Towards Alleviating Human Supervision for Document-level Relation Extraction

by

Yuxi Feng

B.Sc., Tsinghua University, 2018

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Computer Science)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2024

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Towards Alleviating Human Supervision for Document-level Relation Extraction

submitted by **Yuxi Feng** in partial fulfillment of the requirements for

the degree of **Doctor of Philosophy** in **Computer Science**.

Examining Committee:

Laks V.S. Lakshmanan, Professor, Computer Science, UBC
Supervisor

Giuseppe Carenini, Professor, Computer Science, UBC
Supervisory Committee Member

Mark Schmidt, Associate Professor, Computer Science, UBC
University Examiner

Jane Z. Wang, Professor, Electrical Engineering and Computer Engineering, UBC
University Examiner

Xifeng Yan, Professor, Computer Science, University of California, Santa Barbara
External Examiner

Additional Supervisory Committee Members:

Raymond Ng, Professor, Computer Science, UBC
Supervisory Committee Member

Abstract

Motivated by various downstream applications, there is tremendous interest in the automatic construction of knowledge graphs (KG) by extracting relations from text corpora. Relation Extraction (RE) from unstructured data sources is a key component for building large-scale KG. In this thesis, I focus on the research centered on *Document Level Relation Extraction*. One challenge of *Document Level Relation Extraction* is the lack of labeled training data since the construction of a large in-domain labeled dataset would require a large amount of human labor. To alleviate human supervision on document-level relation extraction, I propose 1) an unsupervised RE method CIFRE which enhances the recall of pipeline-based approaches while keeping high precision; 2) a semi-supervised RE method DuRE when few labeled data are available, by leveraging self-training to generate pseudo text. In order to improve the quality of pseudo text, I also propose two methods (DuNST and KEST) to improve the controllability and diversity of semi-supervised text generation, solving the challenges of inadequate unlabeled data, over-exploitation, and training deceleration. Comprehensive experiments on real datasets demonstrate that our proposed methods significantly outperform all baselines, proving the effectiveness of our methods in unsupervised and semi-supervised document-level relation extraction.

Lay Summary

Relation Extraction (RE) from unstructured data sources is a key component for building large-scale Knowledge Graphs. In this thesis, I focus on the research centered on *Document Level Relation Extraction*. One challenge of *Document Level Relation Extraction* is the lack of labeled training data since the construction of a large in-domain labeled dataset would require a large amount of human labor. To address this, I introduce two methods: CIFRE, an unsupervised RE approach to extract more relation in pipeline-based methods while maintaining accurate extraction, and DuRE, a semi-supervised method that generates additional training data by an additional text generator with given relation triples and reduces the effort of human annotation. Additionally, I propose two text generation methods (DuNST and KEST) to generate more diverse text that follows the instructions better. Comprehensive experiments on real datasets demonstrate the effectiveness of our proposed methods compared to baselines.

Preface

This thesis is a product of a continuous research collaboration with my supervisor Prof. Laks V.S. Lakshmanan. The chapters are based on papers that are either published or are under review.

- Chapter 2 is based on a research paper under submission. This work is collaborated with Michael Simpson and Laks V.S. Lakshmanan, where Michael Simpson was a postdoc at UBC. For this work, I proposed the key idea about missing entity recognition and Chinese idiom resolution, developed the full pipeline, analyzed its properties, conducted the experiments, and wrote the paper. Laks V.S. Lakshmanan and Michael Simpson provided valuable feedback during this research and helped polish the paper.

- The work in Section 3.3 is done during my internship at Microsoft Research Asia under the supervision of Xiaoyuan Yi and Laks V.S. Lakshmanan. Xiaoyuan Yi was a researcher at Microsoft Research Asia. The research paper is published in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), 2023*[25]. In this paper, Xiaoyuan Yi provided the initial idea of self-training on natural language generation, helped prove theorem 1, and polished the writing of the paper. I proposed the noisy algorithm, designed the details of the algorithm, conducted all the experiments, analyzed the performances, and wrote the paper. Laks V.S. Lakshmanan provided valuable feedback during the research and paper writing. Xiting Wang and Xing Xie also collaborated on this project by joining our discussion and providing feedback.

- The work in Section 3.4 is published in *International Joint Conferences on Artificial Intelligence (IJCAI), 2023*[24]. This paper is also collaborated with Xiaoyuan Yi and Xing Xie from Microsoft Research Asia. In this paper, Xiaoyuan Yi proposed the issue of deceleration of self-training, provided computing resources, helped prove theorem 2, and polished the writing of the paper. I proposed to use non-auto regressive module to accelerate the training of self-training, designed the details of the algorithm, conducted all the experiments, analyzed the performances, and wrote the paper. Laks V.S. Lakshmanan provided valuable feedback regarding experiment design and paper writing.

- The work in Chapter 4 is currently under submission and is accepted to *2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)*. I proposed the key idea to address long-tail entity problem by self-training, developed the full pipeline, analyzed its properties, conducted the experiments, and wrote the paper. Laks V.S. Lakshmanan provided ample feedback in the research process and paper writing,

For the work mentioned in Chapter 2, Sections 3.3, Section 3.4, and Chapter 4, I was the lead investigator, responsible for all major areas of concept formation, statement of research questions, data collection, implementation as well as paper composition. Laks V.S. Lakshmanan was the supervisory author on all projects and was involved throughout the project in concept formation, discussions, and paper composition.

Portions of the abstract and introductory text in Chapter 1 are an aggregation of the papers described above.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First, I want to express my sincere gratitude to my supervisor, Prof. Laks V.S. Lakshmanan, for his efforts in guiding me through all the research works presented in this thesis. Needless to say that his unparalleled research ethics and passion are the cornerstones of this work. His valuable insights have paved my path toward higher-level research and his deep expertise has played a vital role in shaping the trajectory of this work. I am also deeply grateful to my supervisory committee: Prof. Giuseppe Carenini and Prof. Raymond Ng for their insightful comments and suggestions. Additionally, I wish to express my heartfelt gratitude to my examination committee, comprising Mark Schmidt, Jane Z. Wang, Paul Gustafson, and Xifeng Yan. Their probing questions and thoughtful suggestions have enriched my perspective by offering a diverse array of viewpoints.

Besides, I want to express my appreciation to my intern mentor Dr. Xiaoyuan Yi, my manager Xing Xie, and all my colleagues in the Social Computing Group at MSRA. I am also grateful to Huawei Research for the excellent opportunity in the team of Dr. Yong Zhang and Weiwei Zhang. I would like to thank all the collaboraters in Huawei: Zhenan Fan, Linzi Xing, Xinglu Wang, Raymond Li, Mohammadreza Pourreza, Morgan Heisler, and Chunyu Chen. A grateful thank you to all the staff in the Department of Computer Science, particularly Joyce Poon and Lara Hall, for their help in dealing with the administrative issues. Another special memorial of Prof. Ruben Zamar for helping me at the beginning of my study at UBC.

I would like to thank my DMM lab buddies for making my Vancouver life such amazing and for all the support: Dujian Ding, Ganesh Jawahar, Prithu Banerjee, Glenn Bevilacqua, Rudra Ranajee Saha, Sarah Elhammadi, Gorisha Agarwal, Md Tawkat Islam Khondaker, Farnoosh Hashemi, Swati Kanwal, Xiang Zhang, and Dorna Dehghani. Special thanks to my roommates Yihan Zhou, Weidong Yin, Rui Huang, and Lemin Li for the support during the Covid pandemic.

The acknowledgments would not be complete without mentioning all of my friends and family members for making this journey meaningful and joyful. Most importantly, I am greatly indebted to my parents. Their unconditional love and care mean everything to me. I am grateful for their steadfast presence by my side, offering guidance and love every step of the way. They have been my pillars of support and my greatest sources of inspiration. Special thanks to Wanxin Li for the unwavering support and encouragement in my last year of Ph.D. study.

Embarking on a Ph.D. journey is often perceived as a path to academic achievement and intellectual fulfillment, yet the reality is far from a constant state of joy and ease. Throughout my process of pursuing a doctoral degree, the journey is the pervasive nature of failure. Setbacks and failures are not anomalies but rather intrinsic components of academic pursuit. Accepting failure as a customary aspect of the PhD journey is pivotal in navigating its complexities and ultimately achieving scholarly growth and development. Fortunately, I have my supervisors, parents, collaborators, and friends who encourage me during moments of anxiety and despair. Their support is crucial for me to complete my Ph.D. study.

Chapter 1

Introduction

There exists a vast amount of *unstructured* text on the web, including blogs, governmental documents, email communications, and chat logs. People are interested in relations between the entities that appear in the texts, like person, organization, and location. To help people understand these relations, a popular idea is to turn these *unstructured* text into *structured* data by annotating semantic information. A knowledge graph (KG) is a knowledge base that uses a graph-structured data model or topology to represent and operate on data. Driven by applications such as fact-checking [15], question answering [82, 121], semantic search [118], and recommendations [105], recent years have witnessed a surge of interest in the automatic construction of knowledge graphs (KG) [19, 51, 75]. Figure 1.1 shows the usual pipeline for constructing knowledge graphs. Among these parts, Relation Extraction (RE) from unstructured data sources is a key component for building large-scale KG. Traditionally, the input of relation extraction module is a text with a list of entities in the text, and the output is a set of (subject, relation, object) triples. Prior works often focus on *sentence-level* attribute classification [41–43] where relation triples are extracted from a single sentence. Recently, the task of *Document-level Relation Extraction* has been proposed, where the task is to extract relations from documents. Due to the significant challenges in modeling long text spans and obtaining high-quality supervision signals, document-level relation extraction has been relatively underexplored.

Current document-level relation extraction methods [89, 138] can discover the semantic relation that holds between two entities under supervised learning. However, these methods typically require lots of manually labeled data for model training, while in practice, these labeled data would be labor-intensive to obtain and error-prone due to human subjective judgments.

This thesis focuses on alleviating human supervision in the task of document-level relation extraction. More specifically, this thesis focuses on two lines of methods that do not require labeled data in training, or that require fewer labeled data, namely *unsupervised*[98] and *semi-supervised* [31] methods, respectively. Among semi-supervised methods, this thesis mainly discusses a typical method Self-Training (ST)[91]. Self-training builds a

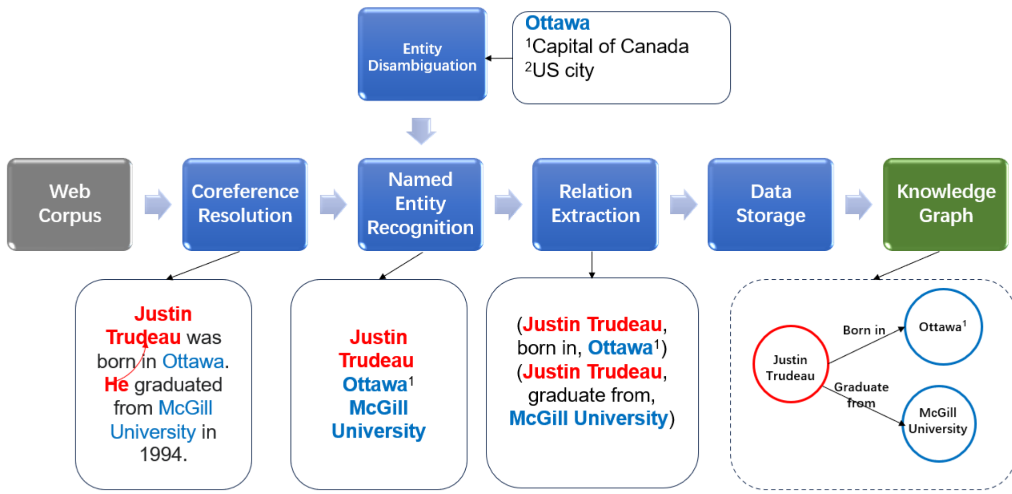


Figure 1.1: A pipeline of automatic construction of knowledge graph.

classifier on a set of labeled data, applies the classifier on a set of unlabeled data to generate pseudo labels, and then uses the pseudo-labeled data to update the classifier parameters. Recently, self-training (ST) has also been proposed in semi-supervised relation extraction [42, 129]. However, ST in document level RE is still underinvestigated.

