Learning to Classify Events from Human Needs Category Descriptions

Haibo Ding and **Zhe Feng**
Bosch Research and Technology Center, Sunnyvale, CA 94085, USA
{haibo.ding,zhe.feng2}@us.bosch.com

Abstract

We study the problem of learning an event classifier from human needs category descriptions, which is challenging due to: (1) the use of highly abstract concepts in natural language descriptions, (2) the difficulty of choosing key concepts. To tackle these two challenges, we propose LEAPI, a zero-shot learning method that first automatically generate weak labels by instantiating high-level concepts with prototypical instances and then trains a human needs classifier with the weakly labeled data. To filter noisy concepts, we design a reinforced selection algorithm to choose high-quality concepts for instantiation. Experimental results on the human needs categorization task show that our method outperforms baseline methods, producing substantially better precision.

1 Introduction

Training accurate text classifiers often requires a large amount of manually labeled data, which is expensive to collect. In contrast, humans can often perform well on a classification task by only reading category descriptions, which are easy to obtain. It is desirable if computers can automatically learn classifiers from class descriptions. In this work, we aim to learn an event classifier automatically from unlabeled events by using human needs category descriptions as supervision. Human needs categories have been proposed to explain why an event is positive or negative (Ding and Riloff, 2018a; Li and Hovy, 2017). For example, event "*I had cancer*" is negative because it violates *Health needs*, while "*I had steak*" is usually positive because it matches *Physiological needs*. Human needs categorization of events (Ding and Riloff, 2018a) is a task to classify events into eight categories associated with human needs (Maslow et al., 1970) in psychology: *Physiological*, *Health*, *Leisure*, *Social*, *Finance*, *Cognition*, *Emotion*, and *None*.

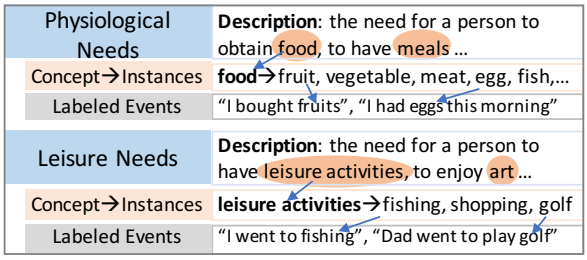

Figure 1: Examples of human need descriptions, selected key concepts, and labeled events with prototypical instances of key concepts.

However, learning a classifier from human needs category descriptions directly is challenging. First, human needs category descriptions often consist of highly abstract concepts. As shown in Fig. 1, the Physiological and Leisure needs are defined using abstract concepts (e.g., "*food*", "*leisure activities*") to cover all instances of them. As demonstrated in our experiments, it is not easy to represent the meanings of these abstract concepts accurately using existing methods. Second, it is not clear how to automatically choose key concepts without accessing manual labels.

In this work, we tackle these two challenges, and propose LEAPI, a method to automatically **Lea**rn a classifier from human need descriptions with **P**rototypical **I**nstantiation. As shown in Fig.1, we first generate candidate key concepts from human needs descriptions (e.g., "*food*"). Then we automatically assign human needs category labels to events that contain prototypical instances of key concepts with the hypothesis that prototypical instances are accurate representations of abstract concepts. For example, we may assign "Physiological Needs" class label to event "⟨*I, had, eggs,* ⟩" because "*egg*" is a prototypical instance of the key concept "*food*". Finally, we train a human needs classifier using the weakly labeled data. Since the

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4698–4704

automatically generated concepts are noisy (e.g., "*person*" is a general term, and may not be a good key concept for recognizing Physiological Needs), we propose a reinforced concept selection algorithm to automatically choose high-quality concepts for instantiation. Experimental results show that our method outperforms baselines, producing substantially better precision.

