Improving Few-Shot Domain Transfer for Named Entity Disambiguation with Pattern Exploitation

Philip Blair and **Kfir Bar**
BasisTech
1070 Broadway
Somerville, MA, USA
{pblair,kfir}@basistech.com

Abstract

Named entity disambiguation (NED) is a critical subtask of entity linking, which seeks to connect knowledge base entities with textual mentions of those entities. Naturally, the performance of a model depends on the domain it was trained on; thus, reducing the amount of data required to train models is advantageous. In this work, we leverage recent research on pattern exploitation for NED and explore whether it can reduce the amount of data required for domain adaptation by reformulating the disambiguation task as a masked language modeling problem. Using ADAPET (Tam et al., 2021), which implements a new approach for few-shot learning using fine-tuned transformer-based language models, we produce an NED model which yields, without any sacrifice of in-domain accuracy, a 7% improvement in zero-shot cross-domain performance as evaluated on NEDMed, a new NED dataset of mental health news which we release with this work.

1 Introduction

In order to understand a piece of text, it is often valuable to understand entities which are referred to by that text. While named entity recognition (NER) is an important aspect of this problem, a large variety of applications (e.g. financial credit risk monitoring and open source intelligence gathering) need to further connect these entities to known entities in a knowledge base (KB). This task of linking textual mentions of entities to their KB entries is known as entity linking.

Entity linking is typically further decomposed into two subtasks: candidate generation and named entity disambiguation (NED). The former is responsible for discovering a set of *possible* mentions of entities in a given document (for example, producing the candidates for Cambridge, MA, USA and Cambridge, UK from the surface form “Cambridge”). An NED system then takes these lists of candidates and selects which, if any, is the correct referent.

Named entity disambiguation systems achieve this by utilizing a number of pieces of information, such as related entities, the type of each candidate entity (person, location, etc.), and semantic descriptions of each entity (e.g. a snippet from the candidate’s Wikipedia page). As this association is usually statistically learned from some training dataset, the performance of an NED system on a given document depends on how closely that document’s domain is to that of the training data, in terms of vocabulary, syntax, and the types of entities. Because these systems are often specialized in specific domains, it is therefore necessary to curate sufficient amounts of training data for each of these applications, which is often costly.

Our work seeks to reduce the amount of data required via leveraging pretrained language models (LMs). LMs are often well-suited to assisting low-resource task setups (Tam et al., 2021), for modern language models are sufficiently powerful that their predictive distributions can be interpreted as a basic form of “common-sense reasoning".

Our contributions are as follows: (1) we take a state-of-the-art baseline NED system (Yang et al., 2019) and augment it with an additional signal from an LM fine-tuned with the ADAPET (Tam et al., 2021) procedure, which adapts language models to few-shot learning natural language processing problems. We show that this augmented system, called DCA-Prompt, achieves similar performance to the baseline in both the same and closely-related domains, but demonstrably outperforms when adapting to a new, dissimilar domain. (2) Additionally, we are releasing a new named entity linking dataset, called NEDMed, which is based on mental health news data.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6826 - 6839
December 7-11, 2022 ©2022 Association for Computational Linguistics

2 Related Work

Named entity disambiguation has a storied history, stemming from the work in Bunescu and Paşca (2006), which utilized a support vector machine (SVM) (Cortes and Vapnik, 1995) kernel based on a similarity measure between input documents and the Wikipedia articles for each candidate.

Cucerzan (2007) and Kulkarni et al. (2009) were seminal works in incorporating the KB topology into this decision-making process (e.g. looking at links between candidate entities across the document), but the computational cost of these techniques was a major limitation. Numerous additional authors later provided their own approximations for this problem; a recent success in this area is known as *dynamic context augmentation*, or DCA (Yang et al., 2019). This technique opts to sequentially process the mentions in the document, using context related to previous extractions (the linked entity itself along with entities related to that entity) to inform subsequent extractions.

While pretrained language models have a long history (Devlin et al., 2019) and are traditionally fine-tuned using masked-language modeling in order to improve their modeling ability on new domains, *pattern exploitation training* (PET) (Schick and Schütze, 2021) and *a densely supervised approach to pattern exploitation training* (ADAPET) (Tam et al., 2021) are relatively recent applications of this fine-tuning approach. These techniques utilize the linguistic information contained in language models to solve natural language processing tasks by formulating them as cloze-style phrases. Tam et al. (2021) demonstrate its efficacy on a range of SuperGLUE (Wang et al., 2019) tasks, showing state-of-the-art or competitive performance on few-shot textual entailment and the BoolQ (Clark et al., 2019) question-answering dataset. To our knowledge, we are the first to apply pattern exploitation to NED.

