OpenStance: Real-world Zero-shot Stance Detection

Hanzi Xu, Slobodan Vucetic and Wenpeng Yin
Temple University
{hanzi.xu; slobodan.vucetic; wenpeng.yin}@temple.edu

Abstract

Prior studies of zero-shot stance detection identify the attitude of texts towards unseen topics occurring in the same document corpus. Such task formulation has three limitations: (i) Single domain/dataset. A system is optimized on a particular dataset from a single domain; therefore, the resulting system cannot work well on other datasets; (ii) the model is evaluated on a limited number of unseen topics; (iii) it is assumed that part of the topics has rich annotations, which might be impossible in real-world applications. These drawbacks will lead to an impractical stance detection system that fails to generalize to open domains and open-form topics.

This work defines OpenStance: open-domain zero-shot stance detection, aiming to handle stance detection in an open world with neither domain constraints nor topic-specific annotations. The key challenge of OpenStance lies in the open-domain generalization: learning a system with fully unspecific supervision but capable of generalizing to any dataset. To solve OpenStance, we propose to combine indirect supervision, from textual entailment datasets, and weak supervision, from data generated automatically by pre-trained Language Models. Our single system, without any topic-specific supervision, outperforms the supervised method on three popular datasets. To our knowledge, this is the first work that studies stance detection under the open-domain zero-shot setting. All data and code are publicly released.1

1 Introduction

Stance detection differentiates the attitude (e.g., support, oppose, or neutral) of a text towards a topic (Walker et al., 2012a). The topic can be a phrase or a complete sentence. The same text can express the author’s positions on many different topics. For example, a tweet on climate warming may also express attitudes about environmental policies as well as the debate between electric or fuel cars. Such compound expression can be seen on all online platforms, including News outlets, Twitter, blogs, etc. Therefore, stance detection can be a complicated task that is essential for developing the inference capability of NLP models as well as other disciplines such as politics, journalism, etc.

Since the textual expressions and the size of topics in the real world are unpredictable, zero-shot stance detection has become the mainstream research direction in this area: topics in the test set are unseen during training. For example, Mohammad et al. (2016) created a dataset SemT6 based on tweets with six noun phrases as topics. One of the topics was reserved for testing and the remaining were used for training. Allaway and McKeown (2020) extended the topic size on the domain of news comments by covering 4,000 topics in training and 600 unseen topics in testing.

However, despite the change in the domain and topic size, there are three major limitations in previous studies which make the task not a real zero-shot task: (i) the dataset only contains texts from a single domain, such as news comments in VAST (Allaway and McKeown, 2020) and tweets in SemT6 (Mohammad et al., 2016); (ii) most literature studied only a limited size of topics with a single textual form (either noun phrases or sentential claims), e.g., (Mohammad et al., 2016; Conforti et al., 2020); (iii) rich annotation for at least part of the topics is always required, which is not possible in real-world applications because data collection can be very time-consuming and costly (Enayati et al., 2021). Those limitations lead to an impractical zero-shot stance detection system that cannot generalize well to unseen domains and open-form topics.

In this work, we re-define what a zero-shot stance detection should be. Specifically, we define OpenStance: an open-domain zero-shot stance detection.
detection, aiming to build a system that can work in the real world without any specific attention to the text domains or topic forms. More importantly, no task-specific supervision is needed. To achieve this, we propose to combine two types of supervision: indirect supervision and weak supervision. The indirect supervision comes from textual entailment—we treat the stance detection problem as a textual entailment task since the attitude toward a topic should be inferred from the input text. Therefore, the existing entailment datasets, such as MNLI (Williams et al., 2018), can contribute supervision to the zero-shot setting. To collect supervision that is more specific to the OpenStance task, we design two MASK choices (MASK-topic and MASK-text) to prompt GPT-3 (Brown et al., 2020) to generate weakly supervised data. Given an input text and a stance label (support, oppose, or neutral), MASK-topic predicts what topic is appropriate based on the content; given a topic and a label, MASK-text seeks the text that most likely holds this stance. The collection of weakly supervised data only needs the unlabeled texts and the set of topics that users want to include. The joint power of indirect supervision and weak supervision will be evaluated on VAST, SemT6 and Perspective (Chen et al., 2019), three popular datasets that cover distinct domains, different sizes and diverse textual forms of topics. Experimental results show that although no task-specific supervision is used, our system can get robust performance on all three datasets, even outperforming the task-specific supervised models (72.6 vs. 69.3 by mean F1 over the three datasets).

