Type-Aware Decomposed Framework for Few-Shot Named Entity Recognition

Yongqi Li (/profile?id=~Yongqi_Li3), Yu Yu (/profile?id=~Yu_Yu4), Tieyun Qian (/profile?id=~Tieyun_Qian1)

Published: 08 Oct 2023, Last Modified: 02 Dec 2023 EMNLP 2023 Findings Everyone Revisions (/revisions?id=Rvz7LvHcdX) BibTeX

Submission Type: Regular Long Paper
Submission Track: Information Extraction
Submission Track 2: NLP Applications
Keywords: Named Entity Recognition, Few-Shot Learning
TL;DR: A decomposed framework for solving few-shot named entity recognition.

Abstract:
Despite the recent success achieved by several two-stage prototypical networks in few-shot named entity recognition (NER) task, the over-detected false spans at span detection stage and the inaccurate and unstable prototypes at type classification stage remain to be challenging problems. In this paper, we propose a novel Type-Aware Decomposed framework, namely TadNER, to solve these problems. We first present a type-aware span filtering strategy to filter out false spans by removing those semantically far away from type names. We then present a type-aware contrastive learning strategy to construct more accurate and stable prototypes by jointly exploiting support samples and type names as references. Extensive experiments on various benchmarks prove that our proposed TadNER framework yields a new state-of-the-art performance.

Supplementary Materials: zip (/attachment?id=Rvz7LvHcdX&name=supplementary_materials)
Online Version: https://arxiv.org/abs/2302.06397 (https://arxiv.org/abs/2302.06397)
Confirmation Of Submission Requirements: Your submission has at most 9 pages (long) or 5 pages (short) of content., The PDF has been run successfully through the ACL Pubcheck tool https://github.com/acl-org/aclpubcheck (https://github.com/acl-org/aclpubcheck), Your submission uses the unchanged Word or LaTeX template for EMNLP 2023., You have read our Multiple Submission Policy and your paper does not violate it, You have read our Code of Ethics and your paper does not violate the policy., All authors of this paper have provided their up-to-date OpenReview profiles.
Submission Number: 1228

Paper Decision

Decision
by Program Chairs (emnlp-2023-pc@googlegroups.com (/profile?id=emnlp-2023-pc@googlegroups.com), juancitomiguelito@gmail.com (/profile?id=juancitomiguelito@gmail.com), juancarabina@meta.com (/profile?id=juancarabina@meta.com), hbouamor@cmu.edu (/profile?id=hbouamor@cmu.edu), +2 more (/group/info?id=EMNLP/2023/Conference/Program_Chairs))
08 Oct 2023, 03:38 (modified: 02 Dec 2023, 05:17) Everyone Revisions (/revisions?id=wXhiNk4bF5)

Decision: Accept-Findings
Comment:
This article primarily enhances few-shot named entity recognition tasks by utilizing label information and span filtering to address two key challenges: the issue of over-detected false spans and the problem of inaccurate and unstable prototypes. The proposed method demonstrates strong performance in the realm of few-shot named entity recognition tasks, as substantiated by a comprehensive array of experiments. After serious discussion and consideration, we appreciate this solid

work but think that the innovation appears somewhat limited. Furthermore, the explanation of filtering false spans is unclear, making it difficult to determine the positive impact of label information on the model and whether it filters out useful knowledge.

Overall, we appreciate both the efforts of reviewers and authors. This work is ready to publish, and the only concern is the novelty. We hope the authors can continue to improve the paper according to the comments and your response.

Meta Review of Submission1228 by Area Chair pbAD

Meta Review by Area Chair pbAD 16 Sept 2023, 04:31 (modified: 02 Dec 2023, 04:21) Everyone

Revisions (/revisions?id=33Y3oFKGdF)

Metareview:

This article primarily enhances few-shot named entity recognition tasks by utilizing label information and span filtering to address two key challenges: the issue of over-detected false spans and the problem of inaccurate and unstable prototypes. The proposed method demonstrates strong performance in the realm of few-shot named entity recognition tasks, as substantiated by a comprehensive array of experiments. After serious discussion and consideration, we appreciate this solid work but think that the innovation appears somewhat limited. Furthermore, the explanation of filtering false spans is unclear, making it difficult to determine the positive impact of label information on the model and whether it filters out useful knowledge.

Overall, we appreciate both the efforts of reviewers and authors. This work is ready to publish, and the only concern is the novelty. We hope the authors can continue to improve the paper according to the comments and your response.

Recommendation: 3: Sound but not Exciting Enough: Accept to Findings

Official Review of Submission1228 by Reviewer WiN1

Official Review by Reviewer WiN1 04 Aug 2023, 11:25 (modified: 02 Dec 2023, 04:04) Everyone, Reviewer WiN1

Revisions (/revisions?id=xlFJULBSFp)

Paper Topic And Main Contributions:

The paper proposes a novel approach called TadNER for few-shot named entity recognition (NER), which aims to overcome two main challenges in this task: over-detected false spans and inaccurate and unstable prototypes. TadNER incorporates a type-aware span filtering strategy and a type-aware contrastive learning strategy.

1)The type-aware span filtering strategy is designed to filter out false spans by identifying and removing those that are semantically far away from type names. This is achieved by considering the distance between each span and the nearest type name and filtering out those that exceed a predefined threshold.

2)The type-aware contrastive learning strategy is designed to construct more accurate and stable prototypes by leveraging type names and support samples as references. This is achieved by jointly optimizing a contrastive loss function that encourages the prototypes to be close to their corresponding support samples and far from other samples in the same type or other types.