The gold standard dataset for NED was defined in Hoffart et al. (2011b). This work extends the CoNLL 2003 shared task's NER dataset (Tjong Kim Sang and De Meulder, 2003) to contain links to entities from the YAGO KB (Hoffart et al., 2011a). This dataset is known in literature as the AIDA CoNLL-YAGO dataset, and is discussed further in Section 4. Additional datasets include AQUAINT (Milne and Witten, 2008), MSNBC (Cucerzan, 2007), ACE2004 (Ratinov et al., 2011), and CWEB (Guo and Barbosa, 2014).

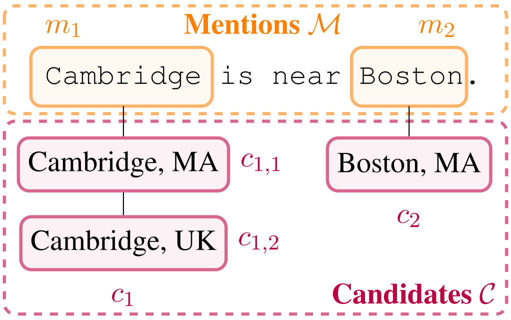


Figure 1: Glossary of terms in the named entity disambiguation task.

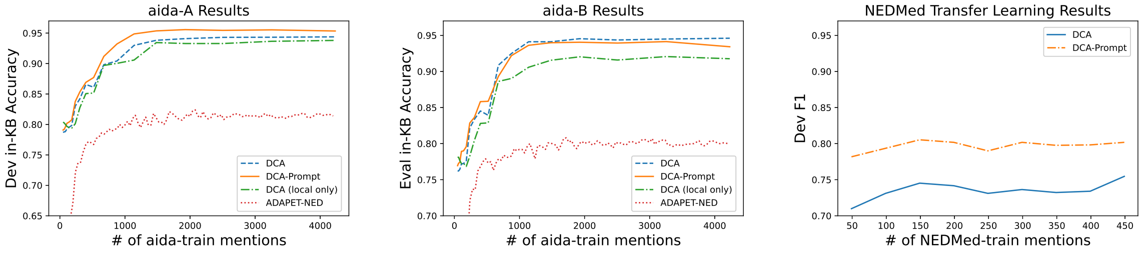
```
[CLS] <LCTX> <MENTIONTXT> <RCTX>.
Description: <ENTDESC>. Is this
      a good description of
   <MENTIONTXT>? [MASK] [SEP]
```

Figure 2: Pattern used during evaluation. Items in square brackets ([]) are special tokens, and items in angle brackets (<>) are substituted with information from the problem input. LCTX and RCTX are the text to the left and right of the entity mention (respectively), MENTIONTXT is the surface form of the mention, and ENTDESC is the description of the candidate under consideration. Colors represent different logical segments of each piece of the input, and [MASK] is what we prompt the language model to substitute into the input.

3 Methodology

In NED (Figure 1), we presume that, for a given input text containing a set of n *mentions* (collectively denoted $\mathcal{M}$), which are the textual surface forms of entities, and a set of *candidates* c_i (collectively, $\mathcal{C}$) for each mention m_i. An NED system ranks these candidates to select the one most likely to be the referent entity. A full formal description of the task is given in Appendix A.

ADAPET solves natural language understanding tasks by "filling-in-the-blank" in natural language patterns. To solve a downstream task, such as classification or question-answering, one formulates the problem instance as some form of prose containing a masked token. A fine-tuned language model is then used to infer which word makes the most sense to substitute this masked token. In order to apply ADAPET to NED, it was necessary to formulate an appropriate pattern. Our system fine-tunes the HuggingFace (Wolf et al., 2020) `bert-large-uncased` model using the pattern shown in Figure 2, treating the task as a binary classification problem (answering, "does this candidate link to this mention?"). The develop-



(a) (b) (c)

Figure 3: Few-shot learning curves for ADAPET-NED, DCA, and DCA-Prompt models. "DCA (local only)" measures performance on the DCA model without entity context enabled, which is more directly comparable to the ADAPET-only line. The x-axis indicates the number of mentions of aida-train used while training.

ment process through which we chose this base model and pattern is described in Appendix B. As done in the codebase provided by Tam et al. (2021), we fine-tune the model to produce the word "true" for correct mention-candidate pairs and the word "false" for incorrect ones. The entity description (`[ENTDESC]`) may be something like a paragraph from an encyclopedia entry or a verbalized form of its knowledge base relationships, as described in Mulang' et al. (2020). This representation should capture enough context to uniquely identify the entity. In our work, we use the first paragraph of the candidate's Wikipedia page.