Our contributions are threefold: (i) we define OpenStance, an open-domain zero-shot stance detection task, that fulfills real-world requirements while having never been studied before; (ii) we design a novel masking mechanism to let GPT-3 generate weakly supervised data for OpenStance. This mechanism can inspire other NLP tasks that detect relations between two pieces of texts; (iii) our approach, integrating indirect supervision and weak supervision, demonstrates outstanding generalization among three datasets that cover a wide range of text domains, topic sizes and topic forms.

2 Related Work

Stance detection. Stance detection, as a recent member of the NLP family, was mainly driven by newly created datasets. In the past studies, datasets have been constructed from diverse domains like online debate forums (Walker et al., 2012b; Hasan and Ng, 2014; Abbott et al., 2016), news comments (Krejzl et al., 2017; Lozhnikov et al., 2018), Twitter (Mohammad et al., 2016; Küçük, 2017; Tsakalidis et al., 2018), etc.

Zero-shot stance detection. Recently, researchers started to work on zero-shot stance detection in order to build a system that can handle unseen topics. Most work split the collected topic-aware annotations into train and test within the same domain. Allaway and McKeown (2020) made use of topic similarity to connect unseen topics with seen topics. Allaway et al. (2021) designed adversarial learning to learn domain-independent information and topic-invariant representations. Similarly, Wang and Wang (2021) applied adversarial learning to extract stance-related but domain-invariant features existed among different domains. Liu et al. (2021) utilized common sense knowledge from ConceptNet (Speer et al., 2017) to introduce extra knowledge of the relations between the texts and topics. Most prior systems worked on a single domain and were tested on a small number of unseen topics. Li et al. (2021) tried to test on various unseen datasets by jointly optimizing on multiple training datasets. However, they still assumed that part of the topics or domains has rich annotations. In contrast, our goal is to design a system that can handle stance detection in an open world without requiring any domain constraints or topic-specific annotations.

Textual entailment as indirect supervision. Textual entailment studies if a hypothesis can be entailed by a premise; this was proposed as a unified inference framework for a wide range of NLP problems (Dagan et al., 2005). Recently, textual entailment is widely utilized to help solve many tasks, such as few-shot intent detection (Xia et al., 2021), ultra-fine entity typing (Li et al., 2022), coreference resolution (Yin et al., 2020), relation extraction (Xia et al., 2021; Sainz et al., 2021), event argument extraction (Sainz et al., 2022), etc. As far as we know, our work is the first one that successfully leverages the indirect supervision from textual entailment for stance detection.

Weak supervision from GPT-3. As the currently most popular and (arguably) well-behaved pre-trained language model, GPT-3 (Brown et al., 2020) has been a great success on few-shot and
zero-shot NLP. As an implicit knowledge base fully in the form of parameters, it is not surprising that researchers attempt to extract knowledge from it to construct synthetic data, e.g., (Yoo et al., 2021; Wang et al., 2021). We use GPT-3 to collect distantly supervised data by two novel masking mechanisms designed specifically for the OpenStance.

3 Problem definition

OpenStance has the following requirements:

- An instance includes three items: text \(s\), topic \(t\) and a stance label \(l\) \((l \in \{\text{support}, \text{oppose}, \text{neutral}\})\); the task is to learn the function \(f(s, t) \rightarrow l\);

- The text \(s\) can come from any domain; the topic \(t\) can be any textual expressions, such as a noun phrase “gun control” or a sentential claim “climate change is a real concern”;

- All labeled instances \(\{(s, t, l)\}\) only exist in test; no train or dev is provided;

- Previous work used different metrics for the evaluation. For example, VAST (Allaway and McKeown, 2020) used macro-averaged F1 regarding stance labels, while studies on SemT6 (Allaway et al., 2021; Liang et al., 2022) reported the F1 scores per topic. To make systems be comparable, we unify the evaluation and use the label-oriented macro F1 as our main metric.