3)The proposed TadNER approach is evaluated on various benchmarks, and the results demonstrate that it achieves a new state-of-the-art performance, outperforming existing approaches.

Reasons To Accept:

The paper demonstrates a clear structure and presents promising experimental results.

Reasons To Reject:

1. The overall logic of the paper is relatively clear, but this paper lacks sufficient novelty and appears to be a limited improvement over the DecomposedMetaNER paper.
2. In the introduction of part of the work, the article lacks details and does not elaborate clearly. These issues will be elaborated on in the following question and answer section.

Questions For The Authors:

In the reading of this paper, there are the following questions: 1)It seems that on the basis of the mainstream span method, the paper only adds label names to inject information into the model, but does not carry out greater innovation in the method or model. The results of the experiment appear to be improved, but there is no good formula or theoretical basis for the effectiveness of the information added or removed. 2)The article proposes to delete the type far from the type name. In the example Figure 1 given in the article, there is only one word between 1976 and California. If we want to delete the influence of 1976, the window distance is very small, which seems to be helpful only for this particular case, but for most of the samples, Whether it is appropriate, and whether it will cause some important boundary information to be missed. Because in most cases, we want to be able to detect more information from a long text. 3)The paper wants to use type names to obtain richer semantics and prototype representation, but in contrast, the mainstream large models such as GPT contain richer semantics. What is the semantic advantage of the proposed type names compared with the big model? Besides, why does the paper not use the big model to get the semantics? 4)The article's description of the type of deletion is too simple, and there is no detailed description of the basis and distance of deletion, so there is no way to judge the

effectiveness of this step. The article seems to give the threshold value in Formula 11 but does not explain the value in the formula, indicating that the description is very vague. 5)Regarding the fine-tuning proposed in the article, may I ask how to divide the support data of the test set during the fine-tuning process? In the case of 1shot, it cannot be divided and fine-tuned the support, which will lead to type loss. How the fine-tuning proposed in the article solves this problem is not explained in the article. 6)The article argues that adding type names can be helpful for prototyping. In the result and analysis part of the article, only the influence on acc after adding the name is given, and the effect of the prototype is not intuitively seen. It seems that the article should give a vector diagram of the prototype to illustrate the correction or aggregation effect of the prototype. 7) Is the model proposed in this paper an experiment based on the DecomposedMetaNER model? If so, is the classification of the training set, verification set, and test set the same as that of DecomposedMetaNER? Can you show more details of the processing

Soundness: 3: Good: This study provides sufficient support for its major claims/arguments, some minor points may need extra support or details.

Excitement: 3: Ambivalent: It has merits (e.g., it reports state-of-the-art results, the idea is nice), but there are key weaknesses (e.g., it describes incremental work), and it can significantly benefit from another round of revision. However, I won't object to accepting it if my co-reviewers champion it.

Reproducibility: 2: Would be hard pressed to reproduce the results. The contribution depends on data that are simply not available outside the author's institution or consortium; not enough details are provided.

Ethical Concerns: No

Reviewer Confidence: 5: Positive that my evaluation is correct. I read the paper very carefully and I am very familiar with related work.

Rebuttal by Authors

Rebuttal

by Authors (Yongqi Li (/profile?id=~Yongqi_Li3), Yu Yu (/profile?id=~Yu_Yu4), Tieyun Qian (/profile?id=~Tieyun_Qian1))

28 Aug 2023, 23:33 (modified: 09 Dec 2023, 01:07) Everyone

Revisions (/revisions?id=wcPowVggmW)

Rebuttal:

We thank the reviewer for the constructive comments. We hope the following clarifications can address the reviewer's concerns.

> Q1. It seems that on the basis of the mainstream span method, the paper only adds label names to inject information into the model, but does not carry out greater innovation in the method or model.
> The results of the experiment appear to be improved, but there is no good formula or theoretical basis for the effectiveness of the information added or removed.

1. To the best of our knowledge, no related work has explicitly identified and addressed the issues of "over-detected false spans" and "inaccurate and unstable prototypes" in the context of few-shot NER. In other words, these are new problems for few-shot NER. In order to solve these two issues, we propose to inject type name information and design the type-aware span filtering strategy and the type-aware contrastive learning strategy, respectively.

Although our approach may seem simple and straightforward, this does not necessarily imply a lack of innovation. Firstly, recognizing a new problem is always critical to the research community, and it might be more important than solving the problem itself. Secondly, our recognized two problems may appear in the few-shot scenarios of other NLP tasks, such as the few-shot event detection task (extracting the event trigger word first and then classifying the event type based on the trigger word). Our discoveries and solutions could provide some insights to the researchers when tackling these tasks, too.

2. Early probabilistic models such as Semi-CRFs [1] had good formalized analysis. Recent data driven deep learning methods can achieve better performance yet often lacking the theoretical basis. To compensate for this shortcoming, for the added type name information, we conducted as much experimental analysis as possible. As can be seen, in Section 3.5, there are a large number of experiments with "Synonyms", "Meaningless" and "Misleading" variants of type names to demonstrate the effectiveness of type name injection.

[1] Sarawagi S, Cohen W W. Semi-markov conditional random fields for information extraction. Advances in neural information processing systems, 2004, 17.