In order to accurately judge this approach versus state-of-the-art NED systems, we additionally augment an existing state-of-the-art baseline system with an input signal from our ADAPET-tuned LM. The baseline system we selected is based on DCA, and is described by Yang et al. (2019). In short, this model, is a feedforward neural network which accepts features based on the similarity of the mention context and trained entity embeddings (Ψ_C), entity type information (Ψ_T), and coherence between the candidates and both previously linked entities and entities linked to those entities (Φ and Φ'). For further information, readers are referred to Yang et al. (2019). Further technical details regarding our training setup can be found in Appendix C.

Our research focused on the following questions:

RQ1. Does using ADAPET for NED perform similarly to the baseline cross-domain scenarios for similar domains?

RQ2. Does using ADAPET for NED reduce the amount of data required to learn the task?

RQ3. Does using ADAPET for NED improve the ability for a NED system to transfer to another *dissimilar* domain?

4 Results

To answer our first two research questions, we compared ADAPET-NED (ADAPET trained on our NED prompt) against three versions of the baseline DCA system: one which operates as normal (ETZH-Attn+DCA-SL from Yang et al. (2019)), denoted "DCA", one which only has "local" features enabled (i.e. no coherence-based features, which makes it more directly comparable to ADAPET-NED), denoted "DCA (local only)", and a version of the DCA system which has been augmented to receive the output from the ADAPET model's "yes" prediction as an input, denoted "DCA-Prompt". We model the (2-way k-shot) few-shot learning curves of these models in Figures 3(a) and 3(b), evaluating on in-domain datasets (training on different sized subsets of AIDA's aida-train training split and evaluating on aida-A and aida-B development and evaluation splits). As expected, this graph shows that ADAPET-NED greatly underperforms the baseline system.

To explore **RQ1**, we measured cross-domain performance on various publicly available datasets and compare to existing benchmarks. Specifically, we looked at F1 performance on MSNBC, AQUAINT, ACE2004, and CWEB (described in Section 2), as done in Yang et al. (2019). These datasets are all general knowledge corpora, based on either news or encyclopedia pages. We train our models on aida-train and evaluate across all datasets. The results are shown in Table 1. We see that, while ADAPET-NED is not competitive, augmenting DCA with features from ADAPET (DCA-Prompt) yields performance ranging from comparable to superior, with state-of-the-art performance on the ACE2004 dataset. Additionally, for **RQ2**, we find that all three of these models have similar data requirements. We quantify this by using the Kneedle al-

System	MSNBC	AQUAINT	ACE2004	CWEB
ETZH-Attn (Yang et al., 2019)	91.97	84.06	86.92	70.07
ETZH-Attn + DCA-SL (Yang et al., 2019)	**94.57** ± 0.2	87.38 ± 0.5	89.44 ± 0.4	73.47 ± 0.1
ETZH-Attn + DCA-RL (Yang et al., 2019)	93.80 ± 0.0	**88.25 ± 0.4**	90.14 ± 0.0	**75.59 ± 0.3**
ADAPET-NED (ours)	78.70 ± 0.1	78.50 ± 0.1	81.00 ± 0.1	65.60 ± 0.1
DCA-Prompt (ours)	92.37 ± 0.1	87.59 ± 0.4	**91.34** ± 0.0	74.60 ± 0.1

Table 1: Comparison of cross-domain performance of various systems. The best results and our work are in bold. F1 scores are shown, with other systems' scores taken from Yang et al. (2019). Confidence intervals are shown (DCA measured over 5 runs; ours over 3). We observe no degradation in performance over the baseline.

gorithm (Satopaa et al., 2011) to locate the "knees" in each of the curves in Figure 3, which showed diminishing returns at around 1,000 mentions for all three.

In order to assess **RQ3** in a real-world context, we adapted our trained models to a dataset tailored to the medical domain, which is quite different from AIDA's general news domain. We created a dataset, denoted NEDMed[1], containing 110 internet articles on mental health news, which were partitioned into 66 training documents (NEDMed-train) and 44 evaluation documents (NEDMed-dev), containing 2,839 and 1,841 mentions, respectively. Documents were manually annotated for person, location, and organization types, along with a variety of others. For a full list of types and further details on this dataset, see Appendix D. For our experiments, we only utilize entities which have Wikipedia links (4,342 mentions, or roughly 92% of the total 4,680).