2 Related Work

There is a growing interest in studying affective events. Some of the previous work (Goyal et al., 2013; Deng and Wiebe, 2014; Ding and Riloff, 2016; Reed et al., 2017; Ding and Riloff, 2018b) aim to recognize the affective polarity of events. Recently, there have been many research work focusing on studying human needs and motives (Paul and Frank, 2019; Rashkin et al., 2018; Ding and Riloff, 2018a; Ding et al., 2019; Otani and Hovy, 2019) to achieve a deeper understanding of sentiment and emotion. However, all these work focused on building classifiers using manually labeled data, or using manual mapping rules (Ding et al., 2018) from existing lexicons such as LIWC (Pennebaker et al., 2007), which requires a significant amount of manual effort.

Our work is related to zero-shot learning for text classification (Yin et al., 2019). As Yin et al. (2019) pointed out there are two different settings of zero-shot learning: (1) *label-partially-unseen*: in which part of the labels are still available for training, and many methods (Zhang et al., 2019; Xia et al., 2018; Rios and Kavuluru, 2018) have been proposed under this setting; (2) *label-fully-unseen*: in which all labels are unseen, and it is also called dataless classification in previous work (Chang et al., 2008; Song and Roth, 2014). Our work of learning a classifier from human needs category descriptions is similar to the second setting of *label-fully-unseen*.

Researchers have also proposed methods (Srivastava et al., 2017; Hancock et al., 2018) to learn classifiers from natural language explanations. These methods require both crowdsourced labels and corresponding explanations of the labels, which are not directly applicable to our problem. One key difference between their work and ours is that their methods convert explanations to logical forms or labeling rules as supervision literally, while our work aims to learn classifiers from conceptual descriptions by considering the hyponyms of abstract concepts. For example, in our work we need to understand that the concept "*food*" in Fig.1 means all instances of food, not just the word "*food*".

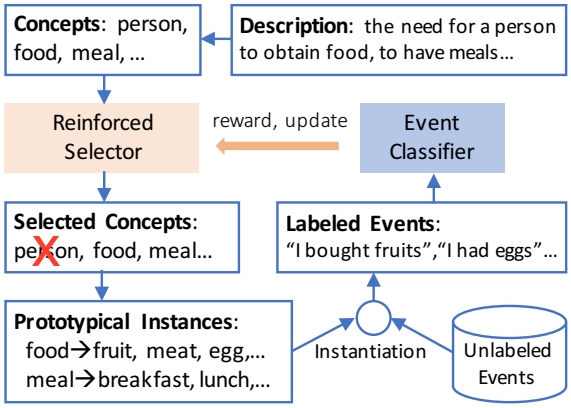

Figure 2: Flow of our method LEAPI

Our work is also related to reinforcement learning which has been used in many NLP applications such as relation classification (Feng et al., 2018; Qin et al., 2018), and sentiment analysis (Wang et al., 2019).

3 Learning Classifiers from Descriptions

Our goal is to design an automatic method to learn a classifier from human needs descriptions and unlabeled events. The key idea is to generate weak labels by instantiating abstract concepts with prototypical instances. Fig.2 shows the basic flow of our method. First, we generate candidate concepts from category descriptions and collect prototypical instances for each concept. Then, we automatically assign human needs labels to unlabeled events that contain prototypical instances of concepts corresponding to the labels. Finally, we train a classifier using the weakly labeled events. To filter noisy concepts, we also design a reinforced selection method to choose high-quality concepts for instantiation.

3.1 Concept Extraction

We hypothesize that key concepts mentioned in the human needs category descriptions can be used to categorize events for their implied human needs. In our work, we use the human needs categories proposed by Ding et al. (2018), which are motivated by Maslow's Hierarchy of Needs (Maslow et al., 1970) and Fundamental Human Needs (Max-Neef et al., 1991) in psychology. We use the manual annotation guidelines described in (Ding et al., 2018) as our human needs category descriptions. Since the original guidelines are short and brief, we rewrote them into self-contained sentences. We include the

descriptions in the Appendix.