> Q2. The article proposes to delete the type far from the type name. In the example Figure 1 given in the article, there is only one word between 1976 and California. If we want to delete the influence of 1976, the window distance is very small, which seems to be helpful only for this particular case, but for most of the samples, Whether it is appropriate, and whether it will cause some important boundary information to be missed. Because in most cases, we want to be able to detect more information from a long text.

We guess that the reviewer might have misunderstood the main reason why "1976" was misclassified as "LOC". The reviewer may think that the misclassification of "1976" into "LOC" was influenced by the semantic of the nearby token "California". However, the error is mainly due to the bias brought by the domain gap in the few-shot scenario.

First of all, we are focusing on the NER task setting with a fixed set of entity types (which is also the most common in its practical applications). For the example in Figure 1 (a) and (b), we only aim to extract entities of the "ORG" and "LOC" type. So even if "1976" is an entity of the "DATE" type, we don't want to extract it.

Secondly, assuming under a specific few-shot NER setting (Figure 2), the entity type sets of the training and test sets are {"PER", "DATE"} and {"ORG", "LOC"}, respectively. The model needs to be trained on the training set first, and then be fine-tuned using few support samples in the test set. Since entities of the "DATE" type have appeared in the training set, the span detection module will be trained to extract the span of the "DATE" type entity. During the inference on the test set, as shown in the Figure 1 (a), "1976" is extracted at the span detection stage. Note that **for this test set, our predefined entity type set does not include the "DATE" type**, i.e., we do not want to retain the entity span of "1976". Continuing to classify it would force it to be assigned an entity type of either "ORG" or "LOC" (as these are the only two types in this test set), resulting in an obviously incorrect classification result. Therefore, the type-aware span filtering strategy is designed to alleviate this problem.

> Q3. The paper wants to use type names to obtain richer semantics and prototype representation, but in contrast, the mainstream large models such as GPT contain richer semantics.
> What is the semantic advantage of the proposed type names compared with the big model? Besides, why does the paper not use the big model to get the semantics?

Firstly, we suppose the reviewer's concern is about "using a larger model to get a richer semantics may perform better than incorporating type names". However, the two main issues we've identified and tried to address, i.e., "over-detected false spans" and "inaccurate and unstable prototypes", are not caused by the lack of rich semantic information in the LM. Instead, they are caused by the bias from the domain gap and the scarcity of samples under few-shot settings, respectively.

For the issue of "over-detected false spans", the main reason is that at the entity span detection stage, the span information learned by the model from the source domain is inevitably transferred to the target domain. When the entity type sets of the source domain and target domain are different, some entity spans that only belong to the source domain type set rather than the target domain type set will be incorrectly extracted (this is also discussed in question 2 above). Therefore, this problem is due to the bias from the domain gap, not a lack of rich semantic information obtained through the LM.

For the issue of "inaccurate and unstable prototypes", the main reason is that the prototype under the few-shot settings is constructed with very few samples (1/5-shot), which may deviate from the actual category center in some scenarios (Figure 1c). This would lead to a reduction in the model's performance (Section 3.6). Hence, even using a larger model to obtain the representation of entity words, there may still exist such a prototype bias caused by the scarcity of samples.

For these two reasons, we think it is unnecessary to conduct experiments with a big model just for obtaining richer semantics. Therefore, we only used the BERT model in our experiments, which is frequently adopted by community researchers, including all the baselines we compared, to validate the effectiveness of our proposed method.

Secondly, we are not sure whether the reviewer's concern is about the necessity of small PLMs like BERT for this problem. If this is the case, the answer is YES because LLMs cannot perform better than fine-tuned SLMs, e.g., 5-shot + settings in our task, due to the lack of source-domain NER data. A newly published paper [1] in ACL 23 also get the similar conclusion with the help of a large corpus in the source domain.

[1] Chen et al., 2023. Learning In-context Learning for Named Entity Recognition. ACL 2023, pages 13661–13675.

> Q4. The article's description of the type of deletion is too simple, and there is no detailed description of the basis and distance of deletion, so there is no way to judge the effectiveness of this step.
> The article seems to give the threshold value in Formula 11 but does not explain the value in the formula, indicating that the description is very vague.

Sorry for the confusion caused by the lack of detailed descriptions, which is mainly due to space constraints that prevent us from elaborating on the corresponding information in the main body.

1. For detailed process of the type-aware span filtering (deletion), please refer to lines 20-31 of the algorithm in Appendix A.1. Per your suggestion, we will add a more detailed description of the type-aware span filtering strategy in the corresponding methodology section.
2. The symbols used in Eq. 11 were mentioned in the previous paragraphs. Specifically, e_i^s refers to the i_{th} entity token in the support set (line 130), and the Map() function is for converting the original class labels into corresponding type names, e.g., "PER"->"person" (page 3, footnote 4). The main purpose of this formula is to obtain a type name-aware threshold using samples in the support set, which would be used for the false span filtering. We will add detailed descriptions of these symbols in Eq. 11 to make it more clear to read.

> Q5. Regarding the fine-tuning proposed in the article, may I ask how to divide the support data of the test set during the fine-tuning process?
> In the case of 1 shot, it cannot be divided and fine-tuned the support, which will lead to type loss. How the fine-tuning proposed in the article solves this problem is not explained in the article.

We suppose the reviewer's concern about "how to divide the support set into a training set and a validation set" during the fine-tuning process.