In the remaining part of this chapter, an overview of relation extraction problem is given in Section 1.1, followed by an overview of self-training method presented in Section 1.2. Then in Section 1.3, I highlight some fundamental shortcomings of the existing research works and take a further look into the motivation of the research described in this thesis. Lastly, Section 1.4.2 summarizes how this thesis addresses the existing shortcomings and presents a brief outline of this thesis.

1.1 Overview of Relation Extraction Problem

It is typical to categorize relation extraction tasks along two dimensions: (1) *open-domain (OpenIE)* or *domain-specific*, (2) *sentence-level* or *document-level*. Traditionally, an OpenIE system (e.g., Stanford OpenIE [1]) is leveraged to extract cross-domain relations. OpenIE systems [2] are traditionally based on dependency parsing and constituency parsing. Recently, there have been approaches that leverage deep learning in extracting relations in the general domain and even in different languages, simultaneously [35, 88]. Self-supervision [41] and Causal models [61] are also applied for OpenIE. There have also been a number of efforts at creating domain-targeted KGs. Examples include the Aristo Tuple KB [72] in the elementary science domain and CPIE [106] in the biomedical domain. In addition, in the financial domain, Benetka et al. [3] extracted quintuples of monetary transactions covering economic events. More recently, Elhammadi et al. [23] built a knowledge graph in the financial domain based on Semantic Role Labeling (SRL) and apposition detection.

Prior works often focus on sentence-level attribute classification [41–43] where relation triples are extracted from a single sentence. Recently, the task of *Document-level Relation Extraction* has been proposed, where the task is to extract relations from documents. Due to the significant challenges in modeling long text spans and obtaining high-quality supervision signals, document-level relation extraction has been relatively underexplored.

Current document-level relation extraction methods [89, 138] can discover the semantic relation that holds between two entities under supervised learning. To address the issue of long input documents to the transformer-based models [101] which usually have a fixed maximum length of input tokens, Zhou et al. [138] made use of a sliding window approach over the input and separate it into several chunks. Zeng et al. [130] leveraged a double-graph network to model the entities and relations within a document. To address the multilabel problem of Document-level RE, Zhou et al. [138] proposed using adaptive thresholds to extract all relations of a given entity pair. Zhang et al. [132] developed the DocUNET model to reformulate document-level RE as a semantic segmentation task and used a U-shaped convolutional neural network architecture to implicitly learn the interdependency among the multiple triples in one context. Tan et al. [94] proposed the use of knowledge distillation and focal loss to denoise the distantly supervised data for DocRE. Wang et al. [109] proposed a positive-unlabeled learning algorithm under incomplete annotation scenario. However, these methods typically require lots of manually labeled data for model training, while in practice, these labeled data would be labor-intensive to obtain and error-prone due to human subjective judgments.

Two directions of work have been explored in the literature to alleviate human supervision in relation extraction: *Pipeline-based approach* and *Learning-based approach*. *Pipeline-based approach* [3, 23] combines different information extraction techniques (like OpenIE [2] and Semantic Role Labeling[30]) trained on *open domain* and uses a domain-specific dictionary to extract relations in the given domain. These methods do not require any domain-specific labeled data, the only human supervision needed is the build of the domain-specific dictionary and pre-set patterns. However, pipeline-based approaches are usually prone to error propagation.

For *Learning-based approach*, researchers have developed methods that learn from unlabeled data in order to alleviate human supervision. Traditionally researchers have proposed pseudo-labeling for relation extraction, e.g., distant supervision. Distant supervision [71] leverages external knowledge bases to obtain annotated triplets for supervision. These methods make a strong assumption that the relation between entity pairs should not depend on the context, which leads to context-agnostic label noises and sparse matching results. To solve this problem, self-ensembling [74] has been proposed to filter noisy examples by distant supervision. Self-ensembling methods assume that predictions on the unlabeled data by the model should remain unchanged, even if there are perturbations in the model parameters or training data. Self-ensembling methods usually suffer from insufficient supervision. When labeled data is limited, the model typically fails to acquire new relation knowledge that could be learned from the unlabeled data, thus impeding further improvements.

On the other hand, since a large amount of in-domain text is usually accessible, we can tackle relation extraction in a semi-supervised learning manner, i.e., there is a small number of labeled data and a large number of unlabeled text in the same domain. Recently, self-training (ST) has also been proposed in semi-supervised relation extraction [42, 129]. ST improves the predictive ability of the model by obtaining high-confidence labels from unlabeled data incrementally and retraining the model on the updated labeled data. However, using self-training directly may introduce noisy pseudo labels inevitably, which hurts the model performance, known as the *gradual drift problem*. To solve this problem, Hu et al. [42] adopts meta-learning to reduce the influence of noisy pseudo labels. Yu et al. [129] treat the ambiguous instances as partially-labeled instances. All these works [42, 129] focus on *sentence-level RE*. However, it is still unclear whether self-training would address challenges for *document-level RE*. First, the quality of pseudo-labels for document-level RE is lower, which might negatively influence the result of self-training. Compared to sentence-level RE, the input of document-level RE contains a more complicated structure and the text span between two given entities is usually longer. This makes document-level RE a more challenging task, and thus the quality of base model (e.g., F1 score) is worse. Self-training on pseudo labels predicted by these base models are generally more noisy. Besides, the above mentioned ST methods do not solve the challenge of *inadequate training data for rare relations* (See Sec. 1.3). More detailed review of self-training method is shown in Section 1.2.

Recently large language models (LLMs) like ChatGPT [1] or GPT4 [76] have raised people's attention with their emergent ability [113] to learn from a few examples in the context, which is so-called in-context learning (ICL). The key idea of in-context learning is to learn from analogy to examples in a given prompt. Different from supervised learning requiring a training stage that uses backward gradients to update model parameters, ICL does not conduct parameter updates and directly performs predictions on the pre-trained language models. Few-shot ICL and zero-shot performance of LLMs have been proven to achieve comparable performance with supervised fine-tuning on smaller models on numerous tasks like mathematical reasoning [114].

Wadhwa et al. [103], Li et al. [53], and Xu et al. [120] evaluated the performance of LLMs in *sentence-level* relation extraction tasks. Based on their findings, in-context learning on LLMs can achieve comparable performance for RE with tuning relatively small PLMs in sentence-level relation extraction. Besides, Xu et al. [120] finds that combining data generated from LLMs with original training data can yield better RE performance than from traditional data augmentation approaches. Li et al. [53] proposed a majority-vote-based method to aggregate the answers based on questions in different forms, improving the zero-shot performance of LLMs on sentence-level relation extraction.

Since the input of document-level RE has a long length, ICL for document-level RE can be challenging, since the fixed input length of LLM can only contain few ICL examples. According to Ozyurt et al. [77], the F1 score of ICL on GPT-JT[2] (6B parameters) is only 35.09 on DocRED dataset, which is much lower than ATLOP[138] with supervised-training on RoBERTa(base)[63] model (63.40 F1 with around 125M parameters).

Here is an example of zero-shot performance of document-level relation extraction with ChatGPT in the biomedical domain. (Fig 1.2). We can find out that ChatGPT has the following issues: 1) ChatGPT treats relation extraction in an *OpenIE* manner, which is usually in a complicated form and does not reveal the relations between two entities. For example, in *(eradication rate with standard triple therapy, was unsatisfactory, probably because of widespread bacterial resistance due to unrestricted antibiotic use)*, the subject and object contain many modifiers and the object is not an entity. 2) ChatGPT sometimes cannot figure out the relation between two

[1] `https://openai.com/blog/chatgpt`

[2] `https://huggingface.co/togethercomputer/GPT-JT-6B-v1`

biomedical concepts (H. pylori and widespread bacterial resistance), no matter what kind of prompt is given (predicting relation given subject and object; predicting object given subject and relation; predicting subject given object and relation).

In summary, currently zero-shot and few-shot learning for LLM-based method cannot achieve comparable performance to supervised fine-tuning on smaller encoder-based models (e.g., BERT[17] and RoBERTaLiu et al. [63]).

1.2 Overview of Self-Training

Figure 1.3 depicts the paradigm of standard self-training. First, a supervised learning algorithm is trained based on the labeled data only. This classifier is then applied to the unlabeled data to generate more labeled examples as input for the supervised learning algorithm. In practice, usually, only the labels the classifier is most confident (Top-K pseudo data) in are added at each step. Self-training method is first proposed in Scudder [91]. In the field of natural language processing, some early works have applied self-training to word sense disambiguation [127] and parsing [86]. Recently, Self-training has flourished again by iteratively generating pseudo labels and augmenting the tuning of data-hungry pre-trained language models, showing great advantages in further enhancing natural language understanding (NLU) [4, 10, 21, 70, 102] and Neural Machine Translation (NMT) [37, 46] where massive unlabeled input text exists. Vu et al. [102] finds out that using a better base model leads to better self-training results. Besides classical ST, diverse follow-up modifications have been developed for further improvement, which generally fall into two lines.

Sample Selection These methods select only a part of unlabeled instances, given that there are works that leverage full unlabeled data [102]. The sample selection methods can be grouped into the following three categories.

- **Confidence score**. These methods model confidence to avoid over-noisy pseudo labels [4, 93]. Usually, the model confidence is computed by the output of the final softmax layer [4].
- **Uncertainty**. These methods obtain informative instances and enhance performance on the hard ones [46, 73]. Jiao et al. [46] find out that NMT models benefit more from uncertain monolingual sentences for self-training, since easy patterns in monolingual sentences with

deterministic translations may not provide additional gains over the self-training teacher model.

- **Label balance**. Wei et al. [111] re-samples the pseudo-labels generated by models based on the frequencies of the training samples to benefit minority classes.

Noisy Labeling These methods [37, 117] inject synthetic noise into the pseudo data, *e.g.*, token shuffle or image distortion to propagate local smoothness and improve model robustness. Specifically, He et al. [37] shows that Monte-Carlo dropout is a crucial ingredient to prevent self-training from falling into the same local optimum as the base model, which is responsible for the gains in neural machine translation tasks. In this way, Self-training with noisy data can be viewed as an extension of data augmentation [85], where modified copies of a dataset are artificially increased in the training set. Vu et al. [102] also finds out that strong base models benefit from including even significantly noisy pseudo-labels in self-training.

