EMNLP'19

Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach

Wenpeng Yin, Jamaal Hay, Dan Roth
Cognitive Computation Group
Department of Computer and Information Science, University of Pennsylvania
{wenpeng,jamaalh,danroth}@seas.upenn.edu

Abstract

Zero-shot text classification (0SHOT-TC) is a challenging NLU problem to which little attention has been paid by the research community. 0SHOT-TC aims to associate an appropriate label with a piece of text, irrespective of the text domain and the aspect (e.g., topic, emotion, event, etc.) described by the label. And there are only a few articles studying 0SHOT-TC, all focusing only on topical categorization which, we argue, is just the tip of the iceberg in 0SHOT-TC. In addition, the chaotic experiments in literature make no uniform comparison, which blurs the progress.

This work benchmarks the 0SHOT-TC problem by providing unified datasets, standardized evaluations, and state-of-the-art baselines. Our contributions include: i) The datasets we provide facilitate studying 0SHOT-TC relative to conceptually different and diverse aspects: the "topic" aspect includes "sports" and "politics" as labels; the "emotion" aspect includes "joy" and "anger"; the "situation" aspect includes "medical assistance" and "water shortage". ii) We extend the existing evaluation setup (*label-partially-unseen*) – given a dataset, train on some labels, test on all labels – to include a more challenging yet realistic evaluation *label-fully-unseen* 0SHOT-TC (Chang et al., 2008), aiming at classifying text snippets without seeing task specific training data at all. iii) We unify the 0SHOT-TC of diverse aspects within a textual entailment formulation and study it this way. [1]

1 Introduction

Supervised text classification has achieved great success in the past decades due to the availability of rich training data and deep learning techniques. However, zero-shot text classification (0SHOT-TC) has attracted little attention despite its great potential in real world applications, e.g., the intent recognition of bank consumers. 0SHOT-TC is challenging because we often have to deal with classes that are compound, ultra-fine-grained, changing over time, and from different aspects such as topic, emotion, etc.

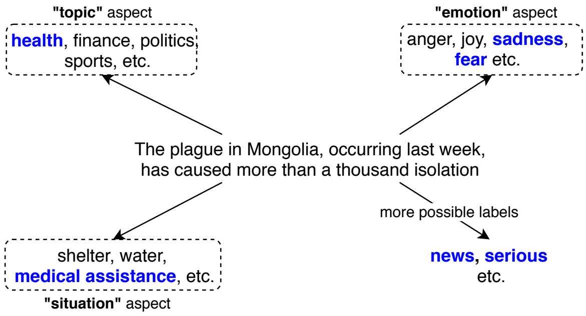

Figure 1: A piece of text can be assigned labels which describe the different aspects of the text. Positive labels are in blue.

Existing 0SHOT-TC studies have mainly the following three problems.

First problem. The 0SHOT-TC problem was modeled in a too restrictive vision. Firstly, most work only explored a single task, which was mainly topic categorization, e.g., (Pushp and Srivastava, 2017; Yogatama et al., 2017; Zhang et al., 2019). We argue that this is only the tiny tip of the iceberg for 0SHOT-TC. Secondly, there is often a precondition that a part of classes are seen and their labeled instances are available to train a model, as we define here as *Definition-Restrictive*:

Definition-Restrictive (0SHOT-TC). Given labeled instances belonging to a set of seen classes S, 0SHOT-TC aims at learning a classifier $f(\cdot) : X \rightarrow Y$, where $Y = S \cup U$; U is a set of unseen classes and belongs to the same aspect as S.

In this work, we formulate the 0SHOT-TC in a

[1] http://cogcomp.org/page/publication_view/883

broader vision. As Figure 1 demonstrates, a piece of text can be assigned labels which interpret the text in different aspects, such as the "topic" aspect, the "emotion" aspect, or the "situation" aspect described in the text. Different aspects, therefore, differ in interpreting the text. For instance, by "topic", it means "this text is about {health, finance $\cdots$}"; by "emotion", it means "this text expresses a sense of {joy, anger, $\cdots$}"; by "situation", it means "the people there need {shelter, medical assistance, $\cdots$}". Figure 1 also shows another essential property of 0SHOT-TC – the applicable label space for a piece of text has no boundary, e.g., "this text is news", "the situation described in this text is serious", etc. Therefore, we argue that we have to emphasize a more challenging scenario to satisfy the real-world problems: seeing no labels, no label-specific training data. Here is our new 0SHOT-TC definition:

*Definition-Wild (*0SHOT-TC*).* 0SHOT-TC aims at learning a classifier $f(\cdot) : X \to Y$, where classifier $f(\cdot)$ never sees Y-specific labeled data in its model development.

