A Hybrid Model for Few-Shot Attribute Extraction Using Prototypical Networks and k-Nearest Neighbors

Taiyu Zhang[1], Yuqing Ni[1], Ziyang Guo[2]

1. Key Laboratory of Advanced Control for Light Industry Processes (Ministry of Education), School of Internet of Things Engineering, Jiangnan University, Wuxi, 214122, China

2. Huawei Technologies Company Limited, Shenzhen, 518129, China
E-mail: yuqingni@jiangnan.edu.cn

Abstract: In the accelerated progress of globalization, China customs play a vital role in safeguarding national sovereignty and public health, especially in the inspection and quarantine of imported goods. However, varying regulatory standards across countries and the emergence of new types of substances pose challenges for customs in detecting and identifying relevant substances. Attribute extraction, a key task in Natural Language Processing (NLP), is crucial for extracting compliance-related information from unstructured text for customs supervision. This paper proposes a few-shot attribute extraction model that combines Prototypical Networks, k-Nearest Neighbors (kNN), and Conditional Random Fields (CRF) decoding. The model uses Bidirectional Encoder Representations from Transformers (BERT) to generate embeddings, constructs attribute prototypes, and enhances the perception of class boundaries with kNN scores. The CRF decoder integrates these components to optimize label sequence prediction. Experiments show that the model achieves higher F1 scores than baseline models such as BERT-CRF and ProtoNet under the 5-way-10-shot setting.

Key Words: Few-Shot Learning, Natural Language Processing, k-Nearest Neighbors, Attribute Extraction

1 Introduction

In the accelerated progress of globalization, the scale and frequency of international trade are constantly increasing. China customs, as the key guardian of import and export goods, bear the multiple missions of safeguarding national sovereignty, maintaining social order, and promoting economic prosperity. This is especially true in the inspection and quarantine of imported food and pharmaceutical products, where the work of customs is of vital importance, as these products directly relate to the life and health of the general public. However, due to significant differences in legal frameworks and regulatory standards among countries, customs face many complex and thorny challenges in the implementation of regulatory tasks.

Take Canada as an example. The country officially implemented the Cannabis Act on October 17, 2018, legalizing the use of recreational cannabis, becoming the second country in the world to fully lift cannabis restrictions after Uruguay. However, in China, cannabis and its main active ingredient, tetrahydrocannabinol (THC), are explicitly classified as strictly controlled drugs, and their import and use are strictly prohibited. In 2022, customs in Hong Kong, China, seized a batch of food imported from Canada, which contained cannabidiol (CBD) that is legally sold in Canada. Although CBD is allowed for food additives in Canada, it is still considered as a prohibited item in China due to its association with cannabis. Similar regulatory conflicts also occur in the management of melatonin. In the United States, melatonin is widely sold as an over-the-counter dietary supplement, while in China, it is strictly classified as a drug and must be approved by the National Medical Products Administration before it can be legally marketed.

This work was financially supported by the National Key R&D Program of China (2023YFF1104900), the National Natural Science Foundation of China (Grant 62303196), and the Natural Science Foundation of Jiangsu Province (Grant BK20231036).

In the actual operation of customs supervision, accurate detection and identification of relevant substances are crucial. This requires customs officers to have a clear understanding of the properties, components, uses, and legal status of various substances, and to use corresponding detection methods for precise inspection. However, with the continuous emergence of new types of substances, data on these substances are often scarce and difficult to obtain, and traditional attribute extraction methods fail to meet the customs' needs for rapid and accurate identification. Therefore, how to efficiently extract key attributes from limited information has become an important issue that needs to be urgently resolved.

Attribute extraction is a core task in the field of Natural Language Processing (NLP), aimed at extracting attribute information of entities from unstructured text. In scenarios such as customs inspection, extracting attribute information such as the existence of specific substances or their legal status is crucial to ensure compliance. However, under the condition of limited training samples, attribute extraction faces many challenges: the lack of training data makes it difficult for models to capture semantic patterns, while the complexity of the task itself requires models to have strong representation and reasoning capabilities. Therefore, how to achieve efficient and accurate attribute extraction under the condition of limited samples has become an important direction for current research.