**OpenStance vs. prior zero-shot stance detection.** Prior studies of zero-shot stance detection worked on a single dataset \(D^i\) in which all texts \(s\) come from the same domain. Topics \(t\) in the dataset are split into train, dev and test disjointly. The main issue is that a model that fits \(D^i\) does not work well on a new dataset \(D^j\) that may contain \(s\) of different domains and unseen \(t\). For example, a model trained on VAST can only get F1 49.0% on Perspectrum, which is around the performance of random guess. OpenStance aims at handling multiple datasets of open domains and open-form topics without looking at their train and dev.

**OpenStance vs. textual entailment.** Stance detection is essentially a textual entailment problem if we treat the text \(s\) as the premise, and the stance towards the topic \(t\) as the hypothesis. This motivates us to use indirect supervision from textual entailment to deal with the stance detection problem. Nevertheless, there are two distinctions between them: (i) even though we can match \(l\) of stance detection with the labels of textual entailment: support \(\rightarrow\) entailment, oppose \(\rightarrow\) contradict and neutral \(\rightarrow\) neutral, whether a topic \(t\) in stance detection can be treated as a hypothesis depends on the text form of \(t\). If \(t\) is a noun phrases such as “gun control”, \(t\) cannot act as a hypothesis alone as there is no stance in it; if \(t\) is a sentential claim such as “climate change is a real concern”, inferring the truth value of this hypothesis is exactly a textual entailment problem. This observation motivates us to test OpenStance on topics of both phrase forms and sentence forms; (ii) Zero-shot textual entailment means the size of the annotated instances for labels is zero, while OpenStance requires the topics have zero labeled examples.

4 Methodology

This section introduces how we collect and combine indirect supervision and weak supervision to solve OpenStance.

**Indirect Supervision.** As we discussed in Section 3, stance detection is a case of textual entailment since the stance \(l\) towards a topic \(t\) should be inferred from the text \(s\). To handle the zero-shot challenge in OpenStance, textual entailment is a natural choice for indirect supervision.

Specifically, we first cast stance detection instances into the textual entailment format by combining \(l\) and \(t\) as a sentential hypothesis \(h\), such as “it supports topic \(p\)”, and treating the \(s\) as the premise \(p\); then a pretrained model on MNLI (Williams et al., 2018), one of the largest entailment dataset, is ready to predict the relationship between the \(p\) and \(h\). An entailed (resp. contradicted or neutral) \(h\) means the topic \(t\) is supported (resp. opposed or neutral) by the text \(s\).

Unfortunately, the indirect supervision from textual entailment may not perform well enough in real-world OpenStance considering the widely known brittleness of pretrained entailment models and the open domains and open-form topics in OpenStance. Therefore, in addition to the indirect supervision from textual entailment, we will collect weak supervision that is aligned better with the texts \(\{x\}\) and the topics \(\{t\}\).
Weak Supervision. For the next step, we would like to create some weakly supervised data using easily available resources to obtain a better understanding of the target task. We used GPT-3 (Brown et al., 2020), a pre-trained autoregressive language model that can perform text completion at (arguably) a near-human level, to help us create some weakly labeled instances.

We form incomplete sentences using prompts, and let the GPT-3 complete them. Since a stance label \( l \) connects the text \( s \) and the topic \( t \) and such connection is unavailable in a zero-shot setting, the construction of incomplete sentences is driven by two questions: (i) given an input text \( s \) and a stance, e.g., support, what topics are supported by \( s \)? (ii) given a topic and a stance, for example, support, what texts support this topic? As a result, there are two kinds of prompts: \( \text{MASK-Topic} \) and \( \text{MASK-Text} \). To implement the two masking mechanisms, we need to prepare three sets: the raw texts \( \{s\} \), a set of topics \( \{t\} \), and the known stance labels \{support, oppose, neutral\}. It is noteworthy that no topic-specific human annotations are used here.