Second problem. Usually, conventional text classification denotes labels as indices $\{0,1,2, \cdots, n\}$ without understanding neither the aspect's specific interpretation nor the meaning of the labels. This does not apply to 0SHOT-TC as we can not pre-define the size of the label space anymore, and we can not presume the availability of labeled data. Humans can easily decide the truth value of any upcoming labels because humans can interpret those aspects correctly and understand the meaning of those labels. The ultimate goal of 0SHOT-TC should be to develop machines to catch up with humans in this capability. To this end, making sure the system can understand the described aspect and the label meanings plays a key role.

Third problem. Prior work is mostly evaluated on different datasets and adopted different evaluation setups, which makes it hard to compare them fairly. For example, Rios and Kavuluru (2018) work on medical data while reporting R@K as metric; Xia et al. (2018) work on SNIPS-NLU intent detection data while only unseen intents are in the label-searching space in evaluation.

In this work, we benchmark the datasets and evaluation setups of 0SHOT-TC. Furthermore, we propose a textual entailment approach to handle the 0SHOT-TC problem of diverse aspects in a unified paradigm. To be specific, we contribute in the following three aspects:

Dataset. We provide datasets for studying three aspects of 0SHOT-TC: topic categorization, emotion detection, and situation frame detection – an event level recognition problem. For each dataset, we have standard split for *train*, *dev*, and *test*, and standard separation of seen and unseen classes.

Evaluation. Our standardized evaluations correspond to the *Definition-Restrictive* and *Definition-Wild*. i) *Label-partially-unseen evaluation*. This corresponds to the commonly studied 0SHOT-TC defined in *Definition-Restrictive*: for the set of labels of a specific aspect, given training data for a part of labels, predicting in the full label set. This is the most basic setup in 0SHOT-TC. It checks whether the system can generalize to some labels in the same aspect. To satisfy *Definition-Wild*, we define a new evaluation: ii) *Label-fully-unseen evaluation*. In this setup, we assume the system is unaware of the upcoming aspects and can not access any labeled data for task-specific training.

Entailment approach. Our *Definition-Wild* challenges the system design – how to develop a 0SHOT-TC system, without accessing any task-specific labeled data, to deal with labels from diverse aspects? In this work, we propose to treat 0SHOT-TC as a textual entailment problem. This is to imitate how humans decide the truth value of labels from any aspects. Usually, humans understand the problem described by the aspect and the meaning of the label candidates. Then humans mentally construct a hypothesis by filling a label candidate, e.g., "sports", into the aspect-defined problem "the text is about ?", and ask ourselves if this hypothesis is true, given the text. We treat 0SHOT-TC as a textual entailment problem so that our model can gain knowledge from entailment datasets, and we show that it applies to both *Definition-Restrictive* and *Definition-Wild*.

Overall, this work aims at benchmarking the research of 0SHOT-TC by providing standardized datasets, evaluations, and a state-of-the-art entailment system. All datasets and codes are released.

2 Related Work

ZERO-STC was first explored by the paradigm "Dataless Classification" (Chang et al., 2008).

Dataless classification first maps the text and labels into a common space by Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007), then picks the label with the highest matching score. Dataless classification emphasizes that the representation of labels takes the equally crucial role as the representation learning of text. Then this idea was further developed in (Song and Roth, 2014; Chen et al., 2015; Li et al., 2016a,b; Song et al., 2016).

With the prevalence of word embeddings, more and more work adopts pretrained word embeddings to represent the meaning of words, so as to provide the models with the knowledge of labels (Sappadla et al., 2016; Yogatama et al., 2017; Rios and Kavuluru, 2018; Xia et al., 2018). Yogatama et al. (2017) build generative LSTM to generate text given the embedded labels. Rios and Kavuluru (2018) use label embedding to attend the text representation in the developing of a multi-label classifier. But they report R@K, so it is unclear whether the system can really predict unseen labels. Xia et al. (2018) study the zero-shot intent detection problem. The learned representations of intents are still the sum of word embeddings. But during testing, the intent space includes only new intents; seen intents are not covered. All of these studies can only meet the definition in *Definition-Restrictive*, so they do not really generalize to open aspects of 0SHOT-TC.