There are also theoretical analyses about self-training. Wei et al. [112] prove that the minimizers of population objectives based on self-training and input-consistency regularization will achieve high accuracy with respect to ground-truth labels, under some specific assumptions. Frei et al. [27] compute a bound on the classification error of Bayes-optimal classifier with gradient descent algorithm. Such theoretical guarantees ensure the general soundness of self-training algorithm.

As mentioned in Section 1.1, Self-training has also shown superiority in the task of sentence-level Relation Extraction [41, 129] and document-level Relation Extraction [96]. I will state the challenges of current self-training methods in Section 1.3.

1.3 Motivation and Research Questions

As discussed in the Introduction (Chapter 1), it is expensive to label a large amount of training data for document-level Relation Extraction. According to Tan et al. [95], suppose there are N entities in one document and the label space of interest contains R relations. The decision space for human annotation is $N \times (N-1) \times R$. In particular, for an average case of DocRED [126] ($N = 20$, $R = 96$), an annotator will need to make 36,480 classification decisions for one document. As a result, it takes around 30 minutes by experts to annotate one document. The cost of labeling motivates *unsupervised approaches* for relation extraction.

Unsupervised approaches may rely on distance constraints to generate candidate relation triples [5] or normal forms of the dependency parse trees [45]. However, these dependency-based methods cannot identify the semantic roles of arguments in an extraction, which is essential for high-quality relation extraction.

Meanwhile, Semantic Role Labeling (SRL) [30] aims at detecting argument structures associated with verb predicates and labeling their semantic roles. Semantic roles make it possible to impose structural and semantic constraints on entity types to ensure high-quality extractions. Based on that, SRL-based approaches [3, 23] built a knowledge graph in the financial domain based on semantic role labeling (SRL) with dictionary-guided predicate filtering. The knowledge bases built in Benetka et al. [3] used a pipeline that consisted of named entity recognition (NER), semantic role labeling (SRL), and date/money parser. Further, Elhammadi et al. [23] combined SRL and pattern-based information extraction to extract domain-targeted noun/verb-mediated relations. However, it is still unknown whether these methods could be extended to other domains and languages. This motivates our first research question (**RQ**).

<u>RQ1 (Unsupervised RE):</u> Can these pipeline-based approaches extend to other domains and languages? Further, can we improve the recall of pipeline-based approaches while keeping high precision?

To answer RQ1, I propose CIFRE (Chapter 2) to another Chinese language. I improve the recall of these pipeline-based approaches by detecting missing entities and resolving Chinese idioms via back-translation. Details of our contributions are discussed in Section 1.4.1.

Another line of work focuses on *semi-supervised* methods, i.e., we have a small number of labeled data and a large number of unlabeled data. In order to get enough labeled data for training, numerous methods have been proposed to generate pseudo-labeled data, including distant supervision and self-training.

Distant supervision aligns triples in a related knowledge graph with the sentences in the input text, in order to automatically generate training data. Distant supervision assumes the responsibility to determine which sentence supports which relation and to what degree it expresses the relation of interest. In other words, distant supervision labels sentences with appropriate relations, and generates an error-prone training set consisting of possibly wrongly-labeled instances, which in turn is used to train relation extraction models.

Self-training methods [42, 129] have been proposed to generate pseudo-labeled data in the semi-supervised setting. Hu et al. [42] propose a semi-supervised learning framework that adopts meta-learning during pseudo label generation and automatically learns to reduce the influence of noisy pseudo labels. Yu et al. [129] further focuses on ambiguous instances, treats the ambiguous instances as partially-labeled instances, and applies set-negative training for them [123].

Although previous works have shown that self-training could improve the results of sentence-level relation extraction [42, 129], it is still unclear if it could work on document-level relation extraction. A significant challenge of ST is *inadequate training data for long-tail relations*. A relation class is considered as long-tail if it consists of items with low frequency or occurrence in the training set compared to other classes. Understanding and properly handling long-tail classes is important where imbalanced datasets can lead to biased models or poor performance on minority classes. In Re-DocRED dataset, the frequent classes include the top 10 most popular relation types which cover around 60% of the training set triples[95]. The other 85 classes are categorized as long-tail classes which cover around 40% of the training set triples. As shown in Fig. 1.4, current document-level RE systems [138] do not perform well on long-tail relations, which hardly appear in the training data. For example, the F1-score for class *located in* is 83.02 while *ethnic group* is only 6.45. The reason could be that the number of appearance in training data (Re-DocRED[95]) is vastly different (20k vs. 155). Assuming that training data and unlabeled data have the same distribution, we cannot expect these long-tail relations to appear sufficiently often in the unlabeled text corpus. To address this, Tan et al. [96] re-sample training set and assign more weight to the classes that have high precision and low recall. However, this method does not bring new information to the relation classifier. As a result, these self-training methods might not be able to improve the RE performance on these rare relations.

On the other hand, self-training (ST) could be applied in both relation classification side and text generation side. Traditionally ST benefits from a vast number of unlabeled instances and extends the generalization bound [112, 133]. Consequently, classical ST only works for a few generation tasks with adequate plain text, like Sequence Labeling [108] and Machine Translation [37, 46]. Nonetheless, it is unresolved how to incorporate self-training into the data-intensive attribute-controllable Natural Language Generation (NLG), i.e., generate a textual sequence satisfying the input attribute label, as opposed to NLU. Since model inputs now are discrete labels, massive high-quality unlabeled target text (*e.g.*, movie reviews for sentiment-controllable NLG) is essential to construct pseudo label-text pairs, which is impractical in low-resource domains, impeding the broad application of ST [21]. This brings three new challenges for ST, which motivates our second research question.

Challenge 1: Limited unlabeled data: since model inputs become discrete labels, there is no massive unlabeled data for the NLG model to extend the learned distribution boundary. With limited unlabeled text, a potential approach to further improve ST performance is to leverage the generative ability of NLG models and produce synthetic (pseudo) text [90, 125] from given labels besides pseudo labels from text. In this case, unfortunately, two other major challenges arise.

Challenge 2: Over-exploitation. Augmented by self-generated text, NLG models are forced to repeatedly fit the already learned text distribution. This gradually homogenizes the generated pseudo text and causes a shrunken (collapsed) generalization boundary, resulting in decreased controllability and generation diversity.

Challenge 3: Training deceleration. We need to re-generate all pseudo text in each ST iteration with updated model parameters, which interrupts the parallelism of Transformer [101]-based models, severely decelerating training and impairing practicality.

RQ2 (Text Generation): How can we improve the controllability and diversity of text generation? Specifically, how do we overcome the three challenges (inadequate unlabeled data, over-exploitation, and training deceleration) for self-training in controllable NLG?

To answer RQ2, I proposed DuNST (Section 3.3) and KEST (Section 3.4) to overcome the proposed three challenges. DuNST aims to tackle Challenge 1 by leveraging both pseudo-labeled data and pseudo text and to tackle Challenge 2 by adding noise to encourage model exploration. KEST tackles Challenge 2 by applying a kernel-based loss function which encourages more diversity and tackles Challenge 3 by applying Non-auto Regressive generation to accelerate text decoding speed. Details of our contributions are discussed in Section 1.4.1.

The challenge of *inadequate training data for long-tail relations* in the task of RE motivates our third research question:

RQ3 (Semi-supervised RE): How can we improve the performance (e.g., F1-score) of semi-supervised RE, especially for long-tail relations?

The answer of RQ2 helps to answer RQ3. With the improvement of controllability and diversity of the text generator, I propose a novel method DuRE (Chapter 4). Unlike previous ST methods [42, 96], I simultaneously train a controllable text generator and relation classifier. The text generator aims to generate diverse outputs given specific relation triples. I also apply contrastive loss and self-adaptive sampling to improve the quality of self-training. Details of our contributions are discussed in Section 1.4.1.

1.4 Contribution and Outline

1.4.1 Solution to the Research Questions

Motivated by the large cost of constructing supervised training data, the thesis seeks to improve the performance of *unsupervised* and *semi-supervised* approaches for document-level relation extraction and addresses specific challenges mentioned in Section 1.3. Figure 1.5 describes the overall framework of this thesis.

This thesis takes the first step to extend the pipeline-based approaches to another domain, answering the two questions of **RQ1(Unsupervised RE)**. The answer to RQ1 is that *pipeline-based approach can be potentially extended to other languages. We can improve the recall of pipeline-based approach with missing entity recognition and idiom resolution.* We present our system, ChInese Financial Relation Extraction (CIFRE) which consists of a modular pipeline, supporting large-scale high-quality KG construction. We are the first to leverage SRL for Chinese RE. We semi-automatically construct a dictionary of semantically and structurally constrained financial predicates in order to retain high-quality and financially relevant SRL extractions. We design patterns specific to Chinese, by leveraging grammatical structures unique to the language, for pattern-based extraction. I improve the recall of pattern-based extraction in Chinese by considering coordinated (COO) relations. Further, I detect and complete extractions with missing entities and resolve Chinese idioms via back-translation to generate additional extractions. Finally, I demonstrate the performance of our pipeline on 2 corpora of Chinese financial news, compared to a suite of baselines.

Then this thesis tries to tackle the three challenges listed in RQ2, since the answer to RQ2 (text generation) contributes to the answer to RQ3 (semi-supervised RE). The general answer to RQ2 is as follows: *Self-Training with dual objective (both classification and generation) is able to improve the controllability and diversity of text generation, with specific strategies to address the issue of over-exploitation (adding noise or kernel-based loss) and training deceleration (non-autoregressive generation).*

To handle *Challenge 1* and *Challenge 2*, I propose a novel **Du**al **N**oisy **S**elf **T**raining (**DuNST**) (Section 3.3, for semi-supervised controllable NLG. DuNST jointly learns to generate text from given attribute labels and predict labels for text, characterizing these two directions as a dual variational generation process. Such duality allows our model to leverage not only generated pseudo text but also pseudo labels predicted for available unlabeled text. Both generation and classification would be augmented by the two kinds of pseudo data and thus gradually refined during the ST process, handling *Challenge 1*. Besides, DuNST incorporates two new types of flexible noise into generated pseudo text, namely softmax temperature and soft pseudo text, to further perturb and escape from the text space learned at the previous ST iteration, which helps propagate local smoothness and enhance robustness [10, 117], addressing *Challenge 2*. Our method can be theoretically regarded as exploring a larger potential space, thus facilitating an extended generalization boundary and improved attribute coverage, balancing exploration and exploitation better. Hence, DuNST could further boost controllability while maintaining comparable generation fluency and diversity.

Further, to tackle *Challenge 2* and *Challenge 3* and address the limitation of DuNST, I propose another novel self-training framework, **K**ernel Distance Based **E**fficient **S**elf **T**raining (**KEST**) (Section 3.4), for improving semi-supervised controllable NLG. Instead of learning from generated pseudo textual sequences with traditional cross-entropy loss, KEST directly fits the approximated text distribution from the last iteration in the embedding space. Such an objective not only relaxes the constraint imposed by the previous ST iteration but also encourages diverse outputs of the current model, addressing *Challenge 2*. Besides, I design a non-autoregressive generation schema to produce soft representations of pseudo text (rather than hard strings) in parallel, greatly reducing time cost and handling *Challenge 3*. Furthermore, such a soft text is naturally a kind of noisy pseudo data [37, 117], which helps the model denoise errors and propagate local smoothness [10, 112].