Few-shot learning (FSL) is an emerging learning paradigm that aims to quickly learn and generalize to new tasks using only a small number of labeled samples[1]. Due to its potential in resource-constrained scenarios, FSL has attracted widespread attention.

Named Entity Recognition (NER) tasks in similar low-data scenarios have also received extensive attention, with a large number of studies devoted to this area. For example, one study focuses on Transformer-based self-supervised pre-trained language models (PLMs) and explores three or-

thogonal approaches to enhance the model's generalization ability in low-sample environments[2].

The challenges of few-shot attribute extraction are mainly reflected in the following aspects: First, due to the limited number of training samples, models are prone to overfitting and struggle to achieve good generalization on new data[3, 4]. Second, attribute extraction involves complex semantic understanding and structured information extraction, requiring models to possess deep language comprehension capabilities. Prototypical Networks, as an effective meta-learning method, can quickly adapt to new tasks by learning category prototypes[5]. The core idea of Prototypical Networks is to embed the support samples of each category into a vector space and calculate the prototype of each category (i.e., the centroid vector of the category). Some studies have proposed Contrastive Prototypical Networks (CPN), which enhance the effectiveness and quality of embeddings through contrastive learning. In addition, it introduces pseudo-prototype strategies and data augmentation techniques to address the problem of data scarcity[6]. When facing new query samples, the model classifies them by calculating the similarity between the query samples and the category prototypes. Meanwhile, the k-Nearest Neighbors (kNN) algorithm, as a classic non-parametric method, has unique advantages in few-shot scenarios because it makes predictions based on similarity measurements[7]. By directly utilizing the representations of the training samples (support set), kNN can effectively complement the shortcomings of parametric models. Therefore, a hybrid model combining Prototypical Networks, the kNN algorithm, and Conditional Random Fields (CRF) can effectively enhance the model's generalization ability and sequence prediction accuracy[8].

To address the task of attribute extraction for emerging substances in customs scenarios where data is scarce and sample sizes are small, and to improve the accuracy of the model, this paper proposes a few-shot attribute extraction model based on Prototypical Networks, kNN, and CRF decoding to address the attribute extraction problem in data-scarce scenarios. The main contributions of this study are as follows:

- A hybrid model combining Prototypical Networks and kNN is proposed, enhancing the model's perception of class boundaries.
- A CRF decoder is introduced to integrate the emission scores from Prototypical Networks and kNN, optimizing the prediction performance of the label sequence.

2 Related Work

2.1 Attribute Extraction

Attribute extraction is an important task in the field of NLP, aimed at identifying and extracting attribute information related to specific entities from unstructured text. These attributes typically include features, states, behaviors, or other important information related to the entity. For example, in medical texts, attribute extraction can extract information about patients' symptoms, treatment effects, or side effects of drugs. Research on attribute extraction can be roughly divided into three categories: rule-based methods, statistical learning-based methods, and deep learning-based methods.

Rule-based attribute extraction methods rely on manually written rules, such as domain-specific lexicons, semantic patterns, and syntactic structures. These rules are usually designed by domain experts according to the needs of specific tasks. For example, in the introduction of pharmaceutical products, the attribute related to "ingredients" can be identified by defining vocabulary associated with it (such as "Astragalus", "Angelica dahurica"). However, rule-based methods have obvious limitations: they require a lot of manual work to build rules and are difficult to cover all possible expressions, especially when facing complex natural language. In addition, these rules often lack understanding of contextual information, which can lead to misjudgments.

Statistical learning-based attribute extraction methods rely on feature engineering, transforming text data into feature vectors. These methods often use traditional machine learning algorithms, such as Hidden Markov Models (HMM), Maximum Entropy Models (ME), Support Vector Machines (SVM), and CRF. These methods learn the relationship between features and attributes through annotated data. For example, features such as part-of-speech tagging and dependency syntax analysis can be used to improve the accuracy of attribute extraction. However, these methods require a large amount of annotated data to train the model and rely on manually designed features, which limits their adaptability and scalability in new fields.