We notice that the subject and object in a description sentence often are key concepts. Therefore, we extract subjects and objects from each description sentence as key concept candidates for each category by using *nsubj* and *obj* dependency relations generated by Stanford CoreNLP tool (Manning et al., 2014). For each pair of concepts `a1` and `a2` corresponding to the subject and the object, we construct 3 concept rules: `Has(a1)`, `Has(a2)`, and `Has(a1∧a2)`. Each rule will assign its class label to an event if the event matches the rule.

3.2 Prototypical Instantiation

Based on the intuition that the meaning of an abstract concept can be represented with its prototypical instances. For each concept, we collect 20 most frequent instances of the concept from Probase (Wu et al., 2012) as its prototypical instances.[1] If a concept is not in Probase, we use the most similar concept from Probase based on cosine similarity computed using word embeddings. Then we automatically assign human needs labels to unlabeled events using the constructed concept rules. In our work, we use the events previously extracted by Ding and Riloff (2018b) as our unlabeled events, in which each event is represented as a 4-field tuple, i.e., ⟨Agent, Predicate, Object, PrepositionPhrase⟩. An event matches a concept rule if its fields are prototypical instances of the concepts in the rule. If an event matches a rule, it will receive a human need label associated with the rule. If an event receives different labels, its final label is the majority vote. These weakly labeled events are used to train the final event classifier.

3.3 Human Needs Classifier of Events

Though the weakly labeled events can be accurate, its coverage may not be high. Therefore, we train a simple logistic regression on the weakly labeled events obtained in the last step using event embedding as features. Same as (Ding and Riloff, 2018a), the embedding of an event is computed as the average of embeddings of words in the event.

3.4 Reinforced Concept Selection

We notice that the automatically generated concepts are noisy. As shown in Fig.1 "*person*" extracted from the definitions of Physiological Needs is a general term, weak labels generated using the rule based on this concept can be very noisy for training the final classifier. Therefore, we propose a reinforced concept selection method to select high-quality concepts for instantiation. Since concepts are used via concept rules, we perform the selection among the concept rules. Specifically, we formulate the concept rule selection as follows: given a set of concept rules c_i and its corresponding human need label l_i pairs, i.e., $\mathcal{C} = \{(c_1, l_1), (c_2, l_2), ..., (c_n, l_n)\}$, our goal is to select a subset $\hat{\mathcal{C}}$ of high-quality concept rules.

State. We use s_i to denote the state of each (concept rule, label) pair (c_i, l_i), and represent it with a dense embedding, which is computed as the element-wise product of concept rule and label embeddings. The embedding of a concept rule is the average of word embeddings of all concept words in the rule. Label embeddings are just the embeddings of label names.

Policy Network. We use two layer neural network as our policy function, which is defined as:

$$\pi_\theta(a_i|s_i) = a_i\sigma(f_\theta(s_i)) + (1 - a_i)(1 - \sigma(f_\theta(s_i)))$$

where $f_\theta = W_2\text{ReLU}(W_1 s_i + b_1) + b_2$, the action a_i indicates whether a concept rule is selected ($a_i = 1$) or not ($a_i = 0$), σ is the sigmoid function, and the parameters are $\theta = \{W_1, b_1, W_2, b_2\}$.

Algorithm 1: Reinforcement Learning Algorithm for Concept Rule Selection

Input: concept rule and label pairs $\mathcal{C}$, max episode M, sampling times T, and learning rate α, and parameters of policy network θ
initialize parameter θ;
for *epoch m=0 to M* **do**
  for *sampling time t=0 to T* **do**
    sample selection action for each pair in $\mathcal{C}$;
    estimate reward r_t with selected concepts;
  end
  estimate the baseline $b = \frac{1}{T}\sum r_t$;
  adjust reward $\hat{r_t} = r_t - b$;
  update $\theta \leftarrow \theta + \alpha \sum_t \sum_i \hat{r}_t \nabla_\theta \log \pi_\theta(s_i|a_i)$
end