\( \text{MASK-Topic} \): In this masking framework, we randomly choose a text from \( \{s\} \) and a stance label from \{support, oppose, neutral\}, then build the prompt as:

\[
\text{S/he claims text, so s/he label the idea of } \text{MASK}
\]

For example, when the text is “Coldest and wettest summer in memory” and the label is oppose, the prompt would be “S/he claims coldest and wettest summer in memory, so s/he opposes the idea of”. Then, this prompt is fed into GPT-3, and the completion “global warming” would be the predicted topic.

\( \text{MASK-Text} \): In this case, we randomly choose a topic from \( \{t\} \) and a stance label towards it, then build the prompt as:

\[
\text{His/her attitude towards topic is label because s/he thinks MASK}
\]

For example, when the topic is “climate change is a real concern”, the label is “oppose”, the completed sentence filled by GPT-3 could be “His attitude towards climate change is a real concern is opposition because s/he thinks the science behind climate change is not settled”.

For any dataset of stance detection, we first collect the three sets (i.e., \( \{s\} \), \( \{t\} \), and \( \{l\} \)) from the label-free training set without peeking at any gold annotations, then use \( \text{MASK-Topic} \) and \( \text{MASK-Text} \) prompts to generate equal number of weakly supervised examples. We will study which masking scheme is more effective in experiments. In addition, to have a fair comparison with supervised methods that learn on the \text{train} of a task, we make sure our generated weakly supervised data has the same size as the \text{train} for any target task.

Although noise is common in weakly supervised data, GPT-3 performs badly on neutral completions for both \( \text{MASK-Topic} \) and \( \text{MASK-Text} \) tasks. This is not a surprise for the \( \text{MASK-Topic} \) since the GPT-3 is asked to provide a topic that the given text has a neutral attitude for, while most texts, obtained from unlabeled \text{train} and originally extracted from social networks, usually express a strong attitude. Furthermore, in \( \text{MASK-Text} \), even though the GPT-3 can output a text given the neutral label towards a topic, the response is very general and does not provide any insights. For example, when the template is "His attitude towards high school writing skills is neutral because he thinks [MASK]", GPT-3 fills out the MASK with "that they are important but not essential." Obviously, it is much easier to generate text with a clear attitude compared to a neutral stance. On the one hand, GPT-3 may not really understand what a neutral stance is. On the other hand, even humans cannot easily write a neutral opinion towards a topic. Since the quality of generated neutral instances is not very promising, we take the same approach as how VAST (Allaway and McKeown, 2020) collected its neutral samples: matching texts with random topics in the dataset.

Training strategy. To keep consistent format and make full use of the entailment reasoning framework, we convert all phrase-form topic in the weak supervision data into a sentence-form hypothesis with the positive stance, i.e., "he is in favor of topic" (note that this does not change the original label). Then, we randomly split the weak supervision data as \text{train} (80\%) and \text{dev} (20\%). Given the entailment dataset MNLI as the indirect supervision data \( D_{\text{ind}} \) and weakly supervised data \( D_{\text{weak}} \) from GPT-3, we first pretrain a RoBERTa-large (Liu et al., 2019) on \( D_{\text{ind}} \), then finetune on \( D_{\text{weak}} \). In inference, we test the final model on the \text{test} of each task, checking the system’s generalization ability on diverse domains without optimizing on any domain-specific \text{train}.
5 Experiments

5.1 Datasets

We choose datasets that can cover (i) multiple domains, (ii) different sizes of unseen topics, and (iii) various textual forms of topics (phrase-form and sentence-form). Therefore, we evaluate on three mainstream stance detection datasets: SemT6 (Mohammad et al., 2016), VAST (Allaway and McKeown, 2020) and Perspectrum (Chen et al., 2019). We discard their training sets and dev sets to satisfy the definition of OpenStance.

**SemT6 (Mohammad et al., 2016)** contains texts from the tweet domain regarding 6 topics: Donald Trump, Atheism, Feminist Movement, Hillary Clinton, and Legalization of Abortion. It is a three-way stance detection problem with labels {support, oppose, neutral}. Note that the prior applications of SemT6 for zero-shot stance detection always trained on five topics and tested on the remaining one. To match the motivation of OpenStance, we treat the whole SemT6 data as test, i.e., all six topics are unseen. When we report the data-specific supervised performance, we follow prior work to regard any five topics as seen and test on the sixth topic; each topic will have the chance to be unseen, and the average performance is reported.