After addressing the three challenges proposed in RQ2, this thesis finally provides the answer to RQ3. The answer to RQ3 is that *Dual Self-training on both classification and generation can improve the F1 score in semi-supervised document level RE, especially for long-tail relations.* I propose a novel method – Dual contrastive self-training for semi-supervised Relation Extraction (DuRE). Unlike previous ST methods [42, 96], I simultaneously train a controllable text generator, generating diverse outputs given specific relation triples. To improve the controllability of the generator, I leverage the signal of the trained RE classifier to label positive and negative generated sequences, and then apply a ranking calibration loss [137] to contrast the positive and negative sequences to improve generation quality. In addition, I propose a self-adaptive way to sample pseudo text from different relation classes. I add noise by increasing generation temperature for relations with higher precision, which introduces diversity to the training set and helps reduce overfitting. Besides, I sample more examples from relations with lower recall. Since long-tail relations usually have a low recall (Fig. 1.4), they are more likely to be sampled, and thus their recall can be increased through training.

In summary, the main contributions of the thesis are outlined below.

1.4.2 Thesis Outline and contributions

The remainder of this thesis is structured around the three research questions proposed in section 1.3.

Chapter 2 focuses on the first research question about unsupervised methods for relation extraction (RQ1). To answer RQ1, I propose to improve the recall of pipeline-based approaches by detecting missing entities and resolving Chinese idioms via back-translation. The key contributions of this chapter are as follows.

- Section 2.3 discusses the modular pipeline I proposed for Chinese financial relation extraction. It is an unsupervised method without labeled data in training. I semi-automatically construct a dictionary of semantically and structurally constrained financial predicates in order to retain high-quality and financially relevant SRL extractions. I design patterns specific to Chinese, by leveraging grammatical structures unique to the language, for pattern-based extraction. I improve pattern-based extraction in Chinese by considering coordinated relations.

- Section 2.4 points out a common issue with SRL-based approach. I present a distance-based solution for the MER task to improve recall of the pipeline.

- Section 2.5 points out a language-specific issue when extracting relation tripes in Chinese. I propose to use back-translation to resolve Chinese idioms to generate additional extractions and thus improve recall of the pipeline.

- The proposed CIFRE pipeline is evaluated with other baselines on two Chinese financial news corpora in Section 2.6. The effectiveness of two novel components (Missing Entity Recognition and Chinese Idiom Resolution) is also discussed in this section.

In Chapter 3, I attempt to tackle the second research question. For RQ2, I propose two self-training-based methods (DuNST (Sec 3.3) and KEST (Sec

3.4)) to improve the controllability and diversity for controllable natural language generation, which could be further applied in generating pseudo texts given specific relation triples for semi-supervised relation extraction. The key contributions of this chapter are as follows.

- Section 3.3.2 proposes the Dual Noisy Self-Training (DuNST) algorithm. To the best of our knowledge, we are the first to incorporate Self-training into semi-supervised controllable language generation and propose a novel and effective ST method.
- Section 3.3.2 proposes two new types of flexible noise into generated pseudo text, namely softmax temperature and soft pseudo text, to further perturb and escape from the text space learned at the previous ST iteration, which helps propagate local smoothness and enhance robustness.
- In Section 3.3.2, I demonstrate that DuNST explores a larger potential text space and extends the generalization boundary, providing a theoretical interpretation for our method.
- In Section 3.3.4, I conduct thorough experiments on three attribute-controllable generation tasks and manifest the superiority of DuNST in improving control accuracy with competitive quality of the generated text, further exploiting the capacity of powerful PLMs for NLG.
- In Section 3.4.2 I dig into the over-exploitation problem of applying self-training to controllable NLG and propose a novel kernel-based ST framework to address this problem.
- In Section 3.4.2 I design a non-autoregressive generation schema to reduce the time cost of producing pseudo text for self-training, making ST more practical for real scenarios.
- In Section 3.4.2 I theoretically show that KEST can explore a larger potential text space.
- In Section 3.4.3 I demonstrate through exhaustive experiments that our model significantly improves controllability with competitive generation diversity and quality, further exploring the capacity frontier of PLMs.

In Chapter 4, I attempt to tackle the third research question. To answer RQ3, I develop a semi-supervised learning framework in document-level relation extraction, leveraging both pseudo data and pseudo text. The key contributions of this chapter are as follows.

- In Section 4.3, I dig into the problem of document-level extraction of long-tail relations and propose to simultaneously train a controllable text generator to address the limitation of previous self-training methods [96] that only leverage pseudo-labeling.

- In Section 4.3.3, I propose a contrastive loss to control the quality of generated pseudo text, improving the generation quality and thus helping to enhance the classification performance of the relation classifier.

- In Section 4.3.3, I conduct comprehensive experiments to show that our model significantly improves F1-score in different RE benchmarks on general and biomedical domains, especially on long-tail relations, addressing the proposed challenge.

In Chapter 5, I summarize the works presented in this thesis and then discuss some of the several interesting future directions about relation extraction and self-training. The challenges that remain to be solved are highlighted and a brief hypothesis of how future research can address those challenges is presented.

Document:
The prevalence of *H. pylori* infection in dyspeptic patients in Yemen is very high, the eradication rate with standard triple therapy was unsatisfactory probably because of *widespread bacterial resistance* due to unrestricted antibiotic use.

ChatGPT results:	**Golden Label not returned by ChatGPT:**
(prevalence of H. pylori infection, is very high in, dyspeptic patients in Yemen) (eradication rate with standard triple therapy, was unsatisfactory, probably because of widespread bacterial resistance due to unrestricted antibiotic use)	(H. pylori, **exhibits**, widespread bacterial resistance)

Questions for ChatGPT:

Q: What is the relationship between H. pylori and widespread bacterial resistance?

A: The given text does not provide a direct relationship between H. pylori and widespread bacterial resistance. However, it suggests that the eradication rate of H. pylori with standard triple therapy was unsatisfactory, probably because of widespread bacterial resistance due to unrestricted antibiotic use. This means that the frequent and unrestricted use of antibiotics might have led to the development of bacterial resistance, which could have contributed to the reduced effectiveness of standard triple therapy in treating H. pylori infection.

Q: What bacterium exhibits widespread bacterial resistance according to the text?

A: The given text does not explicitly mention any bacterium exhibiting widespread bacterial resistance. However, it suggests that the eradication rate of H. pylori infection with standard triple therapy was unsatisfactory, probably because of widespread bacterial resistance due to unrestricted antibiotic use. This means that the frequent and unrestricted use of antibiotics might have led to the development of bacterial resistance, which could have contributed to the reduced effectiveness of standard triple therapy in treating H. pylori infection.

Q: What do H. pylori exhibit according to the text?

A: According to the given text, H. pylori infection exhibits a high prevalence in dyspeptic patients in Yemen. Dyspepsia is a medical condition characterized by upper abdominal pain or discomfort, bloating, and early satiety. Furthermore, the text suggests that the eradication rate of H. pylori infection with standard triple therapy was unsatisfactory, probably because of widespread bacterial resistance due to unrestricted antibiotic use. This means that H. pylori infection might have developed resistance to antibiotics due to their frequent and unrestricted use.

Figure 1.2: An example of document-level relation extraction in Bacteria Biotope at BioNLP Open Shared Tasks [6].

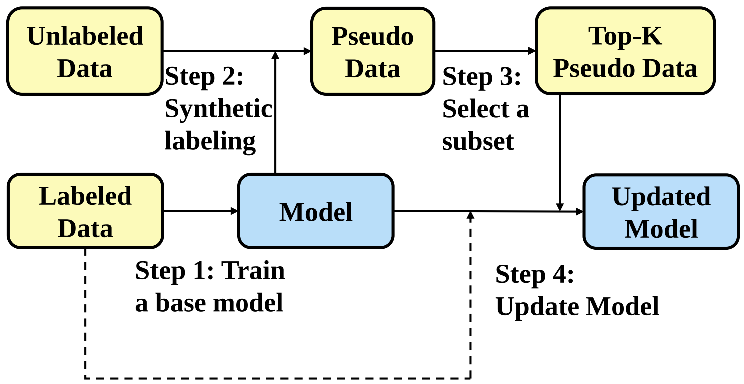


Figure 1.3: Classic Self-training. ST trains a base classification model on a small labeled data set, then iteratively predicts pseudo labels for unlabeled data to augment the original set and finally fits the model to the augmented training set.

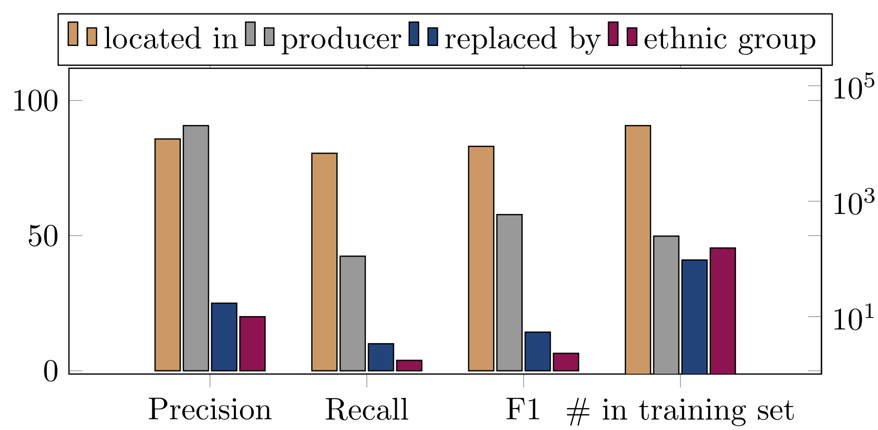


Figure 1.4: Testing set performance (Precision, Recall, F1, and # of examples in training set) of baseline ATLOP [138] trained on Re-DocRED [95] dataset.

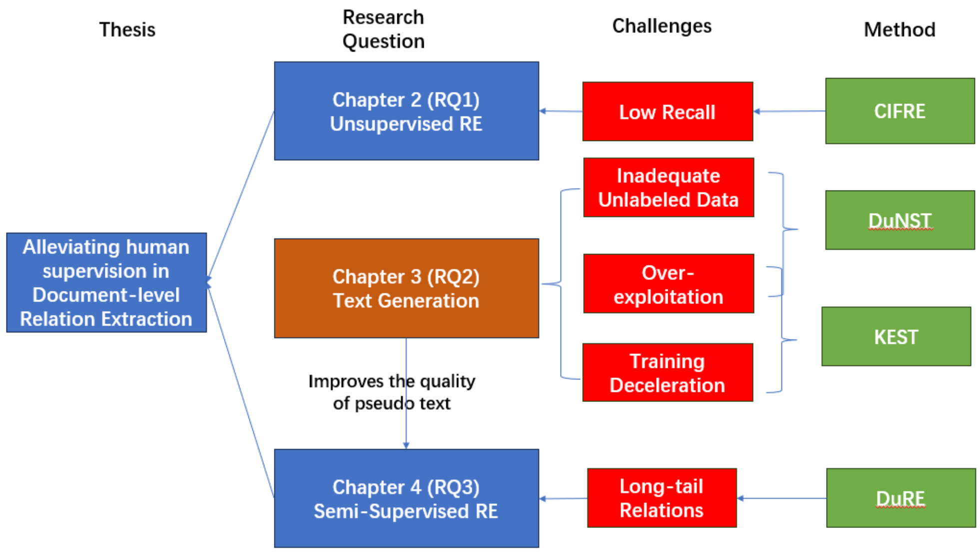


Figure 1.5: The overview of this thesis with corresponding concepts. The left column is the main theme of the thesis, which is supported by the research questions in the middle. The middle columns are organized by chapters to each research question and challenge. The right column describes the methods designed to answer these research questions.