Policy Optimization. We formulate the concept rule selection as a policy optimization problem in which we aim to find a policy that can select a subset of rules with maximum reward $U(\theta)$, where

$$U(\theta) = E_{a_1,...,a_n} r(a_1, ..., a_n|s_1, ..., s_n) - b$$

We define the reward r to be the macro F1 score of event classification on the development dataset. For

[1] We also collect 200 most confident sentiment words from the SemEval-2015 English Twitter Lexicon (Kiritchenko et al., 2014) as prototypical instances of "sentiment words" concept.

each trajectory, we only receive one reward when selection for all concepts are finished. To reduce the variance, we adjust rewards with a baseline b, which is computed as the average of rewards of sampled trajectories. In our experiments, we use the REINFORCE algorithm (Williams, 1992) to optimize our policy network. The detailed concept rule selection algorithm is shown in Algorithm 1.

4 Evaluation

4.1 Experimental Setting

Our experimental setup is same with the *label-fully-unseen* type of zero-shot learning (Yin et al., 2019), in which all labels are unseen. In our experiments, we used the 542 events with officially annotated human needs labels by Ding et al. (2018) as our test set, and used another distinct set of 300 events labeled in preliminary studies as our development set for hyperparameter tuning. We also used 30K[2] unlabeled events[3] as our unlabeled data.

We compared our method with the following methods.

Majority: We used the majority label of human needs classes as the predictions for testing events.

ESA: We implemented ESA (Gabrilovich and Markovitch, 2007) using the 2019/01/20 Wikipedia dump. To predict an event's label, we first map both events and human needs category descriptions into sparse vectors represented using Wikipedia page titles. Then, for each event, we compute its cosine similarity with each category and predict its label as the most similar one.

Word2Vec: We computed the embeddings of events and category descriptions as the average of embeddings of words in them. Then, we predicted an event's label as the most similar category based on its cosine similarities with all categories.

BERT: We used the pre-trained BERT model (bert-base-uncased) (Devlin et al., 2018) to compute the embeddings of words in events and descriptions, then used the average to compute final embeddings. Same as Word2Vec, we used cosine similarity to predict the labels of events.

Entail: We also experimented with three pre-trained entailment models that are trained on: MNLI (Williams et al., 2018), GLUE RTE (Wang et al., 2018), FEVER (Thorne et al., 2018), and their ensemble model proposed in (Yin et al., 2019). We first manually converted both human needs names and descriptions into hypotheses according to Yin et al. (2019), then we used pre-trained entailment models to predict if an event *entails* or *not-entails* any of the hypotheses. If it entails, we assign the corresponding label to the event.

Implementation Details For *ESA*, *Word2Vec*, and *BERT*, we used cosine similarity for prediction. Since the *None* category is defined to categorize events that do not belong to other classes, its category description does not contain key concepts that can be used to identify events for this class. We predicted an event as *None* if its similarities with other categories are $< \tau$, which was selected on the dev set.

For our human needs classifier, we used the LR classifier in scikit-learn (Pedregosa et al., 2011) with default parameters. Since the *None* category description is not meaningful, we randomly selected K events from unlabeled data as training samples for this class. Our reinforced policy network has around 10k parameters, with a hidden layer size of 32. We used Word2Vec (Mikolov et al., 2013) as our word embeddings. In our experiments, the maximum epoch number is M=200, and we manually searched for the hyperparameters on the development set from the following ranges: learning rate $\alpha \in \{$1e-2, 1e-3, 1e-4$\}$, number of *None* class events $K \in \{100, 300, 500\}$, sampling times $T \in \{10, 30, 50\}$. The best hyperparameters are α=1e-3, K=300, and T=30.