In fact, for the few-shot (1-shot or 5-shot) NER task, the number of samples available in the divided validation set is too limited to prevent the overfitting problem. Thus, in our experiments, we do not divide the support set into a training set and a validation set, but use it as a whole for fine-tuning the model. To address the overfitting problem, we employ a loss-based early-stopping strategy, which is a commonly used approach in previous methods like CONTaiNER [1] during the fine-tuning step.

In Appendix A.5, we provide specific implementation details. During the fine-tuning process of the span detection module and type classification module, we monitor the loss. If the loss continues to rise for β times (where β is a hyperparameter), we stop the fine-tuning process to mitigate the risk of overfitting. We will clarify this in the methodology section to ensure clarity and avoid any potential confusion.

[1] Das et al., 2022. CONTaiNER: Few-Shot Named Entity Recognition via Contrastive Learning. ACL (1) 2022: 6338-6353.

> Q6. The article argues that adding type names can be helpful for prototyping. In the result and analysis part of the article, only the influence on acc after adding the name is given, and the effect of the prototype is not intuitively seen. It seems that the article should give a vector diagram of the prototype to illustrate the correction or aggregation effect of the prototype.

Thanks for the suggestion. We conducted visualization experiments as soon as we received comments from the reviewer, i.e., we visualized the distribution of prototypes and test samples without and with type names.

The visualization results show that the entity prototypes of "ORG" and "MISC" types are obviously deviated from the corresponding actual sample centers, which is due to the scarcity of samples in the given support set. In contrast, the type-aware entity prototypes of the "ORG" and "MISC" types, which incorporate type name semantics, are able to alleviate the problem of prototype bias caused by the scarcity of samples.

But, unfortunately, due to the format restriction of the rebuttal, we are unable to show the images here to the reviewer. We will add this visualization analysis in the paper to intuitively show the effect of type names on the prototype correction.

> Q7. Is the model proposed in this paper an experiment based on the DecomposedMetaNER model? If so, is the classification of the training set, verification set, and test set the same as that of DecomposedMetaNER? Can you show more details of the processing?

No, the experiment of our proposed model is not based on that of the DecomposedMetaNER [1] model. The only similarity between our approach and DecomposedMetaNER [1] is that it is also a decomposed framework, i.e., entity spans are extracted first and then entity types are categorized, which is common in NER [2,3].

Besides, the division of the training set, validation set, and test set is the same as that of DecomposedMetaNER [1]. Due to the space limit, we can only place a detailed description of the dataset in Appendix A.3. We apologize for the inconvenience.

1. We adopt the Few-NERD dataset made public by Ding et al. (2021) [4], which discloses the division of training/validation/test sets and few-shot sampling results (link: https://ningding97.github.io/fewnerd/ (https://ningding97.github.io/fewnerd/)).
2. For the Domain Transfer dataset, we use the division and sampling data made public by Das et al. (2022) [5] (link: https://github.com/psunlpgroup/CONTaiNER (https://github.com/psunlpgroup/CONTaiNER)).

Our TadNER and all the compared baselines, e.g., DecomposedMetaNER [1], conduct experiments using these two publicly available data divisions and samplings.

[1] Ma et al., 2022. Decomposed Meta-Learning for Few-Shot Named Entity Recognition. ACL 2022, pages 1584–1596.
[2] Shen et al., 2021. Locate and Label: A Two-stage Identifier for Nested Named Entity Recognition. ACL 2021, pages 2782–2794.
[3] Zhang et al., 2022. Exploring Modular Task Decomposition in Cross-domain Named Entity Recognition. SIGIR 2022. Association for Computing Machinery, New York, NY, USA, 301–311.
[4] Ding et al., 2021. Few-NERD: A Few-shot Named Entity Recognition Dataset. ACL-IJCNLP 2021.
[5] Das et al., 2022. CONTaiNER: Few-Shot Named Entity Recognition via Contrastive Learning. ACL (1) 2022: 6338-6353.

Replying to Rebuttal by Authors

Official Comment by Reviewer WiN1

Official Comment by Reviewer WiN1 03 Sept 2023, 12:16

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

I appreciate receiving the author's response, but it seems that the authors haven't addressed my primary concerns and focal points adequately.

Firstly, regarding the issue of "over-detected false spans," I still have some doubts. In Q2, the authors didn't provide a clear explanation. The paper suggests deleting types that are far from type names, but in Q2, it is mentioned that misclassification is due to domain gap bias, not the influence of nearby tokens. I'm not sure if the authors are trying to emphasize the impact of distance or bias.

Secondly, concerning whether using larger models can lead to more accurate results, the authors didn't present compelling experimental evidence to support their claim. This leaves me skeptical of their response.

Thirdly, the paper asserts that when the entity type sets of the source and target domains differ, some entity spans specific to the source domain type set may be incorrectly extracted due to domain gap bias. This seems more like a common issue of model overfitting and poor generalization in few-shot scenarios rather than a domain-specific problem.

Lastly, the author's explanation of the fine-tuning method isn't entirely clear. What does "as a whole" mean? Additionally, the reference to CONTaiNER as the method for fine-tuning seems misleading, as CONTaiNER primarily employs losses related to text vector representations, and early stopping isn't its main fine-tuning method.

Replying to Rebuttal by Authors

Rebuttal Acknowledgement by Reviewer WiN1

Rebuttal Acknowledgement by Reviewer WiN1 03 Sept 2023, 12:25 (modified: 09 Dec 2023, 02:16)

Everyone Revisions (/revisions?id=uAP1EloyzP)

Acknowledgement: I have read the author rebuttal and made any necessary changes to my review.