Chapter 2

Unsupervised Document-Level Relation Extraction

A version of this chapter is to be submitted. I was the main investigator. Throughout the project, I led the process of defining project goals and key research questions, the model implementation, and designing and running the experiments. The work was done under the supervision of Laks V.S. Lakshmanan.

2.1 Introduction

Driven by applications such as fact checking [15], question answering [82, 121], semantic search [118], and recommendations [105], recent years have witnessed a surge of interest in the automatic construction of knowledge graphs (KG) [19, 51, 75]. Relation Extraction (RE) from unstructured data sources is a key component for building large-scale KG. The recent availability of large corpora of Chinese financial news articles buttresses the motivation for financial relation extraction in Chinese. RE in Chinese has traditionally been tackled with both supervised and unsupervised approaches. However, supervised Chinese RE relies on a pre-specified ontology of relations and large amounts of human-annotated data in training. As such, the scalability and transferability of supervised approaches to *new* relations are limited.

Previous unsupervised approaches may rely on distance constraints to generate candidate relation triples [5] or normal forms of dependency parse trees [45]. However, these dependency-based methods cannot identify the semantic roles of arguments in an extraction, which is essential for high-quality relation extraction. Meanwhile, Semantic Role Labeling (SRL) [30] aims at detecting argument structures associated with verb predicates and labeling their semantic roles. Semantic roles make it possible to impose structural and semantic constraints on entity types to ensure high-quality extractions.

Previous RE approaches restrict the subject and object to be entities, while SRL operates in a "best-effort" manner, thus allowing for incomplete extractions containing missing entities. As a result, there is an opportunity to complete such extractions with the corresponding missing entities to generate a valid extraction. Consider the sentence "据**美国彭博社**报道，由于进行减员，**德国**最大的零售商**麦德龙股份公司**宣布，二季度利润意外下降15%。(*According to* ***Bloomberg News*** *in* ***the United States****, due to job cuts,* ***Metro AG****,* ***Germany****'s largest retailer, announced that profit in the second quarter fell unexpectedly by 15%.*)"[3]. The SRL system outputs the incomplete extraction *fall:{A0: profit, A1: 15%}* where the relevant named entity is not extracted, e.g., "whose" profit. A valid extraction with the correct entity would be *fall:{A0:* ***Metro AG****'s profit, A1: 15%}*. Notice that in the above example, the missing entity resides in the same sentence. There also exist cases where the missing entity is located in previous sentences. In the context of the financial domain, we are especially interested in financially related nouns such as *profit* and *sale*. These financially related nouns can act as the subject in a sentence on their own. As a result, in some cases the SRL system fails to extract the entity being referenced. In sum, identifying such missing entities has the opportunity to improve the recall of a relation extraction system. Our goal is to address the missing entity problem while maintaining comparable precision.

In this work, we develop a Chinese language knowledge extraction pipeline tailored to the financial news domain by combining SRL information extraction with patterns derived from Chinese language grammar structures. The pipeline is guided by a dictionary of semantically and structurally constrained financial predicates. We semi-automatically construct the dictionary by mining financially relevant predicates from a Chinese common-sense knowledge base Hownet [20]. In contrast to supervised learning, our method can seamlessly generalize to new relations by augmenting the dictionary with new relation specifications, avoiding the labor intensive task of generating labelled data for training new relations. Further, we formulate the novel Missing Entity Recognition (MER) task, and apply a dependency-based system to complete SRL extractions with missing entities. Moreover, we overcome a challenge unique to the Chinese language in which idioms are commonly used in news reporting. Specifically, idioms pose a problem for unsupervised RE as they can imply a relationship without referring to the corresponding predicate explicitly. Our main observation is that state-of-the-art translation

[3]For convenience, all the named entities are bolded.

systems typically do not produce English→Chinese translations that contain Chinese idioms. Therefore, we perform a meaning-preserving transformation via back-translation from Chinese→English→Chinese that results in idiom-free sentences from which the SRL system can successfully produce extractions.

In sum, our key contributions are as follows. We present our system, ChInese Financial Relation Extraction (CIFRE) which consists of a modular pipeline, supporting large scale high quality KG construction. We are the first to leverage SRL for Chinese RE. We semi-automatically construct a dictionary of semantically and structurally constrained financial predicates in order to retain high quality and financially relevant SRL extractions. We design patterns specific to Chinese, by leveraging grammatical structures unique to the language, for pattern-based extraction. We improve pattern-based extraction in Chinese by considering coordinated relations. Further, we detect and complete extractions with missing entities, and resolve Chinese idioms via back-translation to generate additional extractions. Finally, we demonstrate the performance of our pipeline on 2 corpora of Chinese financial news, compared to a suite of baselines.

Our comprehensive experiments show that CIFRE extracted 22K distinct SRL-based facts (with a top-100 precision of 81%) and 12K distinct pattern-based facts (with a top-100 precision above 91%) from the SmoothNLP corpus (20K articles), and 36K distinct SRL-based facts (with a top-100 precision of 85%) and 15K distinct pattern-based facts (with a top-100 precision above 88%) from the Netease corpus (20K articles).

2.2 Related Work

Automatic knowledge graph construction seeks to build a KG from unstructured text in a specific domain or across multiple domains, without human intervention. Traditionally, an OpenIE system (e.g., Stanford OpenIE [69]) is leveraged to extract cross-domain relations. OpenIE systems [2] are based on dependency parsing and constituency parsing. Recently, there have been approaches that leverage deep learning in extracting relations in the general domain and even in different languages, simultaneously [35, 88]. Self-supervision [41] and Causal models [61] are also applied for OpenIE.

There have also been a number of efforts at creating domain-targeted KGs. Examples include the Aristo Tuple KB [72] in the elementary science domain and CPIE [106] in the biomedical domain. In addition, in the financial domain, Benetka et al. [3] extracted quintuples of monetary transactions

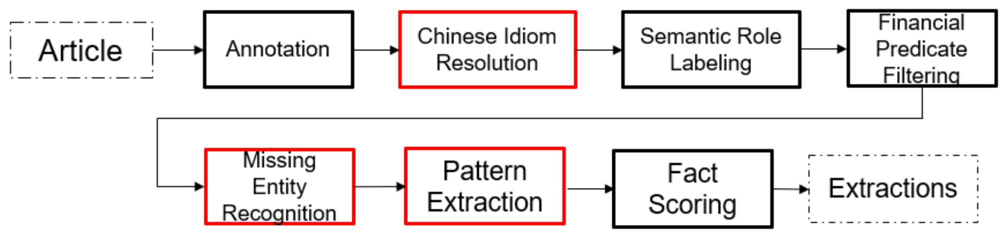


Figure 2.1: Our ChInese Financial Relation Extraction (CIFRE) pipeline

covering economic events. More recently, Elhammadi et al. [23] built a knowledge graph in the financial domain based on SRL and apposition detection. All the above works focus on KG construction from English articles.

Chinese relation extraction is usually handled by models based on supervised machine-learning, e.g., MG-lattice [59] and PGCORE [12]. MG-lattice is trained on three corpora and is targeted at detecting pre-defined latent relationships (where some target predicates are not explicitly stated in the sentence) between two given named entities. Specifically, the entities have to be supplied to MG-lattice as inputs. PGCORE tries to predict subject-predicate-object (SPO) triples using a pointer network. However, these supervised approaches rely heavily on massive manually labeled corpora that are appropriate for the task.

There are also Chinese Open Relation Extraction systems including Tseng et al. [100], UnCORE [5], ZORE [81], and COER [45]. UnCORE exploited distance constraints to generate candidate relation triples ruling out the ability to handle long-term dependencies. COER proposed an unsupervised model based on seven Dependency Semantic Normal Forms (DSNFs). However, as noted in Section 2.1, dependency based methods cannot identify semantic roles of each argument, which is critical for high quality relation extraction.

Recently, CTHE [35] proposed a combined approach incorporating OpenIE and supervised learning in Chinese. If the relation falls into a pre-defined relation set, then they use a supervised learning approach based on BERT [17]; otherwise, they leverage OpenIE to predict a result.

Unlike previous supervised approaches, our CIFRE system does not need labeled data for training. Unlike dependency parsing based systems like COER, UNCORE, and ZORE, our pipeline leverages SRL, leading to more accurate results including the semantic senses for each extraction. We also improve pattern-based extraction in Chinese by considering coordinated relations. Unlike previous SRL-based approaches to financial relation extraction [23] that focus on English, our system is designed for Chinese and includes Chinese-specific patterns (e.g. "chain-of-ATTs") that do not appear in English. Further, absent from Elhammadi et al. [23], we identify and formulate the novel Missing Entity Recognition (MER) task to improve the recall of RE. Finally, to the best of our knowledge, our pipeline is the first that is able to extract relations from Chinese idioms.

2.3 The Knowledge Extraction Pipeline

To begin, we present a high-level description of our CIFRE pipeline (see Fig. 2.1). The inputs to the pipeline include a corpus of articles and two pre-constructed financial dictionaries: one for filtering SRL predicates and the other for identifying valid MER candidates. We have designed a modular architecture whereby different modules can be turned on or off. Compared with Elhammadi et al. [23], who also followed a pipeline-based approach, the modules in red incorporate innovation aimed at generating additional extractions. The pipeline operates at the article level and starts with standard text pre-processing and annotation. Sentences that contain Chinese idioms are passed through a machine translator and translated to English and then translated back to Chinese to get an idiom-free version of the sentence, in order to improve recall. Next, we extract predicate argument structures with the SRL component and use the financial predicate dictionary to filter out noisy extractions and validate structural and semantic constraints on arguments. This step is to ensure the extractions are financially related and to improve precision. After, we detect SRL outputs with missing entities and identify the corresponding entity with a novel solution based on dependency tree distances to complete the extraction, in order to improve recall. Further, we produce additional extractions via high-precision typed patterns (chain-of-noun-ATTs). We maximize the utility of the extractions by minimizing overly specific arguments through argument minimization. Finally, we score the predicate argument structures to reflect our confidence in their precision and conciseness. All language resources we leverage are summarized in Section 2.6.2.

Figure 2.2 shows an example of how our pipeline works. The only inputs to our pipeline are a Chinese News Corpus and a Financial Predicate Dictionary. Our system is able to output both SRL-based and pattern-based extractions.