Zhang et al. (2019) enrich the embedding representations by incorporating class descriptions, class hierarchy, and the word-to-label paths in ConceptNet. Srivastava et al. (2018) assume that some natural language explanations about new labels are available. Then those explanations are parsed into formal constraints which are further combined with unlabeled data to yield new label oriented classifiers through posterior regularization. However, those explanatory statements about new labels are collected from crowd-sourcing. This limits its application in real world 0SHOT-TC scenarios.

There are a few works that study a specific zero-shot problem by indirect supervision from other problems. Levy et al. (2017) and Obamuyide and Vlachos (2018) study zero-shot relation extraction by converting it into a machine comprehension and textual entailment problem respectively. Then, a supervised system pretrained on an existing machine comprehension dataset or textual entailment dataset is used to do inference. Our work studies the 0SHOT-TC by formulating a broader vision: datasets of multiple apsects and evaluations.

Other zero-shot problems studied in NLP involve entity typing (Zhou et al., 2018), sequence labeling (Rei and Søgaard, 2018), etc.

3 Benchmark the dataset

In this work, we standardize the datasets for 0SHOT-TC for three aspects: topic detection, emotion detection, and situation detection.

For each dataset, we insist on two principles: i) **Label-partially-unseen**: A part of labels are unseen. This corresponds to *Definition-Restrictive*, enabling us to check the performance of unseen labels as well as seen labels. ii) **Label-fully-unseen**: All labels are unseen. This corresponds to *Definition-Wild*, enabling us to check the system performance in test-agnostic setups.

3.1 Topic detection

Yahoo. We use the large-scale Yahoo dataset released by Zhang et al. (2015). Yahoo has 10 classes: {“Society & Culture”, “Science & Mathematics”, “Health”, “Education & Reference”, “Computers & Internet”, “Sports”, “Business & Finance”, “Entertainment & Music”, “Family & Relationships”, “Politics & Government”}, with original split: 1.4M/60k in train/test (all labels are balanced distributed).

We reorganize the dataset by first fixing the *dev* and *test* sets as follows: for *dev*, all 10 labels are included, with 6k labeled instances for each; For *test*, all 10 labels are included, with 10k instances for each. Then training sets are created on remaining instances as follows.

For *label-partially-unseen*, we create two versions of Yahoo *train* for 0SHOT-TC:

- *Train-v0*: 5 classes: {“Society & Culture”, “Health”, “Computers & Internet”, “Business & Finance”, “Family & Relationships”} are included; each is equipped with 130k labeled instances.
- *Train-v1*: 5 classes: { “Science & Mathematics”, “Education & Reference”, “Sports”, “Entertainment & Music”, “Politics & Government”} are included; each is equipped with 130k labeled instances.

We always create two versions of *train* with non-overlapping labels so as to get rid of the

		emotions										
		sad	joy	anger	disgust	fear	surp.	shame	guilt	love	none	sum
domains	tweets	1,500	2,150	1,650	50	2,150	880			1,100	1,000	10,480
	events	300	200	400	400	200		300	300			2,100
	fairytales	300	500	250	120	250	220				1,000	2,640
	arti. sent.	200	150	200	30	100	100					780
	sum	2,300	3,100	2,500	600	2,700	1,200	300	300	1,100	2,000	16,000

Table 1: Emotion *test* in 0SHOT-TC

		emotions										
		sad	joy	anger	disgust	fear	surp.	shame	guilt	love	none	sum
domains	tweets	900	1,050	400	40	1,200	370			400	500	4,860
	events	150	150	150	150	150		100	100			950
	fairytales	150	300	150	90	150	80				500	1,420
	arti. sent.	100	100	100	20	100	50					470
	sum	1,300	1,600	800	300	1,600	500	100	100	400	1,000	7,700

Table 2: Emotion *dev* in 0SHOT-TC

model's over-fitting on one of them.

Label-fully-unseen share the same *test* and *dev* with the *label-partially-unseen* except that it has no training set. It is worth mentioning that our setup of *label-partially-unseen* and *label-fully-unseen* enables us to compare the performance mutually; it can show the system's capabilities while seeing different sizes of classes.