In recent years, deep learning-based attribute extraction methods have gradually become mainstream. These methods use neural networks to automatically learn contextual information and semantic features in the text, thereby improving the accuracy and robustness of attribute extraction. For example, Recurrent Neural Networks (RNN) and their variants (such as Long Short-Term Memory networks, i.e., LSTM, and Gated Recurrent Units, i.e., GRU) are widely used to capture sequential information in the text. Convolutional Neural Networks (CNN) are also used to extract local features, thereby identifying patterns related to attributes. In addition, the emergence of pre-trained language models (such as Bidirectional Encoder Representations from Transformers, i.e., BERT) has brought new breakthroughs to the task of attribute extraction[9]. These models, through pre-training on large-scale text data, can capture rich language knowledge and semantic information, thus performing well when fine-tuned for specific tasks[10–13]. For example, BERT-based attribute extraction models can better understand attribute information in context through attention mechanisms, thereby achieving more accurate attribute recognition and classification.

2.2 Few-Shot Learning

Few-Shot Learning aims to rapidly learn and generalize to new tasks using a limited number of labeled samples. In recent years, FSL has garnered significant attention in the field of NLP, especially in scenarios where data resources are scarce. The primary methods include metric learning-based approaches, meta-learning, and transfer learning[14].

- Metric Learning-Based Approaches: These methods achieve classification by learning similarity metrics between samples. For example, Vinyals et al. propose