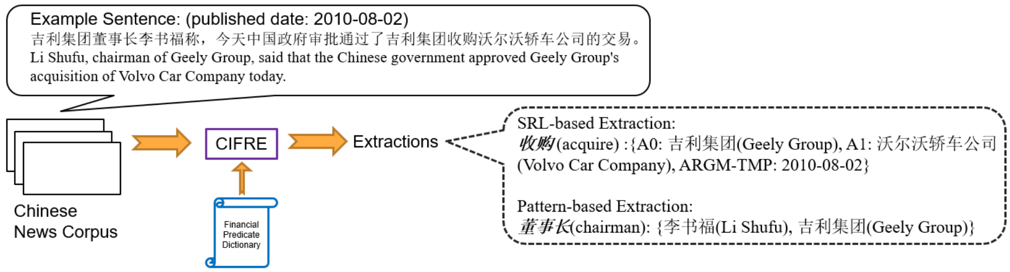


Figure 2.2: Example extractions of our Chinese SRL-based financial Knowledge Extraction (CIFRE) pipeline

Note that our system is also able to parse temporal arguments like "today" to the publishing date (2010-08-02). Finally, the extractions are scored using a binary classifier, with the score reflecting our confidence in the extracted fact.

In the following subsections we describe several components of our pipeline in more detail. The Missing Entity Recognition task is discussed in Section 2.4 and the Chinese Idiom Resolution task is discussed in Section 2.5.

2.3.1 Text Pre-processing and Annotation (TPA)

We start with standard Chinese text pre-processing (e.g., sentence/word segmentation, part-of-speech (POS) tagging, and dependency parsing) using a Chinese NLP toolkit LTP [9]. Next, we detect Named Entities using the NER system provided by LTP. The LTP NER system identifies three types of named entities: people (Nh), organizations (Ni), and locations (Ns).

2.3.2 Semantic Role Labeling (SRL)

We extract semantic relationships between entities using the Semantic Role Labeling system provided by the LTP library. It has been pointed out by Elhammadi et al. [23] that SRL led to strong performance for English financial relation extraction. As described in Section 2.1, information about the semantic roles of arguments of a relation are not captured by traditional dependency-parsing based extractors like COER [45]. In the financial field, temporal information is crucial for meaningful insights. For instance, merely knowing that the BMW group entered the Chinese market (at some time in the past) without knowing precisely when is far less useful. By contrast,

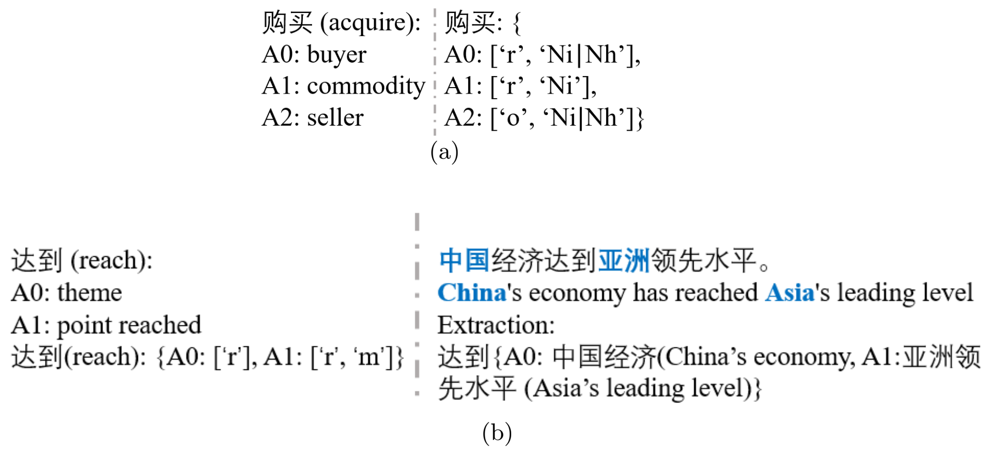


Figure 2.3: (a) Example Chinese Frameset for predicate “购买” (acquire) and an instance in the financial predicate dictionary; “r” is *required* while “o” is *optional*. (b) An example filtered out by predicate filtering. Since the NE type constraint for A1 contains a “m” (number), even if A1 contains another NE “Asia”, it is still filtered out.

SRL not only identifies the correct predicate but also identifies the role of each argument. Correctly identifying the semantic roles of a predicate’s arguments helps us impose *structural* and *semantic* restrictions to improve the precision of the extracted relations. For instance, Figure 2.3(a) shows an example of predicate *acquire*. Specifically, the predicate *acquire* must have at least two arguments, one with the role: *buyer* and the other with the role: *commodity*. We filter out SRL extractions which do not contain a named entity in each argument, in an effort to produce extractions that are well-suited for KG construction.

2.3.3 Financial Predicate Filtering (FPF)

We construct a financial predicate dictionary to ensure the predicates are financially related. Hownet [20] is an online common-sense knowledge base that contains inter-conceptual relationships and inter-attribute relationships of concepts obtained by annotating words in Chinese lexicons with their English equivalents. We leverage the open-source version OpenHowNet [79] to find predicates whose domain is labelled as `finance` or `economy`, which returns 423 candidates. We inspected these candidates and filtered out rare predicates and those that cannot produce a valid extraction due to a lack of arguments. After, we added some common predicates that frequently show up in financial contexts, e.g., "达到" (reach) and "增加" (increase). The resulting dictionary contains 178 predicates.

Following Elhammadi et al. [23], it is vital to identify structural constraints, i.e., whether an argument is required or optional, for each predicate extracted by SRL. It is also important to identify entity type constraints for each argument of an extraction (e.g., whether an argument must contain an organization). In the example from Figure 2.3(a), the type constraints for A0 and A2 in "acquire" are "Ni"(Organization) and "Nh" (People), and the type constraint for A1 is "Ni". A0 and A1 are required, while A2 is optional. Towards this end, we leverage the Chinese Proposition Bank [122] and manually identify the structural and type constraints for each predicate. An example entry in the predicate dictionary is shown in Figure 2.3(a).

We filter out domain irrelevant predicate-argument structures using the financial predicate dictionary constructed previously. The dictionary lists the predicates with the structural constraints of its arguments (`required` vs `optional`), and semantic constraints. The semantic constraints list the valid entity types (e.g., Nh and Ni in Figure 2.3(a) and "m" (i.e., number) in 2.3(b)). Figure 2.3(b) shows an example of an extraction which is filtered out by semantic constraints, since A1 does not contain a required entity of type "number".

2.3.4 Adverb Filtering (AF):

We filter out predicate-argument structures that contain negated arguments since negative statements are unlikely to lead to facts, which tend to be positive. Specifically, we generate a collection of 15 negation words (e.g., 不, 没) and filter out predicate-argument structures with adverbial arguments (ARGM-ADV) that contain any of the negation words. Further, we filter out structures that are in the future tense to guard against erroneously extracting speculative statements as fact. We detect the character "将" (will) in ARGM-ADV and temporal arguments such as "以后" (after) and "未来" (future) in ARGM-TMP. Then, extractions that contain such negation and/or future tense words identified in the ARGM-ADV and/or ARGM-TMP arguments are filtered out.

2.3.5 Temporal Argument Parsing (TAP)

Many of the SRL extracted relations contain temporal arguments ARGM-TMP such as "昨天" *(yesterday)*, "去年" *(last year)*, or "两个月前" *(2*

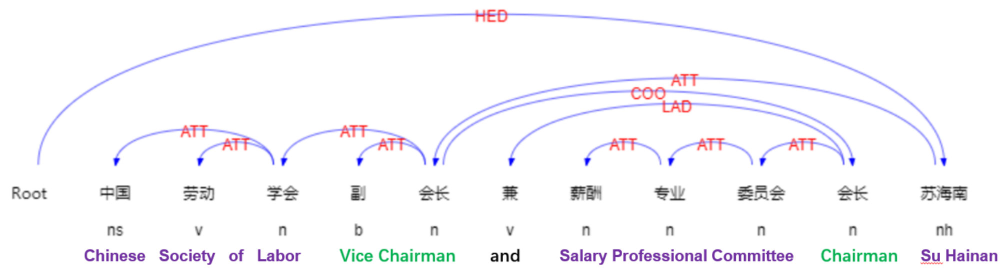


Figure 2.4: Examples of Chain-of-ATT patterns. The words in purple are named entities. The words in green are nouns which have a "COO" relation.

months ago). We pass these arguments to a date parser library[4] that converts localized dates into a standard date format relative to the publication date of the article.

2.3.6 Pattern Extraction (PE)

We observe that there are specific patterns in Chinese that indicate a financially related fact. Specifically, the "Chain-of-ATT's" pattern refers to three or more nouns linked by an attributive (ATT) relation in the dependency parse tree. Figure 2.4 illustrates the chain "学会(Society)$\xrightarrow{ATT}$会长(Chairman)$\xrightarrow{ATT}$苏海南(Su Hainan)". These patterns are similar to the "DSNF1" pattern leveraged in COER [45]. However the following differences exist. First, we add entity type constraints to the first and last items of the chain to ensure the chain connects two named entities. Second, we detect ATT and coordinated (COO) relationships simultaneously to improve recall. Specifically, if two noun phrases, NP_1 and NP_2 have a COO relation, and NP_1 has an ATT relation to an entity NP_3, then NP_2 should also have an ATT relation to NP_3. Consider the example in Figure 2.4: "*Su Hainan, Vice Chairman of the Chinese Society of Labor and Chairman of the Salary Professional Committee*". The COER approach would only extract *(Su Hainan, Vice Chairman, Chinese Society of Labor)*, missing the other relation *(Su Hainan, Chairman, Salary Professional Committee)* since it does not detect the coordinated relationship between *Chairman* and *Vice Chairman*. As a result, our PE module for detecting "chain-of-ATT's" yields high precision extractions through patterns that are commonly found in financial news and specific to Chinese grammar.

Table 2.1 gives 4 types of patterns that we use in pattern detection. The

[4]https://pypi.org/project/dateparser/

Pattern type	**Pattern instance**	**Predicate (subj, obj)**
ORG-PER ATT	耶鲁大学教授陈志武 Yale professor Chen Zhiwu	教授(陈志武,耶鲁大学)
LOC-PER ATT	中国副外长傅莹 Fu Ying, Deputy Foreign Minister of China	副外长(傅莹,中国)
PER-PER ATT	习近平的妻子彭丽媛 Xi Jinping's wife Peng Liyuan	妻子(彭丽媛,习近平)
ORG-ORG Ownership	百度旗下爱奇艺 Baidu's Subsidiary IQiyi	旗下(百度, 爱奇艺)

Table 2.1: Pattern Types, instances and extracted tuples

first 3 patterns do not have restrictions on the relation word. The last pattern (ORG-ORG ownership), restricts the relation word to be "旗下"(subsidiary) to make sure that it correctly indicates ownership between organizations.