3.2 Emotion detection

UnifyEmotion. This emotion dataset was released by Bostan and Klinger (2018). It was constructed by unifying the emotion labels of multiple public emotion datasets[2]. This dataset consists of text from multiple domains: tweet, emotional events, fairy tale and artificial sentences, and it contains 9 emotion types {"sadness", "joy", "anger", "disgust", "fear", "surprise", "shame", "guilt", "love"} and "none" (if no emotion applies). We remove the multi-label instances (appro. 4k) so that the remaining instances always have a single positive label. The official evaluation metric is *label-weighted* F1.

Since the labels in this dataset has unbalanced distribution. We first directly list the fixed *test* and *dev* in Table 1 and Table 2, respectively. They are shared by following *label-partial-unseen* and *label-fully-unseen* setups of *train*.

Label-partial-unseen has the following two versions of *train*:

- *Train-v0*: 5 classes: {"sadness", "anger", "fear", "shame", "love"} are included.
- *Train-v1*: 4 classes: { "joy", "disgust", "surprise", "guilt"} are included.

For *label-fully-unseen*, no training set is provided.

3.3 Situation detection

The situation frame typing is one example of an event-type classification task. A situation frame studied here is a *need* situation such as the need for water or medical aid, or an *issue* situation such as crime violence (Strassel et al., 2017; Muis et al., 2018). It was originally designed for low-resource situation detection, where annotated data is unavailable. This is why it is particularly suitable for 0SHOT-TC.

We use the Situation Typing dataset released by Mayhew et al. (2019). It has 5,956 labeled instances. Totally 11 situation types: "food supply", "infrastructure", "medical assistance", "search/rescue", "shelter", "utilities, energy, or sanitation", "water supply", "evacuation", "regime change", "terrisms", "crime violence" and an extra type "none" – if none of the 11 types applies. This dataset is a *multi-label* classification, and *label-wise weighted F1* is the official evaluation.

The *train*, *test* and *dev* are listed in Table 3.

[2]Please refer to (Bostan and Klinger, 2018) for more details about the constituent datasets.

Summary of 0SHOT-TC datasets. Our three datasets covers single-label classification (i.e., "topic" and "emotion") and multi-label classification (i.e., "situation"). In addition, a "none" type is adopted in "emotion" and "situation" tasks if no predefined types apply – this makes the problem more realistic.

4 Benchmark the evaluation

How to evaluate a 0SHOT-TC system? This needs to review the original motivation of doing 0SHOT-TC research. As we discussed in Introduction section, ideally, we aim to build a system that works like humans – figuring out if a piece of text can be assigned with an open-defined label, without any constrains on the domains and the aspects described by the labels. Therefore, we challenge the system in two setups: *label-partially-unseen* and *label-fully-unseen*.

Label-partially-unseen. This is the most common setup in existing 0SHOT-TC literature: for a given dataset of a specific problem such as topic categorization, emotion detection, etc, train a system on a part of the labels, then test on the whole label space. Usually all labels describe the same aspect of the text.

Label-fully-unseen. In this setup, we push "zero-shot" to the extreme – no annotated data for any labels. So, we imagine that learning a system through whatever approaches, then testing it on 0SHOT-TC datasets of open aspects.

This *label-fully-unseen* setup is more like the dataless learning principle (Chang et al., 2008), in which no task-specific annotated data is provided for training a model (since usually this kind of model fails to generalize in other domains and other tasks), therefore, we are encouraged to learn models with open-data or test-agnostic data. In this way, the learned models behave more like humans.

5 An entailment model for 0SHOT-TC

As one contribution of this work, we propose to deal with 0SHOT-TC as a textual entailment problem. It is inspired by: i) text classification is essentially a textual entailment problem. Let us think about how humans do classification: we mentally think "whether this text is about sport?", or "whether this text expresses a specific feeling?", or "whether the people there need water supply?" and so on. The reason that conventional text classification did not employ entailment approach is it always has pre-defined, fixed-size of classes equipped with annotated data. However, in 0SHOT-TC, we can neither estimate how many and what classes will be handled nor have annotated data to train class-specific parameters. Textual entailment, instead, does not preordain the boundary of the hypothesis space. ii) To pursue the ideal generalization of classifiers, we definitely need to make sure that the classifiers understand the problem encoded in the aspects and understand the meaning of labels. Conventional supervised classifiers fail in this aspect since label names are converted into indices – this means the classifiers do not really understand the labels, let alone the problem. Therefore, exploring 0SHOT-TC as a textual entailment paradigm is a reasonable way to achieve generalization.