Table 2 and Figure 3(c) describe the results of the baseline DCA system, our ADAPET-NED and DCA-Prompt systems on this data. The NEDMed-dev scores on models trained with aida-train (the first group in Table 2) represent zero-shot (cross-domain) scenarios. The models trained on the combined data represent transfer learning scenarios in which we tune an AIDA-trained model on NEDMed data. We additionally report scores on aida-B in order to monitor catastrophic forgetting. Finally, to measure the contribution of aida-train, we trained models using NEDMed-train alone, and found lower NEDMed-dev scores across the board (a roughly 3% drop in F1).

We find that our DCA-Prompt system yields superior performance in both the zero-shot and transfer learning scenarios. Notably, the zero-shot performance of DCA-Prompt is higher than all three metrics of the baseline DCA system.

5 Conclusions and Future Work

These results indicate that pattern exploitation training can effectively be utilized for named entity disambiguation. While results are not state-of-the-art when used in isolation, combining an ADAPET-based classifier with an existing model which can incorporate global context, such as DCA, improves the capacity of that model to flexibly adapt to data from different domains. Our new NEDMed dataset both provides evidence for this and represents a new domain-specific benchmark which can be used by future NED research.

There are a number of ways in which this work could be built upon in the future. This work focused on shifts in domain related to the documents in which mentions are extracted from. Another important type of domain shift relates to large changes in the underlying KB. While we expect the system would be able to adapt, this has not been quantified. Additionally, the optimal strategy for designing patterns for use with ADAPET-style techniques is still an open research question (Liu et al., 2021); as this work relied on human-produced patterns, it is certainly possible that accuracy or data requirements could be improved with more clever pattern design.

5.1 Risks and Limitations

The authors of this paper believe that this work does not introduce any unique risks or limitations; however, we shall note some which are inherent to named entity disambiguation in general. As it is a central inspiration of this work, one of the most noteworthy limitations is that of cross-domain applicability. That is, the performance of our NED system on a given datum remains a function of how closely that datum reflects the data upon which the system was trained. While our work narrows the gap in performance, it remains the case that data

[1] The NEDMed dataset is available to download at `https://github.com/basis-technology-corp/NEDMed`.

Data	Model	aida-B Accuracy	NEDMed-dev F1
AIDA	DCA	**95.2 ± 0.1**	72.8 ± 1.2
	ADAPET	81.0 ± 0.5	53.8 ± 0.2
	DCA-Prompt	95.1 ± 0.1	79.4 ± 0.2
AIDA+NEDMed-train	DCA	95.1 ± 0.1	77.7 ± 0.4
	ADAPET	80.6 ± 0.3	54.6 ± 0.2
	DCA-Prompt	94.9 ± 0.1	**80.2 ± 1.4**

Table 2: Transfer learning performance on aida-B and NEDMed-dev datasets when trained on aida-train ("AIDA") and NEDMed-train. Best scores and our systems are in bold. "ADAPET" denotes our ADAPET-NED system.

from extremely different domains (e.g. a different KB which is dissimilar from Wikipedia) will not be linked as accurately as data from the same domain.

The primary societal risk of NED systems is that of surveillance. While it does not increase the ability to collect data which *may* pertain to a given entity, well-performing NED systems reduce the amount of human labor which is needed to filter through false positives returned by data collection streams. This reduces the total amount of effort required for organizations to precisely aggregate information about specific entities across large quantities of data.

Acknowledgements

The authors would like to thank BasisTech's Karin Lin, James Shatalov-Stein, and Zachary Yocum for their work on the production of the NEDMed dataset, along with Chris Mack and Jimmy Moon for their project management support. This research was done in coordination with Saltlux Inc. and partially funded by the Exobrain Project.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16, Trento, Italy. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of NAACL-HLT 2019*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Mach. Learn.*, 20(3):273–297.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Zhaochen Guo and Denilson Barbosa. 2014. Robust entity linking via random walks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, page 499–508, New York, NY, USA. Association for Computing Machinery.

Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard de Melo, and Gerhard Weikum. 2011a. YAGO2: Exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, page 229–232, New York, NY, USA. Association for Computing Machinery.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011b. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations,* *ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, page 457–466, New York, NY, USA. Association for Computing Machinery.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. *8th International Conference on Learning Representations, ICLR 2020*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ArXiv*, abs/2107.13586.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019*.

David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, page 509–518, New York, NY, USA. Association for Computing Machinery.

Isaiah Onando Mulang', Kuldeep Singh, Chaitali Prabhu, Abhishek Nadgeri, Johannes Hoffart, and Jens Lehmann. 2020. Evaluating the impact of knowledge graph context on entity disambiguation models. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2157–2160, New York, NY, USA. Association for Computing Machinery.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384, Portland, Oregon, USA. Association for Computational Linguistics.

Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pages 166–171.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Derek Tam, Rakesh R. Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. 2021. Improving and simplifying pattern exploiting training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4980–4991, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xiyuan Yang, Xiaotao Gu, Sheng Lin, Siliang Tang, Yueting Zhuang, Fei Wu, Zhigang Chen, Guoping Hu, and Xiang Ren. 2019. Learning dynamic context augmentation for global entity linking. 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 271–281, Hong Kong, China. Association for Computational Linguistics.

A Formal Description of Named Entity Disambiguation

In this section, we provide a formal description of the terminology of named entity disambiguation, building upon the brief outline in Section 1. These terms are shown in Figure 1.

Recalling from Section 3, in NED, we presume that, for a given input text, a candidate generator

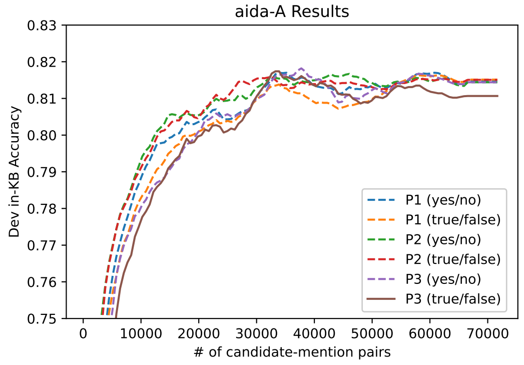


Figure 4: Size-10 moving average of few-shot learning curves for ADAPET patterns shown in Table 3. "(yes/no)" and "(true/false)" indicate the values used for the "[MASK]" token. The x-axis indicates the number of mention-candidate pairs used during training.

detects n *mentions* in the document (denoted $\mathcal{M}$), which are the textual surface forms of entities in the document which may or may not link to a KB entity (much like the entities extracted by a named entity recognition system). A list of *candidates* c_i (collectively denoted $\mathcal{C}$) is associated with each mention m_i in the document. The goal of a named entity disambiguation system is to inspect the output from the candidate generator and determine the correct candidate for each mention. Typically, this is done via scoring each mention-candidate pair. Mathematically, this is done with a scoring function s as follows:

$$f(m_i; \mathcal{M}, \mathcal{C}) \triangleq \underset{c_{i,k}}{\operatorname{argmax}}\, s(m_i, c_{i,k}; \mathcal{M}, \mathcal{C}) \quad (1)$$

Note that s is able to incorporate arbitrary amounts of context in its decision-making process (e.g. other mentions, candidates of other mentions, etc.). In both most recent and this work, s is a neural network trained via gradient descent.

B Pattern and Transformer Analysis

Before we could answer our research questions, it was necessary to understand which choice of pattern makes the most sense for this task. To this end, we needed to first produce a set of patterns to choose from. Unlike many of the SuperGLUE tasks, there was not an obvious choice for what a good pattern may be, so we experimented with a few, shown in Table 3; nonetheless, this served as a good exercise in how to design patterns for use with these systems, which should prove helpful for others. Note that we did experiment with prompt tuning (Lester et al., 2021), but this did not give good results. First, there was a choice of whether to create a binary or n-ary pattern (i.e. "is this the correct candidate?" vs. "which of these are the correct candidate?"). Our work uses the former, as some preliminary empirical results from the latter yielded poor results. For such binary patterns, we need to include two pieces of information: a snippet from the input document (the mention, along with its surrounding context), and some sort of information about the candidate that we want to evaluate. Additionally, these two pieces of information need to be bridged together by the pattern in such a way that we have a masked token which can be filled in to answer the "is this the correct candidate?" question. One notable aspect of these patterns was the decision to utilize a distinct "[MENTION]" token inside of the first pattern, in place of the surface form of the mentioned entity. This is done in order to more directly relate the in-context mention to the question at the end of the pattern, as the candidate description presumably contains many instances of the mention's surface form. To represent the candidates in the model inputs, we rely on the existence of textual descriptions of each entity (ENTDESC).

To determine which would be optimal, we train each of the three patterns from Table 3 on aida-train in order to compare and contrast two pieces of information: the overall accuracy on aida-A when using each pattern and which of these patterns required the least amount of training data to achieve this accuracy. As done in Tam et al. (2021), we tune the transformer for a single epoch over the data, and we sample the aida-A performance every 640 candidate-mention pairs. As is standard, we evaluate aida-A using in-KB accuracy, which is simply the accuracy for all aida-A mentions whose correct answer is in the knowledge base. The results of this analysis are shown in Figure 4. As mentioned in Section 3, we fine-tune various HuggingFace Transformers (Wolf et al., 2020) models to produce the word "yes" or "true" for correct mention-candidate pairs and the word "no" or "false" for incorrect ones. We find that all patterns perform roughly the same (whether using "yes" and "no" or "true" and "false" as the pattern output), with the exception of P3 with a "true"/"false" output (which performs slightly worse). We additionally experimented with ensembling the three patterns together, but this yielded performance worse than using patterns in isolation. As it yielded the greatest overall