2.3.7 Argument Minimization (AM)

We process coordinating conjunctions (e.g., 和 and 与) appearing in an argument to obtain extractions with simpler arguments. We find the entities with coordinated relations by leveraging the dependency relation COO in the sentence's dependency parse tree. Consider the sentence "腾讯收购了搜狗和康盛创想。 *(Tencent acquired Sougou and Comsenz Inc.)*". We split the SRL output *acquire:{A0:Tencent, A1:Sougou and Comsenz Inc.}* into two separate extractions: (1) *acquire:{A0:Tencent, A1:Sougou}* and (2) *acquire:{A0:Tencent, A1:Comsenz Inc.}*. Further, we divide extractions if an adverbial argument is either of "分别(respectively)" or "依次(successively)" and there exist two other arguments connected by the COO dependency relation. Consider the sentence "腾讯和爱奇艺在视频播放器上分别投入三千万和两千万人民币 *(Tencent and IQiyi spent 30m yuan and 20m yuan respectively on video players)*". The resulting extractions are: (1) *"投入(spend): {A0:腾讯(Tencent), A1:30m yuan, A2:视频播放器(video player)}"* and (2) *"投入(spend): {A0:爱奇艺(IQiyi), A1:20m yuan, A2:视频播放器(video player)}"*.

In addition to processing coordinating conjunctions, we minimize the length of arguments even further by identifying and dropping additional tokens that are considered overly specific to shorten the length of an argument, make the extraction concise, and make our extraction of higher utility. We drop tokens that are in parentheses or brackets and remove redundant punctuation. For SRL extractions, we remove the attribute of entities in the arguments. Specifically, we search from the root of the dependency parse tree until we find the first named entity and remove the sub-tree that has an ATT relation connecting to it. For example, consider the SRL extraction "增长*(increase): {A0:*位于广州的富达食品公司营业额*(Fuda Foods Inc. (located in Guangzhou) turnover)，A1:5.7% }*". For the entity "富达食品公司" in A0, we remove the attribute "位于广州的"(located in Guangzhou) while keeping the subsequent text "营业额" (turnover). Thus after argument minimization, the result is the more concise extraction *"*增长*(increase): {A0:*富达食品公司营业额*(Fuda Foods Inc. turnover)，A1:5.7% }"*.

2.3.8 Fact Scoring (FS)

Following previous work in Open RE [23], we score the predicate argument structures to reflect our confidence by training a binary logistic regression classifier using 500 SRL extractions and 500 pattern-based extractions from an independent development corpus that were labeled by us. Facts are considered valid if they are both *precise* and *concise*, i.e., explain only one proposition. We identified a collection of features that are powerful predictors of validity. The features include the presence of a coordinating conjunction or verb, unresolved temporal arguments, pronouns, adjectives, adverbs, and the number of words and named entities in the argument. We classify each valid argument of the extracted fact and take the minimum over all argument scores as the overall confidence score of the fact. We chose the minimum as our aggregate function in order to promote the most precise facts.

2.4 Missing Entity Recognition

In Section 2.1 we identified a limitation of the SRL system, that it is susceptible to extracting partial information in which a named entity can be missing. To address this we formulate the *Missing Entity Recognition* (MER) task, defined as follows. Given an SRL extraction that contains a domain-specific argument with a missing entity, and a list of candidate named-entities: (1) determine if the missing entity is in the list, and (2) if yes, identify the missing entity from among the candidates.

In the financial domain, we require that the missing entity argument must appear in a pre-defined financial noun dictionary. To build this dictionary, we first identify nouns whose label is `economy`, `commercial` or `finance` in the Hownet lexicon. Then, we retain only those nouns that are related to an organization-like entity. The final dictionary contains 179 nouns that can act as standalone subjects or objects in a sentence.

For the generation of candidate named entities, we run the NER system over the k preceding sentences, under the assumption that the entity it refers to must appear before the financial related noun. We choose all named entities of the correct type (e.g., institution) as the candidates. We call these entities *qualified* NEs.

We present the following distance-based solution for the MER task. Our core intuition is that the missing entity for an argument should be the closest qualified entity in the dependency parse tree. Given an SRL argument m with missing entity, for every qualified candidate NE n, if m and n are in the same sentence, the distance $d(m, n)$ is defined as the node distance in the dependency parse tree, $d_{dep}(m, n)$. Otherwise, n and m are in different dependency parse trees. Let the roots of m and n be r_m and r_n, respectively. Then, the distance $d(m, n) = d_{dep}(m, r_m) + d_{dep}(n, r_n) + d_{word}(r_m, r_n)$, where $d_{word}(r_m, r_n)$ is the number of words between r_m and r_n. Finally, we select the qualified NE n^* with the minimum distance to m: $n^* = \arg\min_n d(m, n)$. If two or more candidates have the minimum score, we follow the priority order: 1) prefer entity n acting as a subject; 2) among such entities, prefer n having the minimum word distance $d_{word}(m, n)$. If there are no qualified NEs within the k sentences, we discard the current extraction. In the experiments, we choose k to be 2 since it gives the highest performance.

Using dependency-based distance d_{dep} rather than simple word distance d_{word} can resolve sentences where there are entities appearing in an adverbial or a clause. Consider the following sentence: "恒生银行于港交所公告称，今年中期净利润为69.64亿港元。(***Hang Seng Bank*** *announced on the* ***Hong Kong Stock Exchange*** *that, the mid-year* **net profit** *of this year was 6.964 billion Hong Kong dollars.*)". The dependency-based metric will mark the Chinese word "净利润(net profit)" to *Hang Seng Bank*, rather than *Hong Kong Stock Exchange*, in contrast with choosing the closest entity solely based on word distance.

2.5 Chinese Idiom Parsing

In Chinese lexical system, Chinese idioms are fixed phrases that have been used for a long time and have a fixed structure and complete meaning. One of the design features of most Chinese idioms is the four-character form. Chinese idioms are widely used in writing, even in news reporting. Consider the first example in Table 2.2 containing the idiom "相去甚远". Neither

Sentence	Accurate English Translation	Back-translation	Extraction
但7年以来，日本的通货膨胀率却离2%相去甚远。	But in the past seven years, Japan's inflation rate has been far from 2%.	但在过去七年中，日本通胀率远远低于2%。	低于 (lower than):{A0: 日本通胀率(Japan inflation rate), A1:2%, ARGM-TMP: 在过去七年中(in the past 7 years)}
而深圳交易所的黄金价格也水涨船高，第三季度以来涨幅近10%。	The price of gold on the Shenzhen Stock Exchange has also risen, with an increase of nearly 10% since the third quarter.	自第三季度以来,深圳交易所的黄金价格也大幅上涨近10%。	上涨(rise):{A0: 深圳交易所的黄金价格(The price of gold on the Shenzhen Stock), A1: 近10% (nearly 10%) ARGM-TMP:第三季度以来(since the third quarter) }
上午10点刚过，兴业证券率先揭竿而起，强势上攻。	Just after 10:00 a.m., Industrial Securities took the lead and went up strongly.	就在上午10点钟，公司率先打开柱子并猛烈进攻。	打开(open):{A0: 公司(the company), A1: 柱子(poles)}

Table 2.2: Examples of MBART-50 back-translation. Chinese idioms are underlined. The first two examples lead to a correct relation extraction. The last example is an incorrect back-translation which leads to an incorrect relation extraction. However, the Predicate Dictionary Filtering part in our CIFRE pipeline is able to filter it out.

dependency parsing based approaches, such as COER [45], nor SRL-based approaches are able to extract useful information from this sentence.

A possible solution is to look up the idiom in a lexicon (such as HowNet [20]) and substitute the idiom with its explanation. However, the HowNet explanation for "相去甚远" is "很不同" (differ greatly). Directly substituting this idiom with its explanation results in sentence incoherence. Further, direct substitution does not guarantee that the original meaning will be preserved.

Recently, machine translation has made rapid progress. In particular, to the best of our knowledge, MBART-50 [64] is the state-of-the-art multilingual machine translation algorithm for Chinese-to-English translation. Interestingly, we found that the English-to-Chinese output from MBART rarely contains idioms. In the example above, the idiom "相去甚远" has been replaced in the Chinese back-translation. Thus, the SRL system is able to extract the relation "低于" (lower than):{A0: 日本通胀率(Japan's inflation rate), A1:2%, ARGM-TMP: 在过去七年中(in the past 7 years)}. As shown in the experiment (Section 2.6.5), only 1.6% of the back-translations still contain Chinese idioms. Guided by this observation, we propose leveraging

Chinese-English back translation to resolve Chinese idioms.

Our idiom parsing solution is described as follows. We first determine whether the sentence contains any predicates represented by a Chinese idiom by inspecting the dependency parse tree and detecting the POS tag "i".[5] Then, we substitute the named entities in the original sentences to placeholder strings that will not be changed by the back-translation process. Specifically, we build a one-to-one match for people's names to "{A,B,...}先生" (Person {A,B,...}) and organization names to "{A,B,...}公司" (Company {A,B,...}). Then we pass the sentence to MBART-50, translating the sentence to English (CN-EN) and then translate it back to Chinese (EN-CN). Finally, we substitute the placeholders in the back-translated Chinese sentence with their original entities and pass the back-translated Chinese sentence to the downstream modules in our pipeline.

2.6 Evaluation

2.6.1 Corpora Description

To demonstrate the ability of our CIFRE pipeline to generalize across different datasets, we evaluated it over the following two Chinese financial corpora.

SmoothNLP: Public Chinese Financial News Corpus[6] containing 20K Chinese financial news articles collected from various web sources between January 2015 and May 2019.

NetEase: Financial News Corpus collected from `https://3g.163.com` by the authors. It contains ~20K Chinese financial news articles between May 2010 and March 2015.

2.6.2 Experimental Settings

The language resources we used include the OpenHownet [79] lexicon and Chinese Proposition Bank [122] for financially-related predicate and noun dictionary generation, a Chinese toolkit LTP [9] for basic pre-processing, NER, and SRL, and a machine translation model MBART-50(large-many-to-many) [64]. All the parameters of LTP and MBART-50 are set as default. The corpus we experimented on are SmoothNLP, which is public, and NetEase, which we created ourselves. All public scientific artifacts were consistent

[5] Unlike English, a Chinese idiom is a word that can be tagged with 'i' in POS tagging, which typically consists of four characters. The accuracy of idiom identification is 98% according to experiments on LTP.

[6] `https://github.com/smoothnlp/FinancialDatasets`

Dataset	SmoothNLP	NetEase
# articles processed	20000	19929
# pattern-based facts	15912	19641
# (%) distinct pattern-based facts	12074 (75.9%)	15061 (76.7%)
P@100 for pattern-based facts	88%	91%
# SRL-based facts	22956	37831
# (%) distinct SRL-based facts	21643 (94.3%)	36266 (95.7%)
P@100 for SRL-based facts	81%	85%
# distinct MER-based facts	764	1323
P@100 for MER-based facts	84%	86%

Table 2.3: Statistics and results of running the extraction pipeline on the corpus. P@100 means Precision of top-100 extractions.

Direction	Accuracy
CHS-ENG	66 %
ENG-CHS	48.5%
Back-translation	32%

Table 2.4: Machine Translation quality on sentences containing Chinese idioms.

with their intended use of research. For the corpus we create (NetEase), this can be accessed only for research purposes.

The manual labor is restricted to a semi-automatic method for financial dictionary construction (automatic selection of candidate words from an existing lexicon OpenHownet followed by manual annotation of the required/optional attributes of predicates) and the annotation of 1000 examples for Fact Scoring. The annotation on the resulting 178 predicates is manual and took about 10 person hours. The 1000 examples for Fact Scoring comes from an independent development corpus where any filtering and Argument Minimization are not applied. The annotation took about 5 person hours.