Convert labels into hypotheses. The first step of dealing with 0SHOT-TC as an entailment problem is to convert labels into hypotheses. To this end, we first convert each aspect into an *interpretation* (we discussed before that generally one aspect defines one interpretation). E.g., "topic" aspect to interpretation "the text is about the topic". Table 4 lists some examples for the three aspects: "topic", "emotion" and "situation".

In this work, we just explored two simple methods to generate the hypotheses. As Table 4 shows, one is to use the label name to complete the interpretation, the other is to use the label's definition in WordNet to complete the interpretation. In testing, once one of them results in an "entailment" decision, then we decide the corresponding label is positive. We can definitely create more natural hypotheses through crowd-sourcing, such as "food" into "the people there are starving". Here we just set the baseline examples by automatic approaches, more explorations are left as future work, and we welcome the community to contribute.

Convert classification data into entailment data. For a data split (*train*, *dev* and *test*), each input text, acting as the premise, has a positive hypothesis corresponding to the positive label, and all negative labels in the data split provide negative hypotheses. Note that unseen labels do not provide negative hypotheses for instances in *train*.

		situations											
		search	evac	infra	utils	water	shelter	med	food	reg.	terr.	crim.	none
	total size	327	278	445	412	492	659	1,046	810	80	348	983	1,868
split	test	190	166	271	260	289	396	611	472	51	204	590	1,144
	dev	137	112	174	152	203	263	435	338	29	144	393	724
	train-v0	327	–	445	–	492	–	1,046	–	80	–	983	–
	train-v1	–	278	–	412	–	659	–	810	–	348	–	–

Table 3: Situation *train*, *dev* and *test* split for 0SHOT-TC.

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

Table 4: Example hypotheses we created for modeling different aspects of 0SHOT-TC.

Entailment model learning. In this work, we make use of the widely-recognized state of the art entailment technique – BERT (Devlin et al., 2019), and train it on three mainstream entailment datasets: MNLI (Williams et al., 2018), GLUE RTE (Dagan et al., 2005; Wang et al., 2019) and FEVER[3] (Thorne et al., 2018), respectively. We convert all datasets into binary case: "entailment" vs. "non-entailment", by changing the label "neutral" (if exist in some datasets) into "non-entailment".

For our *label-fully-unseen* setup, we directly apply this pretrained entailment model on the test sets of all 0SHOT-TC aspects. For *label-partially-unseen* setup in which we intentionally provide annotated data, we first pretrain BERT on the MNLI/FEVER/RTE, then fine-tune on the provided training data.

Harsh policy in testing. Since seen labels have annotated data for training, we adopt different policies to pick up seen and unseen labels. To be specific, we pick a seen label with a harsher rule: i) In single-label classification, if both seen and unseen labels are predicted as positive, we pick the seen label only if its probability of being positive is higher than that of the unseen label by a hyperparameter α. If only seen or unseen labels are predicted as positive, we pick the one with the highest probability; ii) In multi-label classification, if both seen and unseen labels are predicted as positive, we change the seen labels into "negative" if their probability of being positive is higher than that of the unseen label by less than α. Finally, all labels labeled positive will be selected. If no positive labels, we choose "none" type.

α = 0.05 in our systems, tuned on *dev*.

6 Experiments

6.1 Label-partially-unseen evaluation

In this setup, there is annotated data for partial labels as *train*. So, we report performance for unseen classes as well as seen classes. We compare our entailment approaches, trained separately on MNLI, FEVER and RTE, with the following baselines.

Baselines.

- *Majority*: the text picks the label of the largest size.
- *ESA*: A dataless classifier proposed in (Chang et al., 2008). It maps the words (in text and label names) into the title space of Wikipedia articles, then compares the text with label names. This method does not rely on *train*.

[3]FEVER is an evidential claim verification problem: given a hypothesis, the system needs to identify evidence sentences as premise, then gives the entailment decision. We use the ground truth evidence as premises in this work.

		topic				emotion				situation			
		v0		v1		v0		v1		v0		v1	
		s	u	s	u	s	u	s	u	s	u	s	u
w/o *train*	Majority	0.0	10.0	10.0	0.0	0.0	13.3	18.5	0.0	0.0	19.7	0.0	16.4
	Word2Vec	28.1	43.3	43.3	28.1	8.1	5.4	6.2	7.3	10.3	24.7	8.6	23.1
	ESA	27.5	30.1	30.1	27.5	6.7	9.7	5.5	9.2	22.8	28.5	22.4	27.7
supervised train	Binary-BERT	**72.6**	44.3	**80.6**	34.9	**35.6**	17.5	**37.1**	14.2	72.4	48.4	63.8	42.9
	our entail												
	MNLI	70.9	**52.1**	77.3	**45.3**	33.4	**26.6**	33.9	21.4	**74.8**	**53.4**	**68.4**	**47.8**
	FEVER	70.2	51.7	77.2	42.7	31.9	24.5	26.0	**22.5**	73.5	47.6	65.7	43.6
	RTE	71.5	45.3	78.6	40.6	32.0	21.8	32.7	21.1	72.8	52.0	65.0	45.2