**VAST (Allaway and McKeown, 2020).** In contrast to SemT6, VAST contains text from the New York Times “Room for Debate” section, and many more topics (4,003 in train, 383 in dev and 600 in test). Those diverse topics, covering various themes, such as education, politics, and public health, are short phrases that are first automatically extracted and then modified by human annotators. Like SemT6, it also has three stance labels, but the neutral topics were randomly picked from the whole topic set. For our OpenStance task, we only use its test to evaluate our system and do not touch the gold labels of its train and dev.

**Perspectrum (Chen et al., 2019)** is a binary stance detection benchmark (label is support or oppose) with two main distinctions with SemT6 and VAST: (i) both its text and topics were collected from several debating websites, and (ii) the topics are sentences rather than noun phrases. Similar to VAST, we do not train our model on its train and dev. The performance on test will be reported. Since there are no neutral samples in this dataset, when the model is pretrained as a 3-way classifier, we set the probability threshold as 1/3 on the oppose label: any prediction that has the oppose probabilities lower than 1/3 will be considered as support. Otherwise, the label would be oppose.

The detailed statistics of the three datasets are listed in Table 1.

| Domain  | #Topics | Topic Form | #Labels |
|---------|---------|------------|---------|
| SemT6   | tweet   | phrase     | 3       |
| VAST    | debate  | 4641/600   | 3       |
| Persp.  | debate  | 541/227    | 2       |

Table 1: Dataset statistics.

5.2 Baselines

There are no prior systems that work on this new OpenStance problem since no training data is available. Here, we consider three baselines that can work on an unsupervised scheme.

**BERT (Devlin et al., 2019).** Given the (text, topic) as input, “BERT-large-uncased” is used as a masked language model to predict the masked token in “text, it [MASK] topic”. BERT will output the probabilities of the three label tokens {support, oppose, neutral} and the label that receives the highest probability would be the predicted stance.

**GPT-3 (Brown et al., 2020).** Given the text and the topic with the instruction telling the model what task we are trying to accomplish, GPT-3 is able to complete the prompt by choosing one of the given labels {support, oppose, neutral}. GPT-3 also has functions designed for classification, but the text completion scheme does a better job on this stance detection task. Our prompt:

Given a topic and a text, determine whether the stance of the text is support, against, or neutral to the topic. Topic: Atheism
Text: Everyone is able to believe in whatever they want. Stance: __________

**Cosine similarity.** We compare the similarities between the text and a hypothesis sentence that combines label and topic, such as “it supports the topic”, “it opposes the
|                                | F1 Score |          |          |
|--------------------------------|----------|----------|----------|
|                                | SemT6    | VAST     | Persp.   |
| random guess                   | 32.0     | 33.3     | 49.8     | 38.3     |
| data-specific supervised learning (prior SOTA) | 38.9 | 78.0 | 91.0 | 69.3 |
| **cross-domain transfer**      |          |          |          |
| SemT6 as *train*               | 38.9     | 28.9     | 47.7     | 38.5     |
| VAST as *train*                | 55.4     | 78.0     | 49.0     | 60.8     |
| Pers as *train*                | 26.7     | 27.0     | 91.0     | 48.2     |
| **open-domain transfer**       |          |          |          |
| baseline                       |          |          |          |
| BERT                           | 22.7     | 36.8     | 36.5     | 32.0     |
| GPT-3                          | 30.5     | 34.2     | 39.9     | 34.9     |
| Cosine                         | 31.5     | 35.9     | 62.7     | 43.4     |
| **ours**                       |          |          |          |
| $D_{\text{ind}}$ and $D_{\text{weak}}$ |          |          |          |
| SemT6-based $D_{\text{weak}}$  | 63.7     | 69.8     | 82.8     | 72.1     |
| VAST-based $D_{\text{weak}}$   | 64.3     | 72.0     | 80.4     | 72.2     |
| Persp-based $D_{\text{weak}}$  | 64.5     | 68.7     | 79.5     | 70.9     |
| joint $D_{\text{weak}}$        | 63.2     | 73.5     | 81.0     | **72.6** |
| w/o indirect                   | 49.6     | 64.6     | 38.2     | 50.8     |
| w/o weak                       | 45.3     | 53.7     | 79.1     | 59.4     |
| w/o MASK-Topic                 | 45.5     | 65.2     | 74.2     | 61.6     |
| w/o MASK-Text                  | 63.4     | 70.8     | 78.2     | 70.8     |