ID	Pattern
P1	[CLS] <LCTX> [MENTION] <RCTX>. <ENTDESC>. Is <MENTIONTXT> [MENTION]? [MASK]. [SEP]
P2	[CLS] <LCTX> <MENTIONTXT> <RCTX>. Description: <ENTDESC>. Is this a good description of <MENTIONTXT>? [MASK] [SEP]
P3	[CLS] <ENTDESC>. In the following, does <MENTIONTXT> refer to this entity? [MASK]. <LCTX> <MENTIONTXT> <RCTX>. [SEP]

Table 3: Patterns used during experiments. Items in square brackets ([ ]) are special tokens, and items in angle brackets (<>) are substituted with information from the problem input. LCTX and RCTX are the text to the left and right of the entity mention (respectively), MENTIONTXT is the surface form of the mention, and ENTDESC is the description of the candidate under consideration. Colors represent different logical segments of each piece of the input. [MENTION] is the special token used in place of the original mention (only in P1), and [MASK] is what we prompt the language model to substitute into the input.

Pretrained Transformer	Elbow	Avg. In-KB Acc.	Final In-KB Acc.
bert-base-uncased	**8304**	75.9	81.4
bert-large-uncased	9584	**78.1**	**82.1**
roberta-base	10224	75.8	81.0
spanbert-base-cased	11504	76.0	80.1
longformer-base-4096	10224	76.3	80.6

Table 4: Comparison of ADAPET results using various pretrained transformers on aida-A dataset. To compute the elbow locations, the Kneedle algorithm (Satopaa et al., 2011) was used on smoothed versions of the curves in Figure 5 (smoothing was done by averaging each data point with its immediate neighbors). Average in-KB accuracy is the mean accuracy across all points in training (higher values indicate better few-shot learning ability).

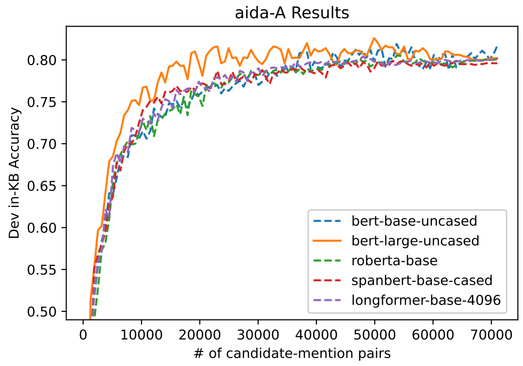


Figure 5: Few-shot learning curves for ADAPET models based on different pretrained transformers. The x-axis indicates the number of candidate-mention pairs used while training.

score, for the analysis in Section 4, we utilize P2 with “true”/“false” as our ADAPET pattern.

Furthermore, we needed to understand which pretrained language model would provide the best performance on this task when fine-tuned with the ADAPET training procedure. To this end, we trained a number of models with different pretrained transformers, with the results in Figure 5 and Table 4. While they all largely converged to a similar in-KB accuracy on the aida-A dataset, the bert-large-uncased model reached this value more rapidly than the other transformer models (quantified by its average accuracy), so it was ultimately chosen for the experiments in Section 4. Notably, all of these models other than Longformer (Beltagy et al., 2020) accept inputs up to a maximum length, so our inputs were trimmed to a maximum length of 256 (the trimming was done in a manner balanced across the color-coded segments of Table 3, with the segment contain the "[MASK]" token not being truncated; this strategy is roughly equivalent to that which is used in Tam et al. (2021)). We note that this is well above the average length of inputs. Additionally, we investigated Longformer as an means of reducing the amount of truncation required, but increasing the length did not yield any noticeable improvement in overall performance over bert-large-uncased.

C Training Details

For the ADAPET model, as described in Appendix B we use HuggingFace’s bert-large-uncased (Lan et al., 2020) model as our base model. Each ADAPET input was truncated to a length of 256, and a batch size of 16 is used for the gradient updates. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate $\gamma = 10^{-5}$ and weight decay of $\lambda = 10^{-2}$. The learning rate is updated according to a linear scheduler.

For the DCA model, we adopt the same hyperparameters as Yang et al. (2019), using the Adam (Kingma and Ba, 2015) optimizer with a learning rate of $\gamma = 2 \times 10^{-4}$. To limit the scope of these experiments, we focus on the best-performing DCA configuration, which is based on a supervised learning training strategy (referred to in the original paper as DCA-SL), with mentions ordered by offset. Reported scores are the best performance measured over up to[2] 500 epochs.