Table 5: Label-partially-unseen evaluation. "v0/v1" means the results in that column are obtained by training on train-v0/v1. "s": seen labels; "u": unseen labels. "Topic" uses *acc.*, both "emotion" and "situation" use *label-wised weighted F1*. Note that for baselines "Majority", "Word2Vec" and "ESA", they do not have *seen* labels; we just separate their numbers into *seen* and *unseen* subsets of supervised approaches for clear comparison.

	topic	emotion	situation	sum
Majority	10.0	5.9	11.0	26.9
Word2Vec	35.7	6.9	15.6	58.2
ESA	28.6	8.0	26.0	62.6
Wiki-based	**52.1**	21.2	27.7	101.0
our entail.				
MNLI	37.9	22.3	15.4	75.6
FEVER	40.1	24.7	21.0	85.8
RTE	43.8	12.6	37.2	93.6
ensemble	45.7	**25.2**	**38.0**	**108.9**

Table 6: Label-fully-unseen evaluation.

We implemented ESA based on 08/01/2019 Wikipedia dump[4]. There are about 6.1M words and 5.9M articles.

- *Word2Vec*[5] (Mikolov et al., 2013): Both the representations of the text and the labels are the addition of word embeddings element-wisely. Then cosine similarity determines the labels. This method does not rely on *train* either.

- *Binary-BERT*: We fine-tune BERT[6] on *train*, which will yield a binary classifier for entailment or not; then we test it on *test* – picking the label with the maximal probability in single-label scenarios while choosing all the labels with "entailment" decision in multi-label cases.

Discussion. The results of *label-partially-unseen* are listed in Table 5. "ESA" performs slightly worse than "Word2Vec" in topic detection, mainly because the label names, i.e., topics such as "sports", are closer than some keywords such as "basketball" in Word2Vec space. However, "ESA" is clearly better than "Word2Vec" in situation detection; this should be mainly due to the fact that the label names (e.g., "shelter", "evaculation", etc.) can hardly find close words in the text by Word2Vec embeddings. Quite the contrary, "ESA" is easier to make a class such as "shelter" closer to some keywords like "earthquake". Unfortunately, both Word2Vec and ESA work poorly for emotion detection problem. We suspect that emotion detection requires more entailment capability. For example, the text snippet "when my brother was very late in arriving home from work", its gold emotion "fear" requires some common-knowledge inference, rather than just word semantic matching through Word2Vec and ESA.

The supervised method "Binary-BERT" is indeed strong in learning the seen-label-specific models – this is why it predicts very well for seen classes while performing much worse for unseen classes.

Our entailment models, especially the one pretrained on MNLI, generally get competitive performance with the "Binary-BERT" for *seen* (slightly worse on "topic" and "emotion" while clearly better on "situation") and improve the performance regarding *unseen* by large margins. At this stage, fine-tuning on an MNLI-based pre-

[4] https://dumps.wikimedia.org/enwiki/

[5] https://code.google.com/archive/p/word2vec/

[6] We always use "bert-base-uncased" version.

	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**

Table 7: Fine-grained *label-fully-unseen* performances of different hypothesis generation approaches "word", "def" (definition) and "comb" (word&definition) on the three tasks ("topic", "emotion" and "situation") based on three pretrained entailment models (RTE, FEVER, MNLI) and the ensemble approach (ens.). The last column *sum* contains the addition of its preceding three blocks element-wisely.

trained entailment model seems more powerful.

6.2 Label-fully-unseen evaluation

Regarding this *label-fully-unseen* evaluation, apart from our entailment models and three unsupervised baselines "Majority", "Word2Vec" and "ESA", we also report the following baseline:

Wikipedia-based: We train a *binary* classifier based on BERT on a dataset collected from Wikipedia. Wikipedia is a corpus of general purpose, without targeting any specific 0SHOT-TC task. Collecting categorized articles from Wikipedia is popular way of creating training data for text categorization, such as (Zhou et al., 2018). More specifically, we collected 100K articles along with their categories in the bottom of each article. For each article, apart from its attached positive categories, we randomly sample three negative categories. Then each article and its positive/negative categories act as training pairs for the binary classifier.