Table 2: Open-domain experiment results on SemT6, VAST and Perspectrum. Our final number is in bold.

topic”, or “it is unrelated to the topic”. We first get the sentential representations by sentence-BERT (Reimers and Gurevych, 2019), then choose the label whose resulting hypothesis obtains the highest cosine similarity score.

In addition to the unsupervised baselines, we further consider the data-specific supervised training as the upperbound, and the following variants of our system: i) only MASK-Text or MASK-Topic; ii) only indirect supervision or weak supervision.

### 5.3 Setting

**GPT-3 for $D_{\text{weak}}$ collection.** The engine we chose for GPT-3 is “curie”, which gives good quality at a reasonable price. There are several parameters that we played with. We set the temperature, which goes from 0 to 1 and controls the randomness of the completion generated, as 0.8 for MASK-Topic and 0.9 for MASK-Text for more diverse results. The randomness for MASK-Text is slightly higher because for some datasets the number of topics is extremely limited, such as SemT6, which only has 6 topics in total; therefore, we want to force diverse responses from GPT-3. The max number of tokens GPT-3 can generate is 6 for MASK-Topic and 150 for MASK-Text. It is worth mentioning that GPT-3 will not necessarily generate as much as the upper bound, sometimes not even close. We let the stop word be "\n", so that it stops generating when it reaches a new paragraph. “top_p” is set as 1, letting all tokens in the vocabulary been used. “frequency_penalty” is 0.3 for MASK-Text to avoid the model producing the same line again and again.

**Training details.** All models are optimized using AdamW (Loshchilov and Hutter, 2019). Learning rate 1e-6, batch size 16, maximal (premise, hypothesis) length is 200. The system is trained for 20 epochs on *train* and the best model on *dev* is kept.

### 5.4 Result

Table 2 lists the main results. We first include “data-specific supervised learning” as the upperbound performance and the “cross-domain transfer” that takes each dataset as the source domain and tests on others respectively. Both settings try to explore the upper limit when we apply human-annotated supervision. Our core task, OpenStance, is evaluated in the last three blocks.

From the baseline block, we can observe that for all domains, baseline methods mostly perform like random guess, except for the slight improvement of the “cosine” approach over Perspectrum. This result indicates the difficulty of
the real-world OpenStance task we proposed. Although BERT and GPT-3 are the top-tier pre-trained language models, they still cannot handle OpenStance well.

Then look at our approach that combines indirect supervision data ($D_{ind}$) and weak supervision data ($D_{weak}$). Note that $D_{weak}$ can be collected based on the label-free train of VAST, SemT6 or Perspectr. We try $D_{weak}$ for each of the task domains and also put them jointly (i.e., “joint $D_{weak}$”). We note that all four versions of $D_{weak}$ result in very consistent performance—mostly around 72% by the “mean”. This clearly supports the robustness of our method: it is less affected by the original domain where text and topic come from, and a single system based on each of the domain or their combination can perform well on all domains.

The last block of Table 2 reports the ablation study, where we discard individual source of supervision (indirect or weak) or individual masking scheme (MASK-Text or MASK-Topic). We observe that i) indirect supervision and weak supervision play complementary roles for the task OpenStance; and they both outperform baselines by large margins, and ii) both masking schemes help, and the MASK-Topic contributes more. This is maybe because MASK-Topic requires the GPT-3 to generate shorter texts than MASK-Text so that MASK-Topic can yield higher-quality data. Additionally, deriving supporting sentences for a given topic sometimes requires substantial background knowledge and solid reasoning, which is still a difficult task for GPT-3.

5.5 Analysis

Next, we conduct a deep analysis for the system robustness towards prompts ($Q_1$), the required size of $D_{weak}$ ($Q_2$), the noise in generated $D_{weak}$ ($Q_3$), and the error patterns made by our system ($Q_4$).