4.2 Experimental Results

Table 1 shows the performance of our method LEAPI and baseline methods that directly used human needs category descriptions for prediction. Results show that Word2Vec performed best among baselines. Without reinforced concept selection (RCS), our method achieved similar F1 score of 31.3 on the dev set as Word2Vec, and F1 score of 35.6 on the test. With RCS, our method averagely selected 56 from the 89 candidate concept rules, and obtained significantly better results than Word2Vec, yielding F1 gains of over +15% on both dev and test sets. Our method also significantly improved the precision from 32.7→55.8 on dev, and from 33.3→51.6 on the test.

Detailed Comparison and Analysis Table 2 shows the performance of the best baseline

[2]We also experimented with more unlabeled events, but found that it did not improve the performance.

[3]Download from http://www.cs.utah.edu/~tianyu/affEvent/affEventKB/

Method	Dev P	Dev R	Dev F1	Test P	Test R	Test F1
Majority	3.1	12.5	5.0	3.0	12.5	4.8
ESA	27.0	21.4	23.9	22.7	22.0	22.3
BERT	37.1	28.4	32.2	27.8	23.9	25.7
Entail	25.7	25.1	25.4	26.4	28.0	27.2
Word2Vec	32.7	30.9	31.8	33.3	28.9	30.9
LEAPI						
–RCS	45.0	24.0	31.3	**52.7**	26.9	35.6
+RCS	**55.8**	**42.9**	**48.6**±1.5	51.6	**42.4**	**46.5**±2.4

Table 1: Macro-averaged results using human need descriptions. +/-RCS indicates if our method uses or not reinforced concept selection. Results of +RCS are the means (also stds for F1) across 10 random seeds.

Word2Vec and our method on the test across human needs categories. Compared with Word2Vec, our method performed better on every category, and obtained large F1 gains of +19.3 on *Health*, +26.4 on *Leisure*, +30.8 on *Social*, and +38.3 on *Emotion* class. We also notice that the performance varies greatly between different categories. Our method obtained relatively small improvement and achieved F1 scores of 33.4 and 28.3 on *Finance* and *Cognition* classes respectively. We examined the predictions on these two categories, and found that the semantic meanings of many events in these two categories are often expressed by event predicates (e.g., "*forgot*" in "*I forgot him*", and "*resign*" in "*I want to resign*"). But, our method only focused on noun concepts, and the concepts of event predicates were not used, which can be improved in future work by extracting and instantiating concepts for event predicates.

Category	Word2Vec P	Word2Vec R	Word2Vec F1	LEAPI P	LEAPI R	LEAPI F1
Physiol.	23.4	57.9	33.3	67.2	38.9	**48.6**
Health	19.2	44.2	26.7	82.6	31.9	**46.0**
Leisure	21.5	42.7	28.6	64.5	48.3	**55.0**
Social	36.7	20.4	26.2	44.9	78.3	**57.0**
Finance	38.1	27.6	32.0	36.9	31.0	**33.4**
Cognition	60.0	11.5	19.4	23.9	36.2	**28.3**
Emotion	50.0	06.2	11.1	63.3	40.5	**49.4**
None	17.7	21.0	19.2	29.4	33.6	**31.3**
MacroAvg	33.3	28.9	30.9	51.6	42.4	**46.5**

Table 2: Results across human needs categories. Our results are the means across 10 random seeds.

Comparison with Manually Selected Concepts To investigate the quality of automatically selected concepts, we also evaluated our method and baselines using the concepts (total 29) that were manually generated by authors[4]. Results are shown in Table 3. The automatic concepts were selected using our method. We find that our method LEAPI achieved much better performance than baselines using manual concepts. It further improved the F1 from 46.5→51.3 compared to that using automatic concepts.

Compared to the results using category descriptions directly (Table 1), both ESA and Word2Vec achieved better performance using both automatically and manually selected concepts, demonstrating the importance of concept selection. We also notice that, with manual concepts, ESA and Word2Vec did not perform better compared to that using automatic concepts. One possible reason is that they can not accurately represent the meanings of the selected concepts, which is the motivation of our work to instantiate abstract concepts with prototypical instances.