For the combined DCA-Prompt model, we first train the ADAPET model on the dataset and then feed its outputs into the DCA model during a separate training session. This effectively means that we train the ADAPET model and freeze its weights when training the final DCA-Prompt model. Future work will aim to model these two components end-to-end.

Experiments were run using a single NVIDIA Tesla T4 GPU on a Google Cloud Platform `n1-standard-8` machine. The ADAPET model takes roughly 18 hours to fully train and evaluate for a single pattern. The DCA model takes roughly four hours to fully train and evaluate.

C.1 Dataset Information

For the bulk of our baseline experiments, we utilize the AIDA CoNLL-YAGO NED dataset (Hoffart et al., 2011b), as provided by Yang et al. (2019). This dataset is split into three pieces: aida-train, containing 18,448 mentions across 942 documents; aida-A, containing 4,791 mentions across 216 documents and typically used as a development set; and aida-B, containing 4,485 mentions across 230 documents and typically used as an evaluation set. Each item from this version of the dataset consists of a mention and a list of candidate Wikipedia entities. Less than 1% of the mentions in aida-A and aida-B do not include the correct candidate in their lists; as with Yang et al. (2019)'s work, these are skipped when evaluating models.

[2] If performance on the development dataset does not improve after 100 epochs, training is terminated early.

D NEDMed Datasheet

This datasheet template is taken from Gebru et al. (2021).

Motivation

For what purpose was the dataset created? *Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

The goal was to create an English named entity linking dataset based on health-related text.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
This annotated dataset was produced by BasisTech.

Who funded the creation of the dataset? *If there is an associated grant, please provide the name of the grantor and the grant name and number.*

The production of this dataset was funded by BasisTech.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? *Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

The items in the dataset are documents annotated with metadata.

How many instances are there in total (of each type, if appropriate)?
There are 110 documents in the dataset. Of which, 66 comprise NEDMed-train and 44 comprise NEDMed-dev. The following is the breakdown of mention types in NEDMed-train:

And the following is the breakdown for NEDMed-dev:

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? *If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not*

Type	#	# in Wikipedia
DISEASE	592	590
LOCATION	319	309
NATIONALITY	129	129
ORGANIZATION	357	339
PERSON	388	226
PRODUCT	96	95
RELIGION	7	7
SUBSTANCE	264	264
SYMPTOM	363	361
TITLE	150	146
TREATMENT	174	174

Type	#	# in Wikipedia
DISEASE	550	550
LOCATION	169	169
NATIONALITY	110	110
ORGANIZATION	185	168
PERSON	173	51
PRODUCT	58	58
RELIGION	6	6
SUBSTANCE	175	175
SYMPTOM	197	197
TITLE	68	68
TREATMENT	150	150

representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

No. This dataset contains a sample of documents taken from https://theconversation.com/ and https://en.wikinews.org/.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? *In either case, please provide a description.*

Each instance is a text document that includes metadata with information such as source, publication date, and language. Entities in the document have been annotated with character offsets, knowledge base identifiers, and types. The possible types according to annotation guidelines were Location, Organization, Person, Product, Nationality, Religion, Title, Disease, Symptom, Substance, and Treatment.

A (visually rendered) example of an annotated sub-section of a document is the following (brackets have been placed around annotated entities, and entities with the same color represent ones which are linked to the same Wikipedia entity):

On Wednesday, the total number of confirmed [deaths] linked to [SARS-CoV-2 coronavirus] [infections] surpassed 100,000 in the [United States], [Johns Hopkins University] data indicated. The [coronavirus] causes [COVID-19], a sometimes-fatal [disease]. The milestone came just under a month after the total number of confirmed [infections] in the [United States] surpassed one million on April 28.

Is there a label or target associated with each instance? *If so, please provide a description.*

Each document in the dataset contains entity mentions, which are associated with a knowledge base identifier (either a Wikidata QID or a custom knowledge base ID) and an entity type.

Is any information missing from individual instances? *If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? *If so, please describe how these relationships are made explicit.*

N/A

Are there recommended data splits (e.g., training, development/validation, testing)? *If so, please provide a description of these splits, explaining the rationale behind them.*

Yes. The dataset is split into a training dataset (NEDMed-train) and a development/evaluation dataset (NEDMed-dev). This was done by randomly partitioning the documents.

