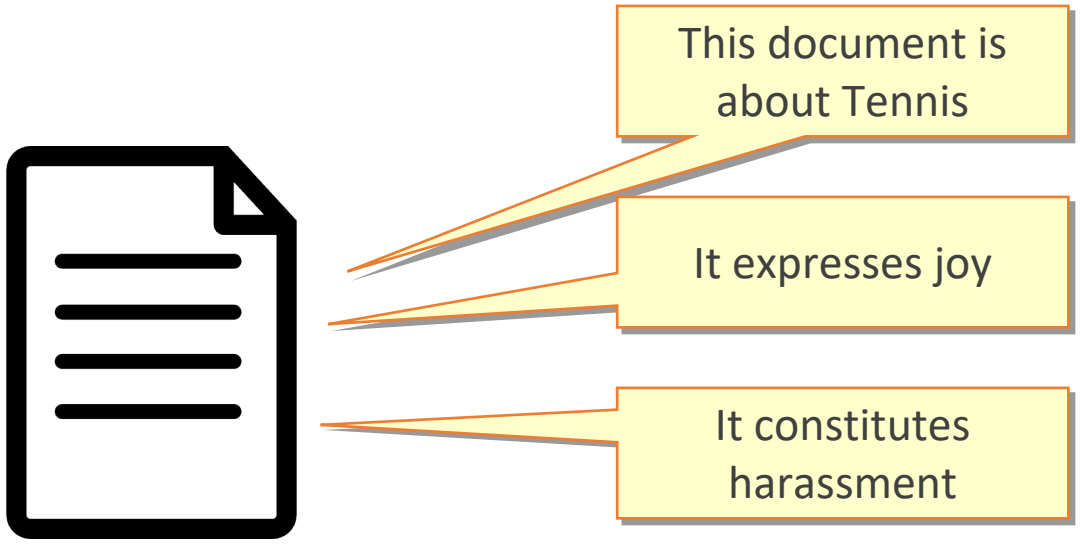
- Label-partially-unseen
  - Based on **Definition 1:** some of the labels have been observed, some were never observed.
    - Evaluate separately on seen and unseen labels.
- Label-fully-unseen
  - Based on **Definition 2:** none of the labels have been observed
  - Following the Dataless protocol from [Chang et al. (AAAI'08)]

We set "label-fully-unseen" because we believe that real-world zero-shot systems should handle the challenge of classifying into a taxonomy without any tasks specific annotation.

A Textual Entailment Approach

Textual Entailment for 0-shot-TC

- Text Classification is inherently a Textual Entailment problem.
- Given a document or a text snippet:
  - Classifying it is equivalent to determining the truth value of a hypothesis
- A program that "knows" how to determine textual entailment can thus classify with respect to any label (hypothesis).



- Assuming, of course, that it "understands" the meaning of the hypothesis.

1. **A** model that "knows" the trick of determining textual entailment – independent of the task at hand

2. **A** model that "understands" the label – part of the input.

(1) Teach it Textual Entailment

- Entailment model learning
  - Tune BERT
  - Data:
    - MNLI (Williams et al., 2018)
    - RTE (GLUE version, Wang et al., 2019)
    - FEVER (Thorne et al., 2018)
  - For “label-partially-unseen”: we fine-tune on the seen annotated data, then evaluate
  - For “label-fully-unseen”: we evaluate directly on the test set

(2) Define the Hypothesis

- We map labels to hypotheses

Labels are templated into a hypothesis as a function of the classification type

aspect	labels	interpretation	example hypothesis	
			word	wordnet definition
topic	sports etc.	this text is about ?	"?"= sports	"?" = an active diversion requiring physical exertion and competition
emotion	anger etc.	this text expresses ?	"?"= anger	"?" = a strong emotion; a feeling that is oriented toward some real or supposed grievance
situation	shelter etc.	The people there need ?	"?"= shelter	"?" = a structure that provides privacy and protection from danger

we first build an interpretation template for each aspect of labels

then we complete the template by filling in the label names or the label's definitions

Challenges of 0shot-TC via Textual Entailment

- How to best map labels to hypotheses?
  - How helpful are label definitions?

	topic				emotion				situation				sum			
	RTE	FEV.	MN.	ens.	RTE	FEV.	MN.	ens.	RTE	FEV.	MN.	ens.	RTE	FEV.	MN.	ens.
word	44.9	42.0	43.4	48.4	12.4	26.7	21.2	18.3	37.7	24.5	14.7	38.3	95.0	93.2	79.3	105.0
def	14.5	25.3	17.2	26.0	3.4	18.7	16.8	9.0	14.1	19.2	11.8	14.4	32.0	63.2	45.8	49.4
comb.	43.8	40.1	37.9	45.7	12.6	24.7	22.3	25.2	37.2	21.0	15.4	38.0	93.6	85.8	81.2	**108.9**

- Directly replacing label names with label definitions needs more attention. It may not work well:
  - label names are mostly common words; definitions are long sentences with, sometimes, complicated words.
  - May introduce some challenges.
  - Example: “sport”: “an active diversion requiring physical exertion and competition”
- Overall, our experiments show that definitions indeed can provide some complementary info, but more advanced models are needed.

Summary

- Thinking about supervision is one of the key tasks we should spend our time on.
- We standardize the study of zero-shot text classification by benchmarking the datasets, evaluation.
- Proposed a textual entailment approach to zero-shot text classification and have shown its effectiveness.
- Dataset & code: https://cogcomp.seas.upenn.edu/page/publication_view/883

Thank You!

- A form of incidental supervision
- Exploits existing TE datasets
- Provides control on how to express (challenging) labels better
- Test-agnostic, real zero-shot