Method	Automatic Concepts P	Automatic Concepts R	Automatic Concepts F1	Manual Concepts P	Manual Concepts R	Manual Concepts F1
BERT	20.6	21.4	20.7±5.1	41.8	23.4	30.0
ESA	46.9	24.2	31.8±3.0	42.6	23.9	30.6
Word2Vec	45.1	33.0	37.7±1.3	34.9	40.0	37.3
LEAPI	51.6	**42.4**	**46.5**±2.4	**58.5**	**45.6**	**51.3**

Table 3: Results on test set using automatically and manually selected concepts. Results of automatic concepts are the means and standard deviations across 10 random seeds.

5 Conclusion

In this work, we proposed a zero-shot learning method to learn a classifier from human needs category descriptions by instantiating abstract concepts with prototypical instances. We also proposed a reinforced concept selection method to select high-quality concepts for instantiation automatically. Our experimental results demonstrate that our method achieved significantly better performance than baselines. In our work, we also noticed that the semantics of some events are composed of several concepts. Therefore, in the future, it would be worthwhile to explore the compositional concepts from category descriptions further to improve the performance of human needs categorization of events.

[4]The candidate concepts, manually selected concepts, and the selected concepts by RCS are in the Appendix.

References

Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *Proceedings of the 23rd National Conference on Artificial Intelligence*.

Lingjia Deng and Janyce Wiebe. 2014. Sentiment propagation via implicature constraints. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Haibo Ding, Tianyu Jiang, and Ellen Riloff. 2018. Why is an event affective? classifying affective events based on human needs. In *AAAI-18 Workshop on Affective Content Analysis*.

Haibo Ding and Ellen Riloff. 2016. Acquiring knowledge of affective events from blogs using label propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Haibo Ding and Ellen Riloff. 2018a. Human needs categorization of affective events using labeled and unlabeled data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2018)*.

Haibo Ding and Ellen Riloff. 2018b. Weakly supervised induction of affective events by optimizing semantic consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Haibo Ding, Ellen Riloff, and Zhe Feng. 2019. Improving human needs categorization of events with semantic classification. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*.

Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*.

A. Goyal, E. Riloff, and H. Daumé III. 2013. A Computational Model for Plot Units. *Computational Intelligence*, 29(3):466–488.

Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Jiwei Li and Eduard Hovy. 2017. Reflections on sentiment/opinion analysis. In *A practical guide to sentiment analysis*, pages 41–59. Springer.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*.

Abraham Harold Maslow, Robert Frager, James Fadiman, Cynthia McReynolds, and Ruth Cox. 1970. *Motivation and personality*, volume 2. Harper & Row New York.

Manfred Max-Neef, Antonio Elizalde, and Martin Hopenhayn. 1991. *Human Scale Development: Conception, Application and Further Reflections*. The Apex Press.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Workshop at ICLR*.

Naoki Otani and Eduard Hovy. 2019. Toward comprehensive understanding of a sentiment based on human motives. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Debjit Paul and Anette Frank. 2019. Ranking and selecting multi-hop knowledge paths to better predict human needs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. Linguistic inquiry and word count: LIWC2007. *Austin, TX: liwc.net*.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2137–2147.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling

naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Lena Reed, JiaQi Wu, Shereen Oraby, Pranav Anand, and Marilyn A. Walker. 2017. Learning lexico-functional patterns for first-person affect. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Yangqiu Song and Dan Roth. 2014. On dataless hierarchical text classification. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Shashank Srivastava, Igor Labutov, and Tom Mitchell. 2017. Joint concept learning and semantic parsing from natural language explanations. In *Proceedings of the 2017 conference on empirical methods in natural language processing*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of NAACL-HLT*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Jingjing Wang, Changlong Sun, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2019. Aspect sentiment classification towards question-answering with reinforced bidirectional attention network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3548–3557.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 481–492.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and S Yu Philip. 2018. Zero-shot user intent detection via capsule neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3905–3914.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. 2019. Integrating semantic knowledge to tackle zero-shot text classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.