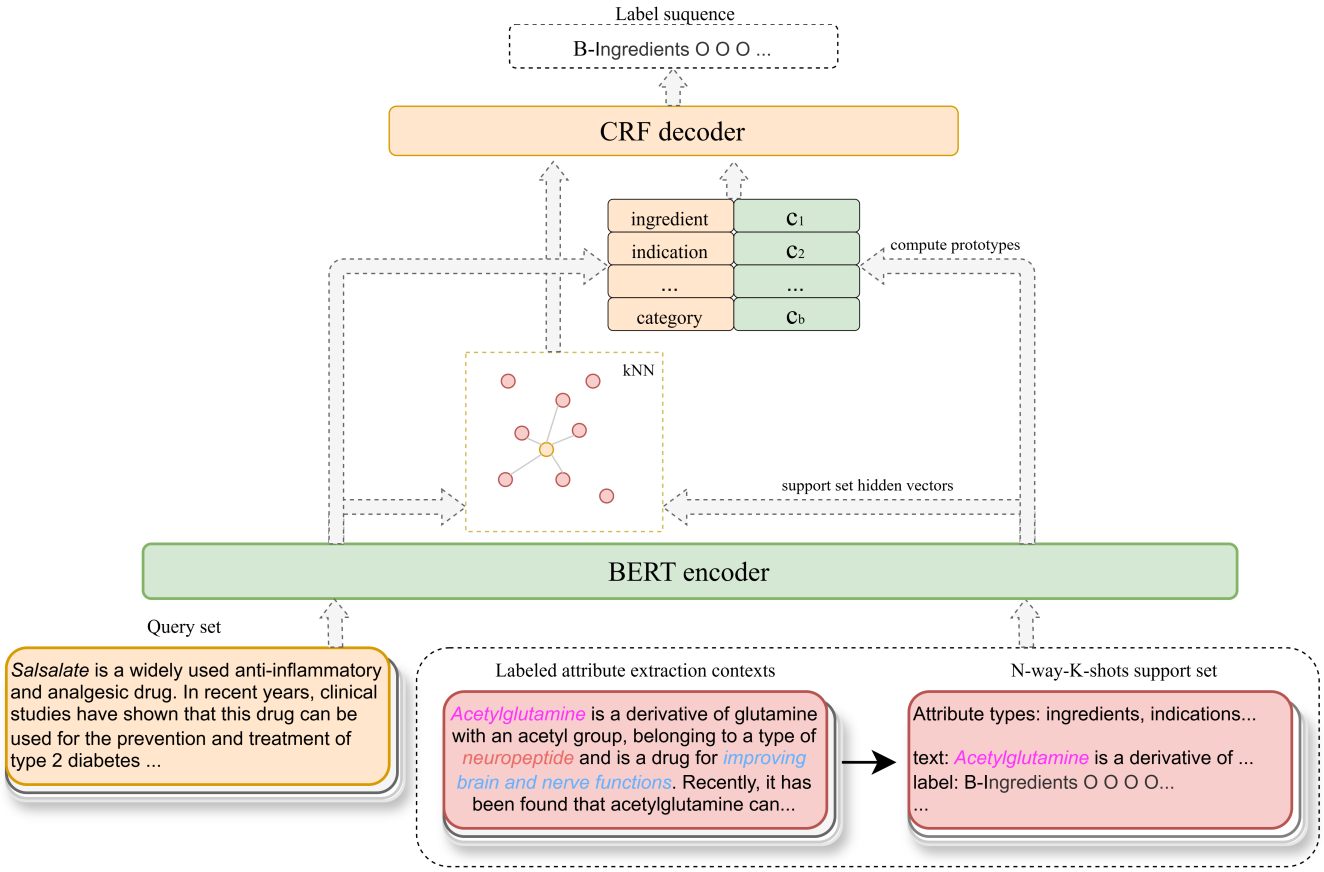

Fig. 1: The framework of this study

Matching Networks, which can do prediction by measuring the similarity between support set samples and test samples[15]. Snell et al. introduce Prototypical Networks, which can do classification by computing prototype representations[5].

- Meta-Learning: This approach focuses on learning how to quickly adapt to new tasks. Finn et al. propose Model-Agnostic Meta-Learning (MAML), which optimizes the initial parameters to enable rapid convergence of models on new tasks[16]. Ravi and Larochelle modele the optimization process as a meta-learning problem[17].
- Transfer Learning: This involves fine-tuning pre-trained models on new tasks. Devlin et al. develope the BERT model, which significantly improve performance in few-shot scenarios through large-scale pre-training[9]. Howard and Ruder propose ULMFiT, which adapts to specific tasks through gradual fine-tuning[18].

2.3 k-Nearest Neighbors

The k-Nearest Neighbors is a classic non-parametric method that performs classification or regression by calculating the similarity between test samples and training samples. In NLP, kNN has been applied to tasks such as text classification and sentiment analysis. For example, Zhang et al. combined kNN with prompt templates and added a position-aware biaffine module to enhance the model's performance in few-shot named entity recognition tasks[19]. Wang et al. also introduced kNN into named entity recognition tasks, determining the entity categories by searching for the nearest neighbors[20].

In recent years, the integration of kNN with deep learning has shown promising results. For instance, Papernot et al. proposed Deep kNN, which combines deep features with kNN to enhance performance. In few-shot learning scenarios, kNN has garnered attention due to its similarity-based prediction capabilities[21].

3 Method

The framework of our work is shown in Figure 1, and this section will provide a detailed explanation of each module.

3.1 Task Definition

In the attribute extraction task, the input data is usually a text description, and the goal is to extract specific attribute values from the text. The input is a text description of a material, represented as a sequence $X = \{x_1, x_2, \ldots, x_n\}$, where x_i is a word or sub-word unit. The output is the corresponding label sequence, denoted as $Y = \{y_1, y_2, ...y_n\}$, where y_i represents the label corresponding to x_i. We implement the attribute extraction task using sequence labeling with BIOE (e.g., B = Begin, I = Inside, O = Outside, E = End) tags, extracting the sentences or phrases containing each attribute value and assigning attribute type labels to each word. We define the task in an N-way-M-shot setting, where each task consists of N attribute types and each attribute type has M support samples[15]. The support set is used to construct the class prototypes, and the query set is

used to evaluate the model performance.