$Q_1$: Robustness of dealing with prompts. Prompt design takes place in both GPT-3 completion and the conversion from stance detection to textual entailment. When generating the prompt for GPT-3, how we construct the prompt in MASK-Topic and MASK-Text can make a huge impact on the completion received. In MASK-Topic, we use the prompt “He said text, so he label the idea of [MASK]”. The reason why we add the idea of” at the end of the prompt is because it helps the model understand that we want a noun phrase. Otherwise, we will see completions like “that”, “it”, etc. Similarly, in MASK-Text, the final prompt we use is “His attitude towards topic is label because he thinks [MASK]”. Considering the freedom of GPT-3 completion, we add “s/he thinks” at the end of the prompt, forcing GPT-3 to generate a reasoning for the given topic/label pair. If we don’t add “he thinks” at the end, it would be common to see GPT-3 repeating the given sentence in the generated completion. In addition, when the label is neutral, such as the prompt “His attitude towards high school writing skills is neutral because he thinks [MASK]”, GPT-3 would output sentences like “he does not have a strong opinion either way” if we don’t have “he thinks” at the end. After the modification, responses would make more sense, such as “that they are important but not essential.” These tricks in prompt design suggest that it is essential to make the sentence structure as clear as possible and provide content that helps to instruct the model on what we want.

When we convert the topic phrase into a sentential hypothesis, we again get involved in the prompt design. During training, we stick with “he is in favor of topic” template to limit the training size, but in the testing, we found the majority voting of four templates (“he/she is in favor of topic” and “he/she opposes topic”) lead to comparable performance with “he is in favor of topic”. This indicates the pre-trained entailment system is considerably robust in dealing with hypotheses derived from different templates.

$Q_2$: How much weakly supervised data is needed? We answer this question by applying $D_{weak}$ alone or together with $D_{ind}$. For each case, we test on sizes varying from 100 to 50,000 and report the average results over 3 random seeds. From the Figure 1, we can see that both settings can reach similar performance when we collect over 10k data
of $D_{\text{weak}}$, but the pretraining on $D_{\text{ind}}$ can dramatically reduce the required size of $D_{\text{weak}}$: from 10k to around 500.

**Q3:** Error patterns of weakly supervised data. We collect typical error patterns in $D_{\text{weak}}$ derived by MASK-Topic and MASK-Text separately.

**MASK-Topic.** Three typical error types.

- **Incomplete generation.** Sometimes GPT-3 fails to give a complete topic phrase and cuts in the middle even though it hasn’t reached the maximum token limit. For example:

  He claims **16 year olds are informed enough to cast a vote, so he supports the idea of GIVING 16-YEAR-OLDS**

  In this example, the topic given by GPT-3 is “giving 16-year-olds”, which is not a complete phrase as we expected. This kind of errors indicate that GPT-3 sometimes stops generating before providing a complete idea even when the word limit is not exceeded.

- **Failure in understanding the stance.** Since we are providing opposite labels (i.e., support and oppose), we hope that GPT-3 would produce distinct topics that hold opposite stances. However, sometimes GPT-3 fails to understand the stances when generating topics. For example:

  He claims **A higher minimum wage means less crime, so he supports the idea of A MINIMUM WAGE**

  He claims **A higher minimum wage means less crime, so he opposes the idea of A MINIMUM WAGE**

  This error type is the most common one in the weakly supervised data (approximately 85% error instances), indicating that GPT-3 is still less effective to interpret negated information.

- **Misunderstanding the text.** The GPT-3 does not always understand the meaning of the sentence correctly. For example:

  He claims **women who are housewives should be paid, so he supports the idea of WOMEN BEING PAID LESS THAN MEN**

  Here, the predicted topic is related but not the main subject of the sentence. Such a mistake is rare but still exists weak supervision.