Replying to Official Comment by Reviewer WiN1

Official Comment by Authors

Official Comment

by Authors (Yongqi Li (/profile?id=~Yongqi_Li3), Yu Yu (/profile?id=~Yu_Yu4), Tieyun Qian (/profile?id=~Tieyun_Qian1))

04 Sept 2023, 10:46 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

We really appreciate the careful and dedicated work of the reviewer. Here we will answer the remaining concerns of the reviewer one by one.

> Firstly, I'm not sure if the authors are trying to emphasize the impact of distance or bias.

In short, for the issue of "over-detected false spans", what we try to emphasize is not only the impact of **bias** which **leads to** this issue, but also the impact of semantic **distance** when **solving** it. In other words, the issue of "over-detected false spans" is caused by domain gap **bias** at the span detection stage, and can be solved by our proposed "type-aware span filtering" strategy using semantic **distance**.

> Secondly, concerning whether using larger models can lead to more accurate results, the authors didn't present compelling experimental evidence to support their claim. This leaves me skeptical of their response.

The first and most important thing is that our proposed TadNER and all the baseline methods compute losses based on text vector representations, which can be easily obtained from BERT-like models. However, the big models like GPT-3 or LLaMA are based on the autoregressive generation paradigm. Such autoregressive models are limited by the single-direction design and the pre-training objective of predicting the next word, making them unsuitable for obtaining text vector representations due to the lack of bi-directional information in the context. Besides, it is hard to conduct experiments with these big models. The reason is that it is not allowed by OpenAI to fine-tune the hidden vectors in the closed-source GPT-3 (we can only access its output text), and it requires excessive computational resources and time to fine-tune the open-source LLaMA (7b-70b).

Nevertheless, to alleviate the reviewer's concern about "whether using larger models can lead to more accurate results", we are conducting experiments based on larger BERT-like models with more parameters, such as RoBERTa-large with 355M parameters (BERT-base has 110M parameters).

However, it will take some time to conduct these experiments. Once they are ready, we will post the experimental results here.

Thanks again for your patience.

> Thirdly, the paper asserts that when the entity type sets of the source and target domains differ, some entity spans specific to the source domain type set may be incorrectly extracted due to domain gap bias. This seems more like a common issue of model overfitting and poor generalization in few-shot scenarios rather than a domain-specific problem.

We agree with the reviewer's opinion of *"This seems more like a common issue of model overfitting and poor generalization in few-shot scenarios rather than a domain-specific problem"*. Indeed, model overfitting or poor generalization in few-shot scenarios is a long-standing problem that has been bothering researchers for many years.

However, "poor generalization" is a very general concept, which has various manifestations, and different manifestations require different analysis and solution ideas.

In the context of few-shot NER, the issue of "over-detected false spans" is one of such manifestations, which is **not just an abstract "generalization problem"**, but is the one that can **be clearly observed and addressed specifically**. More importantly, such issue has not been explicitly pointed out and attempted to be tackled in any previous work. Therefore, we believe that the "over-detected false spans" problem that we have identified and solved is valuable to the community.

> Lastly, the author's explanation of the fine-tuning method isn't entirely clear. What does "as a whole" mean? Additionally, the reference to CONTaiNER as the method for fine-tuning seems misleading, as CONTaiNER primarily employs losses related to text vector representations, and early stopping isn't its main fine-tuning method.

Sorry for the confusion. The term "as a whole" here means "all the samples in the support set are used to calculate the loss function value when fine-tuning".

1. For the span detection, during the fine-tuning process (described in Section 2.2.1), we use the same loss function as the one used for training in the source domain (described in Section 2.1.1).
2. For the type classification, we introduce the loss function used for the fine-tuning process in Section 2.2.2, "Domain Adaption", which calculates the contrastive loss between the samples and type names.

In both fine-tuning processes, even for the 5-way 1-shot setting, there are 5 samples available to calculate the loss. Thus, the problem you mentioned in Q5, "In the case of 1 shot, it cannot be divided and fine-tuned the support", will NOT arise.

In addition, early-stopping, which we described in the rebuttal, is indeed one of the main strategies that CONTaiNER employs when fine-tuning the support set on the test set. See the "Early Stopping" paragraph in Section 3.3 of the CONTainNER paper (https://aclanthology.org/2022.acl-long.439.pdf (https://aclanthology.org/2022.acl-long.439.pdf)).

Replying to Official Comment by Reviewer WiN1

Official Comment by Authors

Official Comment

by Authors (Yongqi Li (/profile?id=~Yongqi_Li3), Yu Yu (/profile?id=~Yu_Yu4), Tieyun Qian (/profile?id=~Tieyun_Qian1))

04 Sept 2023, 23:13 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

We really thank the reviewer for the patience. We present the experimental results below to alleviate the concerns raised by the reviewer in Q3.

As we noted in our previous response:

> "However, the two main issues we've identified and tried to address, i.e., "over-detected false spans" and "inaccurate and unstable prototypes", are not caused by the lack of rich semantic information in the LM. Instead, they are caused by the bias from the domain gap and the scarcity of samples under few-shot settings, respectively.".

To support this claim, we use a larger model with more parameters than BERT (110M parameters), i.e., RoBERTa-large (355M parameters), to obtain the semantics and conduct related experiments.

We introduce the following two variants to conduct ablation experiments:

1. TadNER w/o Span Filtering, which removes the span filtering strategy.
2. TadNER w/o Type Name, which removes the type names when constructing entity prototypes.

The experimental results are shown in Table 1.