3.2 Prototype Network

We use pre-trained language models (such as BERT) to encode the input text and obtain the embedding representations for each word.

BERT encodes the input sequence through multiple layers of Transformer to generate context-related representations. For the input sequence X, the encoder outputs the representation as

$$H = \text{BERT}(X) = \{f_\theta(x_1), f_\theta(x_2), \ldots, f_\theta(x_n)\}$$

where $f_\theta(x_i) \in R^d$ is the hidden representation of the i-th word, and d represents the number of layers in the BERT encoder's hidden layer.

For each attribute type b, we compute the mean of the embeddings of all words of that type in the support set as the prototype for that attribute type:

$$c_b = \frac{1}{|S_b|} \sum_{x_i \in S_b} f_\theta(x_i)$$

where S_b is the set of all words marked as attribute type b in the support set, and $f_\theta(x_i)$ is the embedding representation of word x_i.

3.3 k-Nearest Neighbors Score

First, use BERT to encode each word in the support set, obtaining its embedding representation. Assume the embedding representation of the words in the support set is $\{f_\theta(x_i)\}_{i=1}^{|S_b|}$, where $|S_b|$ is the size of the support set.

Similarly, encode each word in the query set to obtain its embedding representation $\{f_\theta(x_j)\}_{j=1}^{|Q|}$, where $|Q|$ is the size of the query set.

For each word x_j in the query set, calculate the distance between it and all the embeddings of the words in the support set. The Euclidean distance is commonly used:

$$d(x_j, x_i) = \|f_\theta(x_j) - f_\theta(x_i)\|$$

Then, select the k nearest neighbors $\{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\}$, where $x_{i_1} = \arg\min_{x_i}(d(x_j, x_i))$, and so on.

For each query word x_j, extract the label information from its k nearest neighbors. Specifically, we can calculate the frequency or weighted distance of each label in the k nearest neighbors. For instance, for label y, its score can be represented as

$$\text{kNN\_Score}(x_j, y) = \sum_{m=1}^{k} \frac{1}{d(x_j, x_{i_m})} \cdot I(y_{i_m} = y)$$

where I is an indicator function, which equals 1 when $y_{i_m} = y$, otherwise 0. This score represents the strength of the association between the query word x_j and label y.

3.4 CRF Decoder

In the CRF decoder, we combine the emission scores calculated by the original network with the score by kNN to optimize the prediction of the label sequence. The emission score $U(x_i, y_i)$ consists of two parts:

- **Emission score from the original network:**

$$U_{\text{proto}}(x_i, y_i) = -\|f_\theta(x_i) - c_b\|^2$$

  where c_b is the prototype when y_i is a BIOE label associated with entity type b.

The final emission score is a weighted sum of the two:

$$U(x_i, y_i) = \alpha \cdot U_{\text{proto}}(x_i, y_i) + \beta \cdot \text{kNN\_Score}(x_i, y_i)$$

where α, β are non-negative hyperparameters that balance the contribution of the two emission scores.

The transition score $T(y_i, y_{i+1})$ is an element in the transition matrix of the CRF layer, representing the probability of transitioning from tag y_i to tag y_{i+1}[22].

The conditional probability formula for CRF is

$$p(y \mid x) = \frac{\exp\left(\sum_i U(x_i, y_i) + T(y_i, y_{i+1})\right)}{Z(x)}$$

where $Z(x)$ is the normalization factor, ensuring the sum of probabilities for all possible label sequences is 1.

4 Experiment

4.1 Experimental Setup

4.1.1 Datasets

To verify the effectiveness of the proposed method, we conduct experiments on a dataset of chemical product descriptions that we crawled: this dataset includes various attributes of chemical product, including main ingredients, indications(i.e., the symptoms that the medication targets), categories(e.g., ethosuximide is an anticonvulsant.), and odors. We randomly select $N = 5$ attribute types from it, providing $M = 5$ or $M = 10$ support samples for each attribute type to construct an N-way-M-shot task.