**MASK-Text.** Even though GPT-3 can mostly provide a sentence that is related to the topic and align with the correct stance, more than 50% of the time the content is very short and less informative compared to the texts from the datasets. For example:

| Topic: keep weight | Text: “All the medical evidence points to the fact that it’s nearly impossible to keep off weight once lost. The body just won’t let you.” | Gold label: neutral | Predicted label: support |
|-------------------|---------------------------------------------------------------------------------------------------------------------------------|---------------------|-------------------------|

| Topic: musician   | Text: Spotify and Pandora pay usage rates that are much lower than the radio, records and legal downloads that they are replacing. Low enough to where many potential new artists won’t be able to even earn a living. There must be some alternative other than artists simply being forced to accept the new streaming model that destroys royalties. For example, who set streaming royalty rates? Can artists unionize and negotiate collectively with the streaming services? If we don’t sort this out, we will lose a new generation of artists – which is bad for everyone. | Gold label: support |
|-------------------|---------------------------------------------------------------------------------------------------------------------------------|---------------------|-------------------------|

6 Conclusion

In this work, we define OpenStance, a more realistic and challenging zero-shot stance detection problem in an open world. Under such a setting, multiple domains and numerous topics can be involved, while no topic-specific annotations are required. To solve this problem, we proposed to combine indirect supervision from textual entailment and weak supervision collected from GPT-3. Our system, without the help of any task-specific supervision, outperforms the supervised method on three benchmark datasets that cover various domains and free-form topics.
Acknowledgment
The authors appreciate the reviewers for their insightful comments and suggestions.

References
Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn A. Walker. 2016. Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016. European Language Resources Association (ELRA).

Emily Allaway and Kathleen R. McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, pages 8913–8931. Association for Computational Linguistics.

Emily Allaway, Malavika Srikanth, and Kathleen R. McKeown. 2021. Adversarial learning for zero-shot stance detection on social media. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, pages 4756–4767. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Si Hao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 542–557. Association for Computational Linguistics.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won’t-they: A very large dataset for stance detection on twitter. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 1715–1724. Association for Computational Linguistics.

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, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers, volume 3944 of Lecture Notes in Computer Science, pages 177–190. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics.

Saman Enayati, Ziyu Yang, Benjamin Lu, and Slobodan Vucetic. 2021. A visualization approach for rapid labeling of clinical notes for smoking status extraction. In Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances, pages 24–30, Online. Association for Computational Linguistics.

Kazi Saïdul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, a meeting of SIGDAT, a Special Interest Group of the ACL, pages 751–762. ACL.

Peter Krejzl, Barbora Hourová, and Josef Steinberger. 2017. Stance detection in online discussions. CoRR, abs/1701.00504.

Dilek Kucuk. 2017. Stance detection in turkish tweets. In Workshops Proceedings and Tutorials of the 28th ACM Conference on Hypertext and Social Media (HT 2017), Prague, Czech Republic, July 4-7, 2017, volume 1914 of CEUR Workshop Proceedings. CEUR-WS.org.

Bangzheng Li, Wenpeng Yin, and Muhao Chen. 2022. Ultra-fine entity typing with indirect supervision from natural language inference. Trans. Assoc. Comput. Linguistics, 10:607–622.

Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2021. Improving stance detection with multi-dataset learning and knowledge distillation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November.
with multi-round new classes: Formulation, dataset
and system. In Proceedings of the 2021 Conference
of the North American Chapter of the Association
for Computational Linguistics: Human Language
Technologies, NAACL-HLT 2021, Online, June 6-11,
2021, pages 1351–1360. Association for Compu-
tational Linguistics.

Wenpeng Yin, Nazneen Fatema Rajani, Dragomir R.
Radev, Richard Socher, and Caiming Xiong. 2020.
Universal natural language processing with limited
annotations: Try few-shot textual entailment as a start.
In Proceedings of the 2020 Conference on Empirical
Methods in Natural Language Processing, EMNLP
2020, Online, November 16-20, 2020, pages 8229–
8239. Association for Computational Linguistics.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-
Woo Lee, and Woo-Myoung Park. 2021. Gpt3mix:
Leveraging large-scale language models for text aug-
mentation. In Findings of the Association for Com-
putational Linguistics: EMNLP 2021, Virtual Event /
Punta Cana, Dominican Republic, 16-20 November,
2021, pages 2225–2239. Association for Compu-
tational Linguistics.