Models	**Backbone**	**1-shot**					**5-shot**			
		I2B2	CoNLL	WNUT	GUM	Avg.	I2B2	CoNLL	WNUT	GUM
TadNER	RoBERTa-large	**40.69±8.20**	**76.69±7.08**	**44.73±6.12**	**32.13±15.11**	**48.56**	**39.20±12.42**	**85.67±5.96**	**49.00±7.93**	**41.61±15.87**
w/o Span Filtering	RoBERTa-large	35.49±8.68	63.06±4.94	37.22±3.52	22.79±8.98	39.64	29.61±11.04	74.84±3.80	37.81±4.06	29.90±12.34

Models	Backbone	1-shot					5-shot			
w/o Type Name	RoBERTa-large	11.98±4.35	64.19±10.63	30.19±3.53	24.31±5.35	32.67	14.19±5.06	75.16±2.56	38.56±3.32	35.50±2.07

Table 1: Experimental results under the 1-shot and 5-shot Domain Transfer settings using the RoBERTa-large. The mean Micro F1 scores and standard deviations reported in the table are obtained using the 10 sampled support sets, which are the same as those used for the experiments in the paper.

From Table 1, we can observe and conclude that:

1. The removal of the type-aware span filtering strategy leads to a drop in performance. This suggests that with the larger RoBERTa-large model, the issue of "over-detected false spans" caused by domain gap bias at the span detection stage still exists. Our proposed span filtering strategy helps to alleviate this problem well.
2. Removing type names leads to significant performance degradation. This suggests that with the larger RoBERTa-large model, the issue of "inaccurate and unstable prototypes" caused by the scarcity of samples still exists. Incorporating type names when constructing entity prototypes mitigates this problem well.

Overall, the experimental results show that the two issues we are focusing on cannot be solved by a larger model with richer semantics. In other words, these two issues are not caused by the lack of semantic richness in LM, but are more likely due to the two biases that we mentioned above, i.e. the bias from the domain gap and the scarcity of samples under few-shot settings. Therefore, for our experiments in the paper, we only use the BERT model, which is adopted by all the baselines we compared, to fairly verify the validity of our proposed method.

Official Review of Submission1228 by Reviewer RZMJ

Official Review by Reviewer RZMJ 04 Aug 2023, 00:17 (modified: 02 Dec 2023, 04:04) Everyone, Reviewer RZMJ

Revisions (/revisions?id=OEQKIE9Nax)

Paper Topic And Main Contributions:
This paper describes TadNER, a type-aware decomposed framework for named-entity recognition (NER). Its main goal is the enhancement of performance in few-shot NER, i.e. the extension of the set of spans and labels using only few examples. The authors deal with the problem by two points of view: (i) filter out spans semantically far from the accepted type names; (ii) build stable prototypes of labels, to avoid deviation form the class centers.

Reasons To Accept:
The goal of the paper is clear, and the approach seems promising, although it is an application of ideas that already exist for the same kind of task. The use of a big number of baselines and the strong evaluation proposed bring robustness to the whole process. The most interesting part of the study, in my opinion, is the semantic connection between the transformers and the labels (as described in Appendix A.8).

Reasons To Reject:
In some parts, the paper is really hard to follow. In method section, the only real examples are presented in the (small) picture, and it is hard to connect each formula and description to the phases in the image. In general, the LaTeX formatting should be more tidy (see for example Section 2.1.1).

Typos Grammar Style And Presentation Improvements:
There are a lot of errors due to the wrong use of equation in LaTeX (main of them can be found in Section 2.1.1).

Soundness: 4: Strong: This study provides sufficient support for all of its claims/arguments.
Excitement: 3: Ambivalent: It has merits (e.g., it reports state-of-the-art results, the idea is nice), but there are key weaknesses (e.g., it describes incremental work), and it can significantly benefit from another round of revision. However, I won't object to accepting it if my co-reviewers champion it.
Reproducibility: 4: Could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.
Ethical Concerns: No
Reviewer Confidence: 3: Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math, experimental design, or novelty.

Rebuttal by Authors

Rebuttal
by Authors (Yongqi Li (/profile?id=~Yongqi_Li3), Yu Yu (/profile?id=~Yu_Yu4), Tieyun Qian (/profile?id=~Tieyun_Qian1))

28 Aug 2023, 23:33 (modified: 09 Dec 2023, 01:07) Everyone Revisions (/revisions?id=Ze4Kcr0fXs)

Rebuttal:
We thank the reviewer for the constructive comments. We hope the following clarifications can address the reviewer's concerns.

> Q1. In some parts, the paper is really hard to follow. In method section, the only real examples are presented in the (small) picture, and it is hard to connect each formula and description to the phases in the image. In general, the LaTeX formatting should be more tidy (see for example Section 2.1.1).

Sorry for the inconvenience, we will enrich the caption in Figure 2 so that the readers can connect the formula with the real example in the picture more clearly.

> Typos: There are a lot of errors due to the wrong use of equation in LaTeX (main of them can be found in Section 2.1.1).

Thanks for your suggestion! We will fix them per your suggestion.

Replying to Rebuttal by Authors

Rebuttal Acknowledgement by Reviewer RZMJ

Rebuttal Acknowledgement by Reviewer RZMJ 05 Sept 2023, 17:21 (modified: 09 Dec 2023, 02:16)
Everyone Revisions (/revisions?id=PTYw1ZMg2c)

Acknowledgement: I have read the author rebuttal and made any necessary changes to my review.