Are there any errors, sources of noise, or redundancies in the dataset? *If so, please provide a description.*

There is possible noise in the dataset. All data was annotated manually, but the final Krippendorff's α value (pairwise inter-annotator agreement) for the NER annotations was 0.768 and for the linking annotations was 0.767. This means that there remained some level of disagreement among the annotators, which could manifest as noise in the data.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? *If it links to or relies* *on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

The linking information in the dataset refers to Wikidata entities. These are publicly available without restriction and will not change.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? *If so, please provide a description.*

No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? *If so, please describe why.*

Yes. Some articles mention sensitive mental health topics such as suicide.

Does the dataset relate to people? *If not, you may skip the remaining questions in this section.*

Not directly. The articles in the dataset relate to mental health; these may be news stories involving people.

Does the dataset identify any subpopulations (e.g., by age, gender)? *If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

N/A

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? *If so, please describe how.*

N/A

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? *If so, please provide a description.*

N/A

Any other comments?

Collection Process

How was the data associated with each instance acquired? *Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

The document text was collected by searching the sites https://theconversation.com/ and https://en.wikinews.org/ for articles related to mental health. Each document was then annotated by a minimum of two human annotators. In the cases where the annotators disagreed, an adjudication process was used to determine the final set of annotators.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? *How were these mechanisms or procedures validated?*

The unannotated data was collected manually by two employees of the BasisTech data team. Annotators were provided with a set of instructions describing how to annotate for named entities and their links. Annotation was done using an internal proprietary NLP annotation tool, which allows metrics such as inter-annotator agreement to be measured across an annotation project.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The documents were chosen by hand based on their content and metadata in order to target news topics related to health and mental health.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The collection of the raw data was performed by employees of BasisTech. Annotation was performed by three experienced Israeli contractors with whom Basis had worked with prior and were compensated at $15-30 per hour. One contractor was a native English speaker, and the other two were native Hebrew speakers with high levels of English competency. The arbitration process for annotation conflicts was performed by a BasisTech employee who is a native English speaker.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? *If not, please describe the timeframe in which the data associated with the instances was created.*

The unannotated documents were collected between August 5th, 2020 through August 10th, 2020. The original publication of the documents ranged from April 5th, 2005 through May 29th, 2020.

Were any ethical review processes conducted (e.g., by an institutional review board)? *If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

No.

Does the dataset relate to people? *If not, you may skip the remaining questions in this section.*

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
N/A

Were the individuals in question notified about the data collection? *If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

N/A

Did the individuals in question consent to the collection and use of their data? *If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? *If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? *If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? *If so, please provide a description. If not, you may skip the remainder of the questions in this section.*

The documents were tokenized with BasisTech's Rosette® Text Analytics linguistic analysis software before annotation (entity mention annotations align with token boundaries).

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? *If so, please provide a link or other access point to the "raw" data.*

The original text of each collected document is included in each instance in the dataset.

Is the software used to preprocess/clean/label the instances available? *If so, please provide a link or other access point.*

No, proprietary software was used.

Any other comments?

Uses

Has the dataset been used for any tasks already? *If so, please provide a description.*

Yes, this paper.

Is there a repository that links to any or all papers or systems that use the dataset? *If so, please provide a link or other access point.*

Not at present.

What (other) tasks could the dataset be used for?
Named entity recognition (NER).

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? *For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

Not to our knowledge.

Are there tasks for which the dataset should not be used? *If so, please provide a description.*

No.

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? *If so, please provide a description.*

Yes. The data shall be publicly released alongside this paper.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) *Does the dataset have a digital object identifier (DOI)?*

The dataset is available for download on GitHub at `https://github.com/basis-technology-corp/NEDMed`.

When will the dataset be distributed?
It is already distributed.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? *If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

The dataset is released under Creative Commons licenses. See the README file for further details.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

The original unannotated data was released under Creative Commons licenses (CC BY-ND 4.0 and CC BY 2.5).

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?
BasisTech will be supporting this dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
`pblair@basistech.com`

Is there an erratum? *If so, please provide a link or other access point.*

No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? *If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?*

No.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? *If so,* *please describe these limits and explain how they will be enforced.*

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? *If so, please describe how. If not, please describe how its obsolescence will be communicated to users.*

N/A

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? *If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.*

Not officially. While we welcome additional data in this area, NEDMed was curated through a manual process with a specific set of annotators, so we do not feel that it would be appropriate to enable further contributions from external sources. Instead, later datasets in this domain should be used alongside NEDMed.

Any other comments?

E Additional Dataset Information

The AIDA CoNLL-YAGO dataset used to train and evaluate models in this work is released at https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/ambiverse-nlu/aida/downloads under a CC BY 3.0 License. Other (non-NEDMed) evaluation datasets were sourced from https://github.com/YoungXiyuan/DCA.