We notice "Wikipedia-based" training indeed contributes a lot for the topic detection task; however, its performances on emotion and situation detection problems are far from satisfactory. We believe this is mainly because the Yahoo-based topic categorization task is much closer to the Wikipedia-based topic categorization task; emotion and situation categorizations, however, are relatively further.

Our entailment models, pretrained on MNLI/FEVER/RTE respectively, perform more robust on the three 0SHOT-TC aspects (except for the RTE on emotion). Recall that they are not trained on any text classification data, and never know the domain and the aspects in the *test*. This clearly shows the great promise of developing textual entailment models for 0SHOT-TC. Our ensemble approach[7] further boosts the performances on all three tasks.

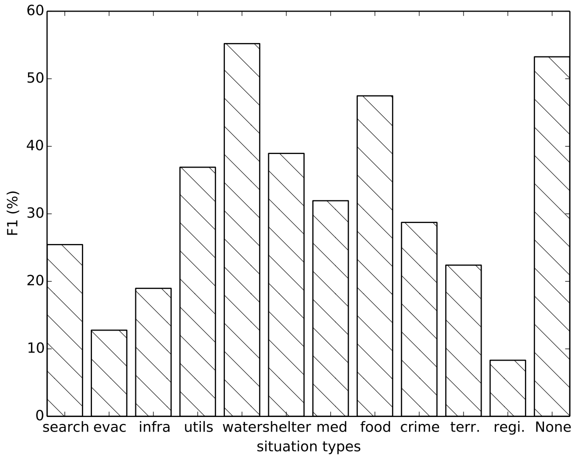

Figure 2: Performance of different situation classes in *label-fully-unseen, predicted by the ensemble model.*

An interesting phenomenon, comparing the *label-partially-unseen* results in Table 5 and the *label-fully-unseen* results in Table 6, is that the pretrained entailment models work in this order for *label-fully-unseen* case: RTE > FEVER >MNLI; on the contrary, if we fine-tune them on the label-partially-unseen case, the MNLI-based model performs best. This could be due to a possibility that, on one hand, the constructed situation entailment dataset is closer to the RTE dataset than to the MNLI dataset, so an RTE-based model can generalize well to situation data, but, on the other hand, it could also be more likely to over-fit the training set of "situation" during fine-tuning. A deeper exploration of this is left as future work.

6.3 How do the generated hypotheses influence

In Table 4, we listed examples for converting class names into hypotheses. In this work, we only tried to make use of the class names and their defini-

[7]For each input pair of the entailment model, we sum up their probabilities after softmax, then do softmax to get new probabilities.

tions in WordNet. Table 7 lists the fine-grained performance of three ways of generating hypotheses: "word", "definition", and "combination" (i.e., word&definition).

This table indicates that: i) Definition alone usually does not work well in any of the three tasks, no matter which pretrained entailment model is used; ii) Whether "word" alone or "word&definition" works better depends on the specific task and the pretrained entailment model. For example, the pretrained MNLI model prefers "word&definition" in both "emotion" and "situation" detection tasks. However, the other two entailment models (RTE and FEVER) mostly prefer "word". iii) Since it is unrealistic to adopt only one entailment model, such as from {RTE, FEVER, MNLI}, for any open 0SHOT-TC problem, an ensemble system should be preferred. However, the concrete implementation of the ensemble system also influences the strengths of different hypothesis generation approaches. In this work, our ensemble method reaches the top performance when combining the "word" and "definition". More ensemble systems and hypothesis generation paradigms need to be studied in the future.

To better understand the impact of generated hypotheses, we dive into the performance of each labels, taking "situation detection" as an example. Figure 2 illustrates the separate F1 scores for each situation class, predicted by the ensemble model for *label-fully-unseen* setup. This enables us to check in detail how easily the constructed hypotheses can be understood by the entailment model. Unfortunately, some classes are still challenging, such as "evacuation", "infrastructure", and "regime change". This should be attributed to their over-abstract meaning. Some classes were well recognized, such as "water", "shelter", and "food". One reason is that these labels mostly are common words – systems can more easily match them to the text; the other reason is that they are situation classes with higher frequencies (refer to Table 3) – this is reasonable based on our common knowledge about disasters.