To simulate the few-shot scenario, we divide each dataset into a support set and a query set. The support set is used to construct class prototypes, and the query set is used to evaluate the model's performance.

4.1.2 Evaluation Metrics

We evaluate the model performance using the following metrics:

- F1 Score: A comprehensive measure that considers both precision and recall, used to assess the overall performance of attribute extraction.
- Accuracy: Used to evaluate the correctness of the model in sequence labeling tasks.
- Micro-F1: In multi-class scenarios, the F1 score is calculated by aggregating the prediction results of all classes.

4.1.3 Implementation Details

We use the BERT model as the pre-trained language model. The class prototypes of the Prototypical Network are calculated by taking the mean of the embeddings of words corresponding to the attribute types in the support set. In the kNN module, the value of k is set to 5, and the distance metric used is the Euclidean distance. The hyperparameters α and β (which balance the weights of the Prototypical

Network and kNN score) of the CRF decoder are optimized through grid search within the range of [0.1, 0.9], and are ultimately set to 0.6 and 0.4, respectively.

4.2 Baseline Model

To compare the performance of the proposed method, we select the following baseline models for comparison:

- BERT-CRF: A sequence labeling model based on BERT, combined with a CRF decoder, fine-tuned directly on the support set.
- ProtoNet: A prototypical network that constructs class prototypes using the support set and performs attribute extraction through nearest neighbor classification.
- BiLSTM-CRF: A traditional sequence labeling model combining bidirectional LSTM with a CRF.

4.3 Results

Table 1 presents a performance comparison between the proposed method and baseline models on the chemical product descriptions Dataset (using a 5-way-10-shot setting as an example).

Table 1: Comparison of different methods

Model	Accuracy (%)	F1 (%)	Micro-F1 (%)
BiLSTM-CRF	68.4	66.2	65.8
ProtoNet	80.5	79.1	78.6
BERT-CRF	84.6	83.5	83.0
This study	**87.2**	**86.1**	**85.7**

From Table 1, it can be seen that our method significantly outperforms the baseline models on the drug product description dataset.

4.4 The Impact of Different M Values

To further analyze the impact of the number of support samples on model performance, we test the performance with $M = 5$, 10, and 20 under the 5-way setting. Figure 2 shows the trend of the F1 scores of our method as M changes.

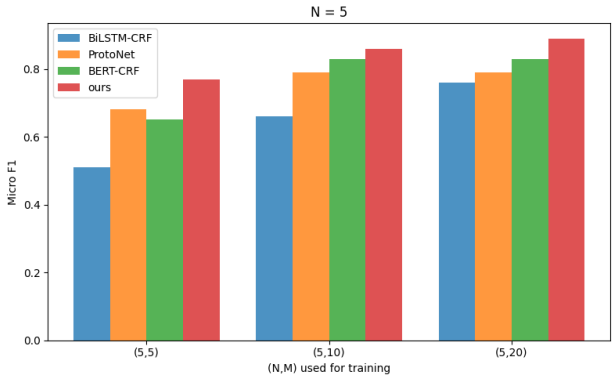

Fig. 2: F1 Scores at different M values

The experimental results indicate that the model's performance improves with the increase of M, but even in the scenario of $M = 5$, our method can still maintain good performance.

5 Conclusion

To tackle the task of attribute extraction for emerging substances in customs scenarios, which are characterized by data scarcity and small sample sizes, and to enhance the accuracy of the extraction, we proposed a novel few-shot attribute extraction model that integrates Prototypical Networks, kNN, and CRF decoder. This hybrid model leverages the strengths of each component to address the challenges of attribute extraction in data-scarce scenarios. Specifically, the Prototypical Networks construct class prototypes to enhance the model's ability to generalize from limited labeled samples, while the kNN algorithm extracts nearest neighbor distance scores to improve the perception of class boundaries. The CRF decoder integrates these scores to optimize the prediction of the label sequence, ensuring consistency and accuracy in sequence labeling tasks.