Official Review of Submission1228 by Reviewer NUFY

Official Review by Reviewer NUFY 31 Jul 2023, 08:07 (modified: 02 Dec 2023, 04:04) Everyone, Reviewer NUFY
Revisions (/revisions?id=5OSiM9ZZ5o)

Paper Topic And Main Contributions:
The authors propose a method for few-shot NER build on top of two-stage methods where first span detection is performed followed by span classification. However, they improve over recent methods by incorporating: (1) an effective way of exploiting the text representations of the labels by embedding them into a common space with the entity tokens, and (2) using these embeddings to filter out false entity span candidates that are too dissimilar to the provided labels. Primarily, they construct entity prototype representations by embedding the entity tokens and the mapped class labels with the same encoder and concatenate them in two variants (i.e. in both directions: entity-label and label-entity). Then, the encoder is trained with a contrastive learning objective that pulls these representations together if they have the same label and pushes them away from each other, if this is not the case.

The authors show that this approach is very effective by conducting few-shot experiments on several datasets (Few-NERD and others) where they outperform previous work (one-stage and two-stage) on a large margin, especially in the realm of very few (1 ~ 2) training examples. Furthermore, they include an extensive ablation study (effect of: span filtering, incorporating the label text, fine-tuning of the encoders for each stage on the support set) and a quantitative as well as a qualitative error analysis.

Reasons To Accept:
The authors propose an effective method for few-shot NER that significantly outperforms previous approaches, especially for the very few shot setting. They explain their method quite well and conduct reasonable experiments to show the effectiveness of their method.

Reasons To Reject:
- see Q1 (re-reported scores seem to be lower than in the cited paper)
- see Q2 (definition of fθ2 is unclear)

Questions For The Authors:
- Q1: In Table 2, where did you get the scores marked as "from Ma et. al (2022c)" from? Maybe I'm mistaken, but I can not find these values in the publication: https://aclanthology.org/2022.findings-acl.124/ (https://aclanthology.org/2022.findings-acl.124/)
- Q2: How is fθ2 defined? Is it really the same as fθ1? But it looks like Map(yi) = t'i can be a token sequence (some of the mapped types in the appendix consist of multiple words). But in this case, equation (4) seems to only work, if fθ1 maps to R^r... in section 2.1.2, it is said that individual entity tokens ei are embedded with fθ2. However, in "2.2.2 Type-Aware Span Filtering" it is said that spans are filtered. Does the filtering happen token-wise? Or are all tokens of one span embedded at once and the result is then filtered (same for the "Inference" section, in both sections you talk about spans)? But if fθ2 takes really a token sequence as input, Figure 2 is very misleading because "Barack Obama" is split into two training instances (("Barack", "person") o ("person", "Barack")) and (("Obama", "person" o ("person", "Obama")) where I had just assumed (("Barack Obama", "person") o ("person", "Barack Obama")).
- Q3: Since your method relies on useful label names: What was the effort to create the respective label name mappings? Could you give some points that this is negligible?

Typos Grammar Style And Presentation Improvements:
- typo: line 127 "datset"
- type: line 938 "doamin"

Soundness: 4: Strong: This study provides sufficient support for all of its claims/arguments.
Excitement: 4: Strong: This paper deepens the understanding of some phenomenon or lowers the barriers to an existing research direction.
Reproducibility: 5: Could easily reproduce the results.
Ethical Concerns: No
Reviewer Confidence: 4: Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Rebuttal by Authors

Rebuttal
by Authors (Yongqi Li (/profile?id=~Yongqi_Li3), Yu Yu (/profile?id=~Yu_Yu4), Tieyun Qian (/profile?id=~Tieyun_Qian1))
28 Aug 2023, 23:34 (modified: 09 Dec 2023, 01:07) Everyone Revisions (/revisions?id=h8et9HoVH0)

Rebuttal:
We thank the reviewer for the constructive comments. We hope the following clarifications can address the reviewer's concerns.

> Q1. In Table 2, where did you get the scores marked as "from Ma et. al (2022c)" from? Maybe I'm mistaken, but I can not find these values in the publication: https://aclanthology.org/2022.findings-acl.124/ (https://aclanthology.org/2022.findings-acl.124/)

This problem is caused by the different versions of the Few-NERD dataset [1] (corresponding to arXiv-V5 and arXiv-V6, respectively). On the publicly available leaderboard (link: https://ningding97.github.io/fewnerd/ (https://ningding97.github.io/fewnerd/)) by the Few-NERD authors, the Few-NERD arXiv-V5 data was deleted by the authors and was replaced with the Few-NERD arXiv-V6 version when we conducted our experiements, so we had to use the Few-NERD arXiv-V6 version.

Results in the camera ready version of Ma et. al's (2022c) [2] paper were obtained using the Few-NERD arXiv-V5 data. For the convenience of comparison with their results, Ma et. al (2022c) [2] reproduced the results using the Few-NERD arXiv-V6 data and published the results on Github. The link is: https://github.com/microsoft/vert-papers/tree/master/papers/DecomposedMetaNER#few-nerd-arxiv-v6-version (https://github.com/microsoft/vert-papers/tree/master/papers/DecomposedMetaNER#few-nerd-arxiv-v6-version), which is also the source of results in our paper which are marked as "from Ma et. al (2022c)".

Due to the space limit, we have to place this explanation in footnote 7 of Appendix A.3. To avoid potential confusion for the readers, we will include this explanation to the main body of the paper.