7 Summary

In this work, we analyzed the problems of existing research on zero-shot text classification (0SHOT-TC): restrictive problem definition, the weakness in understanding the problem and the labels' meaning, and the chaos of datasets and evaluation setups. Therefore, we are benchmarking 0SHOT-TC by standardizing the datasets and evaluations. More importantly, to tackle the broader-defined 0SHOT-TC, we proposed a textual entailment framework which can work with or without the annotated data of *seen* labels.

Acknowledgments

The authors would like to thank Jennifer Sheffield and the anonymous reviewers for insightful comments and suggestions. This work was supported by Contracts HR0011-15-C-0113 and HR0011-18-2-0052 with the US Defense Advanced Research Projects Agency (DARPA). Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

Laura Ana Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *COLING*, pages 2104–2119.

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *AAAI*, pages 830–835.

Xingyuan Chen, Yunqing Xia, Peng Jin, and John A. Carroll. 2015. Dataless text classification with descriptive LDA. In *AAAI*, pages 2224–2231.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop*, pages 177–190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, pages 1606–1611.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *CoNLL*, pages 333–342.

Chenliang Li, Jian Xing, Aixin Sun, and Zongyang Ma. 2016a. Effective document labeling with very few seed words: A topic model approach. In *CIKM*, pages 85–94.

Yuezhang Li, Ronghuo Zheng, Tian Tian, Zhiting Hu, Rahul Iyer, and Katia P. Sycara. 2016b. Joint embedding of hierarchical categories and entities for concept categorization and dataless classification. In *COLING*, pages 2678–2688.

Stephen Mayhew, Tatiana Tsygankova, Francesca Marini, Zihan Wang, Jane Lee, Xiaodong Yu, Xingyu Fu, Weijia Shi, Zian Zhao, Wenpeng Yin, Karthikeyan K, Jamaal Hay, Michael Shur, Jennifer Sheffield, and Dan Roth. 2019. University of pennsylvania lorehlt 2019 submission.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Aldrian Obaja Muis, Naoki Otani, Nidhi Vyas, Ruochen Xu, Yiming Yang, Teruko Mitamura, and Eduard Hovy. 2018. Low-resource cross-lingual event type detection via distant supervision with minimal effort. In *COLING*, pages 70–82.

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78.

Pushpankar Kumar Pushp and Muktabh Mayank Srivastava. 2017. Train once, test anywhere: Zero-shot learning for text classification. *CoRR*, abs/1712.05972.

Marek Rei and Anders Søgaard. 2018. Zero-shot sequence labeling: Transferring knowledge from sentences to tokens. In *NAACL-HLT*, pages 293–302.

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *EMNLP*, pages 3132–3142.

Prateek Veeranna Sappadla, Jinseok Nam, Eneldo Loza Menc'ia, and Johannes Fürnkranz. 2016. Using semantic similarity for multi-label zero-shot classification of text documents. In *24th European Symposium on Artificial Neural Networks*, pages 423–428.

Yangqiu Song and Dan Roth. 2014. On dataless hierarchical text classification. In *AAAI*, pages 1579–1585.

Yangqiu Song, Shyam Upadhyay, Haoruo Peng, and Dan Roth. 2016. Cross-lingual dataless classification for many languages. In *IJCAI*, pages 2901–2907.

Shashank Srivastava, Igor Labutov, and Tom M. Mitchell. 2018. Zero-shot learning of classifiers from natural language quantification. In *ACL*, pages 306–316.

Stephanie M. Strassel, Ann Bies, and Jennifer Tracey. 2017. Situational awareness for low resource languages: the LORELEI situation frame annotation task. In *SMERP@ECIR*, pages 32–41.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In *NAACL-HLT*, pages 809–819.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*, pages 1112–1122.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S. Yu. 2018. Zero-shot user intent detection via capsule neural networks. In *EMNLP*, pages 3090–3099.

Dani Yogatama, Chris Dyer, Wang Ling, and Phil Blunsom. 2017. Generative and discriminative text classification with recurrent neural networks. *CoRR*, abs/1703.01898.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. 2019. Integrating semantic knowledge to tackle zero-shot text classification. In *NAACL*.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*, pages 649–657.

Ben Zhou, Daniel Khashabi, Chen-Tse Tsai, and Dan Roth. 2018. Zero-shot open entity typing as type-compatible grounding. In *EMNLP*, pages 2065–2076.