While our model has shown promising results, there are several directions for future research:

- **Data augmentation:** Incorporating data augmentation techniques to further address the issue of data scarcity in few-shot learning scenarios.
- **Cross-domain adaptation:** Investigating the model's performance in cross-domain few-shot attribute extraction tasks to enhance its applicability in diverse scenarios.
- **Efficiency optimization:** Optimizing the computational efficiency of the model to ensure faster training and inference times.

References

[1] Yisheng Song, Ting Wang, Puyu Cai, Subrota K. Mondal, and Jyoti Prakash Sahoo. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Comput. Surv.*, 55(13s), July 2023.

[2] Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. Few-shot named entity recognition: An empirical baseline study. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10408–10423, November 2021.

[3] Junfan Chen, Richong Zhang, Yongyi Mao, and Jie Xu. Contrastnet: A contrastive learning framework for few-shot text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence, 36*, volume 10 of *AAAI-22 Technical Tracks 10*. AAAI, 2022.

[4] Yiren Jian and Lorenzo Torresani. Label hallucination for few-shot classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7005–7014, 2022.

[5] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. NIPS'17, page 4080–4090, Red Hook, NY, USA, 2017.

[6] Mengjuan Jiang, Jiaqing Fan, Jiangzhen He, Weidong Du, Yansong Wang, and Fanzhang Li. Contrastive prototype network with prototype augmentation for few-shot classification. *Information Sciences*, 686:121372, 2025.

[7] P.K. Syriopoulos, N.G. Kalampalikis, S.B. Kotsiantis, et al. knn classification: a review. *Annals of Mathematics and Artificial Intelligence*, 93:43–75, 2025.

[8] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the*

Eighteenth International Conference on Machine Learning, ICML '01, page 282–289, San Francisco, CA, USA, 2001.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019.

[10] Kiril Gashteovski, Mingying Yu, Bhushan Kotnis, Carolin Lawrence, Mathias Niepert, and Goran Glavaš. BenchIE: A framework for multi-faceted fact-based open information extraction evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4472–4490, Dublin, Ireland, May 2022.

[11] Sabyasachee Baruah and Shrikanth Narayanan. Character attribute extraction from movie scripts using llms. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8270–8275, 2024.

[12] Lin Qiu, Hao Zhou, Yanru Qu, Weinan Zhang, Suoheng Li, Shu Rong, Dongyu Ru, Lihua Qian, Kewei Tu, and Yong Yu. *QA4IE: A Question Answering Based Framework for Information Extraction*, page 198–216. Springer International Publishing, 2018.

[13] M. Sipilä, F. Mehryary, S. Pyysalo, F. Ginter, and Milica Todorović. Question answering models for information extraction from perovskite materials science literature.*arXiv preprint arXiv:2405.15290*, 2024.

[14] Archit Parnami and Minwoo Lee. Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv:2203.04291*, 2022.

[15] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. NIPS'16, page 3637–3645, Red Hook, NY, USA, 2016.

[16] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. ICML'17, page 1126–1135. JMLR.org, 2017.

[17] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*, 2017.

[18] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July 2018.

[19] Mozhi Zhang, Hang Yan, Yaqian Zhou, and Xipeng Qiu. Promptner: A prompting method for few-shot named entity recognition via k nearest neighbor search. *arXiv preprint arXiv:2305.12217*, 2023.

[20] Shuhe Wang, Xiaoya Li, Yuxian Meng, Tianwei Zhang, Rongbin Ouyang, Jiwei Li, and Guoyin Wang. *k*nn-ner: Named entity recognition with nearest neighbor search. *arXiv preprint arXiv:2203.17103*, 2022.

[21] Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018.

[22] Diego F. Silva, Alcides M. e. Silva, Bianca M. Lopes, Karina M. Johansson, Fernanda M. Assi, Júlia T. C. de Jesus, Reynold N. Mazo, Daniel Lucrédio, Helena M. Caseli, and Livy Real. Named entity recognition for brazilian portuguese product titles. In *Intelligent Systems*, pages 526–541, Cham, 2021.