[1] Ding et al., 2021. Few-NERD: A Few-shot Named Entity Recognition Dataset. ACL-IJCNLP 2021.
[2] Ma et al., 2022. Decomposed Meta-Learning for Few-Shot Named Entity Recognition. ACL 2022, pages 1584–1596.

> Q2. How is f_{θ_2} defined? Is it really the same as f_{θ_1}?
> But it looks like Map(y_i) = $t_i^{'}$ can be a token sequence (some of the mapped types in the appendix consist of multiple words).
> But in this case, equation (4) seems to only work, if f_{θ_1} maps to R^r... in section 2.1.2, it is said that individual entity tokens ei are embedded with f_{θ_2}.
> However, in "2.2.2 Type-Aware Span Filtering" it is said that spans are filtered. Does the filtering happen token-wise? Or are all tokens of one span embedded at once and the result is then filtered (same for the "Inference" section, in both sections you talk about spans)?
> But if fθ2 takes really a token sequence as input, Figure 2 is very misleading because "Barack Obama" is split into two training instances (("Barack", "person") o ("person", "Barack")) and ("Obama", "person" o ("person", "Obama")) where I had just assumed (("Barack Obama", "person") o ("person", "Barack Obama")).

Sorry for the confusion due to the unclear formalization of $f_{\theta_1}()$ and $f_{\theta_2}()$ in the paper.
Generally speaking, $f_{\theta_1}()$ and $f_{\theta_2}()$ are the same context encoder with individual parameters. Here, we will formalize $f_{\theta_1}()$ and $f_{\theta_2}()$ in detail and explain the reviewer's questions one by one. For simplicity of description, let's collectively refer to $f_{\theta_1}()$ and $f_{\theta_2}()$ as $f_\theta()$.

For the input X, the output is $V = f_\theta(X)$.

1. If X is a token, then V is the context embedding vector of this token.
2. If X is a word or span composed of N tokens, i.e., the type name "nationality religion" or entity span "Barack Obama", V is the average of the context embedding vectors for these N tokens.
3. If X is a token sequence composed of M tokens $[x_1, \ldots, x_M]$, like the sequence ["Barack", "Obama", "was",..., "1961"] in Figure 2, then V is the sequence of context embeddings $[emb_1, \ldots, emb_M]$=[f_θ(Barack),..., f_θ(1961)], where the dimension of $emb_1, \ldots, emb_M$ is the hidden layer dimension of the encoder $f_\theta()$.

Based on this formalization, below are our explanations for your questions:

1. "Does the filtering happen token-wise? Or are all tokens of one span embedded at once and the result is then filtered."
Explanation: The filtering process is span-based. The entire span is inputted to the encoder $f_{\theta_2}()$ as a whole (corresponding to the second type of input to $f_\theta()$ in the formalization above).

2. "Figure 2 is very misleading because 'Barack Obama' is split into two training instances"
Explanation: The input to $f_{\theta_2}()$ in Figure 2 is a token sequence, so its output is an embedding sequence where each token e_i corresponds to an output embedding $f_{\theta_2}(e_i)$ (corresponding to the third type of input to $f_\theta()$ in the formalization above).

To avoid any potential confusion, we will add this formalization of the encoders $f_{\theta_1}()$ and $f_{\theta_2}()$ into the methodology section.

> Q3. Since your method relies on useful label names: What was the effort to create the respective label name mappings? Could you give some points that this is negligible?

The type name mappings for all the datasets used in our experiments can be seen in Tables 11 and 12 of the Appendix A.8.
As shown in Table 11, for the Few-NERD dataset, we directly use the fine-grained part of the original label as the corresponding type name, such as "art-film"->"film". As shown in Table 12, for the datasets under the Domain Transfer settings, we convert the majority of labels into their most direct natural language form of type names, such as "PER"->"person" and "AGE"->"age".
Therefore, the additional effort required to construct the mappings is almost negligible.

> Typos: line 127 "datset"; line 938 "doamin"

Thanks a lot for your careful reading, we will fix them per your suggestion.

Replying to Rebuttal by Authors

Rebuttal Acknowledgement by Reviewer NUFY

Rebuttal Acknowledgement by Reviewer NUFY 05 Sept 2023, 01:30 (modified: 09 Dec 2023, 02:16)
Everyone Revisions (/revisions?id=HfJD8FtRjj)

Acknowledgement: I have read the author rebuttal and made any necessary changes to my review.

Author License Task by Authors

Author License Task
by Authors (Yongqi Li (/profile?id=~Yongqi_Li3), Yu Yu (/profile?id=~Yu_Yu4), Tieyun Qian (/profile?id=~Tieyun_Qian1))
28 Jun 2023, 15:44 Program Chairs, Authors

Association For Computational Linguistics - Blind Submission License Agreement: On behalf of all authors, we have read and understand the above description for the License Agreement
Intro: On behalf of all authors, we have read and understand the Introduction
Section 1 Grant Of License: On behalf of all authors, we have read and understand Section 1
Section 2 Permission To Publish Peer Reviewers Content: On behalf of all authors, we have read and understand Section 2
Section 3 Attribution And Public Access License: On behalf of all authors, we have read and understand Section 3
Section 4 Effective Date: On behalf of all authors, we have read and understand Section 4
Section 5 Warranty: On behalf of all authors, we have read and understand Section 5
Section 6 Legal Relationship: On behalf of all authors, we have read and understand Section 6
Agreement: On behalf of all authors, I agree