2.6.3 Extraction Results

We ran the CIFRE pipeline on the above two corpora. The evaluation metrics we use are the number of extractions and the precision of the top-k extractions (Precision@k).[7] An extraction is considered a true positive if (i) the relation it describes is accurate and it does not contain additional tokens; and (ii) it describes a fact in the financial domain. To demonstrate the effectiveness of the pipeline, we report the extraction statistics in Table 2.3.

[7]Since it was not feasible to find ground truth in our corpus, instead of recall, we measured the number of extractions.

Sentence	COER extraction	CIFRE extraction
深创投执行总经理刘纲还前往美国与贾跃亭见面。 Liu Gang, Executive General Manager of Shenzhen Capital Group, also went to the United States to meet with Jia Yueting.	[刘纲, 前往, 美国], [刘纲, 见面, 贾跃亭] [Liu Gang, go to, the United States], [Liu Gang, meet, Jia Yueting]	执行总经理:{刘纲,深创投} Executive General Manager:{A0: Liu Gang, A1: Shenzhen Capital Group}
2018年，宝马集团国内销量达到63.995万辆，同比增长7.7%，是宝马集团自1994年正式进入中国市场以来最好的销售记录。 In 2018, the BMW Group's domestic sales reached 639,950 vehicles, a year-on-year increase of 7.7%, which is the best sales record of the BMW Group since it officially entered the Chinese market in 1994.	[宝马集团, 进入, 中国市场] [BMW Group, enter, Chinese market]	达到:{A0: 宝马集团国内销量, A1: 63.995万辆, ARGM-TMP: 2018年} reach:{A0: BMW Group's domestic sales, A1: 639,950 vehicles, ARGM-TMP: in 2018.}

Table 2.5: Some example extractions by COER and our CIFRE. In the first example, COER ignored the "Chain-of-ATT" pattern but extracted predicates "go to" and "meet", which are not financially relevant. In the second example, despite "BMW group enter Chinese market" being informative and financially related, COER still fails to extract important temporal information "in 1994".

We processed 39929 articles in total, and successfully extracted 57909 distinct SRL-based facts and 27135 pattern-based facts. 9% of the pattern-based facts are discovered by the COO relation. We found that 89.7% of the predicate-argument structures that were discarded did not pass the financial predicate filtering step, indicating that the financial dictionary filtering is effective at filtering out financially irrelevant predicates and SRL extractions that do not contain named entities. The MER-based facts increase the number of extractions by 3.5% while even improving the precision slightly (from 84% to 85% overall top-100 precision after adding the MER module). The results show that our new MER module and improvements to PE significantly improve the number of extractions (see Section 2.6.7 for a case study).

2.6.4 Predicate Distribution

Fig. 2.5 shows the distribution of the top 10 financial predicates extracted from the NetEase corpus. The predicate 达到 (reach) is the most frequent

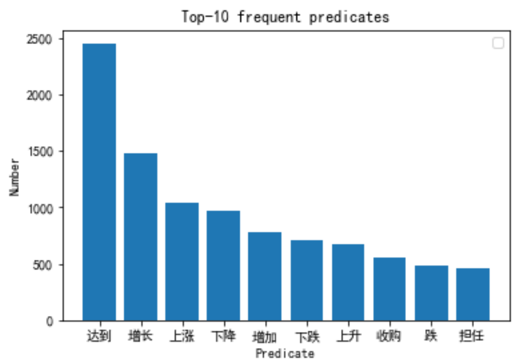


Figure 2.5: Top-10 frequent predicates extracted in Netease corpus

Figure 2.6: An example of failed examples of co-reference resolution by Stanford Corenlp system. Here "its" should refer to "Vitaflo" rather than "Nestle"

predicate. Among these predicates, there are a group of predicates which indicate "ascend or increase" (上升, 上涨, 增加， 增长), and "descend or decrease" (下降, 下跌, 跌) which are usually associated with stocks reporting. The predicate "收购" (acquire) indicates a relation of acquisition between two organizations. The predicate " 担任" (serve as) indicates a relation of people serving in a position in an organization.

2.6.5 Chinese Idiom Parsing Results

In the SmoothNLP and NetEase corpora there are 4598 sentences that contain (i) named entities and (ii) idioms as a predicate (out of 40k articles). We fed these sentences to the MBART-50 [64] model and obtained their back-translations. In the result, only 1.6% of the back-translations still contain Chinese idioms, which verifies our assumption that the MBART-50

Methods	# Distinct Extractions	Precision
CIFRE	161	77.02%
CIFRE+COREF	184	70.11%
CIFRE-MER	150	70.67%
CIFRE-PF	644	24.7%
COER	181	9.9%
MG-lattice	0	N/A

Table 2.6: Method Comparison

model rarely generates Chinese idioms in English to Chinese translation.

We randomly sampled 100 sentences from the back-translated sentences and measured the translation accuracy from Chinese to English. For those sentences that were correctly translated, we then measured the accuracy for its back-translation from English to Chinese. Table 2.4 shows the translation accuracy for both directions.

We found that even for the state-of-the-art MBART-50 model it is relatively difficult to translate English to Chinese. However, interestingly, we are still able to get 84 new extractions with a precision of 74.4% when we pass the back-translated text into our pipeline. We conjecture that our pipeline is able to successfully filter out any bad translations (e.g., nonsensical predicates or missing arguments), even when the overall translation performance is weak. For example, if the back-translation yields an incorrect predicate, our system is able to filter out incorrect relations since they will not pass the predicate dictionary filtering step. In the last example in Table 2.2, the MBART-50 back-translation was incorrect, thus the SRL system outputs a nonsensical extraction: *(the company opens the poles)*. However, since "poles" is not a named entity, it does not pass the predicate dictionary filtering step. Thus, such instances of incorrect translation do not harm the overall precision of our relation extraction.

2.6.6 Comparison with other Relation Extraction Systems

We compare our pipeline with the following variants and baselines.

- CIFRE+COREF: Our pipeline plus Stanford CoreNLP [68] co-reference parser. Resolving pronouns with entities generates additional extractions.
- CIFRE-MER: Our pipeline without MER.
- CIFRE-PF: Our pipeline without predicate dictionary filtering: the

only rule applied is that there must be at least 2 named entities in the SRL extractions.

- COER [45]: A dependency-based Open RE system.
- MG-lattice [59]: A supervised RE system that classifies 44 financial relations given a pair of organizations and a sentence.[8]

We randomly selected 100 articles from the SmoothNLP corpus as a test set. We measured the number of distinct extractions as well as the precision.[9] The precision score computed here gives an *overall* quality of extractions, not only evaluating extractions with high confidence score.

Table 2.6 shows the number of distinct extractions and precision on the test set. We found that activating co-reference resolution, despite generating an additional 23 distinct extractions, only yields 5 additional correct extractions. This is due to error propagation by the Stanford CoreNLP co-reference resolution parser. An example illustrating this module is given in Fig. 2.6. CIFRE+coref extracts "为:{A0:雀巢年度营收(Nestle's annual revenue), A1:4000万瑞士法郎(40 million Swiss francs)}", which is incorrect since A0 should be "Vitaflo's annual revenue". Our final pipeline omits the co-reference resolution module due to the propagation of errors degrading the overall precision.

We found that dropping the MER component leads to fewer extractions and lower precision, thus demonstrating the usefulness of its inclusion in the pipeline. We also found that without predicate dictionary filtering, the SRL system outputs many extractions that are not financially related. COER suffers from a similar issue. Table 2.5 gives 2 examples of extractions by COER and CIFRE. Notice that compared to COER, CIFRE is able to extract more useful relations and the extractions include the time period.

Surprisingly, MG-lattice predicted the "unknown" relation for all inputs considered. While most of the inputs contain relations that are not covered in the FinRE relation set, there are still some overlapping relations between our predicate dictionary and the FinRE relation set, such as "收购"(acquire) and "减持" (share reduction). In the test set, there are 5 extractions by CIFRE that contain such relations. However, all 5 instances are labelled as "unknown" by MG-lattice. Table 2.9 shows a sentence which contains "acquire"

[8] Since the input for MG-lattice is a sentence with two organizations, we applied our NER system and retained those sentences with $m \geq 2$ organizations. We fed the $\mathbf{P}_2^m$ permutations of 2 organizations along with the sentence into the MG-lattice model to get the predicted relations. In total, 1081 org-org-sentence triples were passed to MG-lattice.

[9] The precision here is computed on *all* extractions returned, which is different from the previous evaluation of Precision@k.

as a predicate. This relation is extracted by CIFRE, but not detected by MG-lattice, which is evidence that MG-lattice does not generalize well on our corpus.

2.6.7 MER case study

Table 2.7 shows two successful examples of MER. In the first example, there are three candidate organization-like entities. Our system is able to extract the correct prediction 格力电器 (Gree Electric) because it ignores the entity "QFII" in the adverbial, even if it is closer in word distance. The second example shows that our method is able to correctly predict the entity (*Wenzhou Baiying Real Estate Marketing Planning Co., Ltd.*) which is not in the same sentence as the argument *Heping Zhang*. On the other hand, Table 2.8 shows a failure case where the entity name is only partially extracted "股份"(stock) from "大华股份"(Dahua Co. Ltd.). Most of the incorrect MER predictions are due to mistakes made by the NER system. Specifically, the NER system can output the incorrect name of a named entity or not detect the correct named entity altogether. As a result, our distance-based method may find a different, unrelated entity from previous sentences.

2.7 Conclusion

In this chapter, we described our high-quality relation extraction pipeline CIFRE, which leverages SRL with dictionary-guided filtering and Chinese-specific pattern detection. We introduced the Missing Entity Recognition task and presented a novel distance-based solution that allows our pipeline to detect and complete missing entities in incomplete SRL extractions. Further, we processed Chinese idioms by extracting their implied predicates. We designed CIFRE to be modular allowing individual modules to be updated as the state-of-the-art improves. Comprehensive experiments on two Chinese financial news corpora show that CIFRE achieves the best performance and significantly outperforms the baselines in terms of overall precision. In addition, CIFRE yields 90K extractions in total and achieves a Precision@100 of over 81% on the 2 corpora.

Our system is currently designed for RE on Chinese financial corpora, however, the only ingredient needed to apply our pipeline to other domains is the construction of a domain-specific dictionary. In future work, we plan to investigate ways of extending our pipeline to incorporate textual inferences with which to infer relations when they are not explicitly mentioned in a sentence but are implied (e.g., the latent relations present in the sentences

Sentences	Translation	Entity list	CIFRE extraction	MER prediction
QFII所持有的白色家电则主要由格力电器（000651.SZ）、九阳股份（002242.SZ）构成。格力电器今年以来遭到QFII的持续增持，市值也从当初的8亿元上升至52亿元，市值增加44亿元。	The white goods held by QFII are mainly composed of Gree Electric (000651.SZ) and Joyoung (002242.SZ). Gree Electric's share has been continuously increased by QFII since the beginning of this year, the stock market value also rose from the original 800 million yuan to 5.2 billion yuan, and the market value increased by 4.4 billion yuan.	QFII, 格力电器 (Gree Electric), 九阳股份 (Joyoung)	增加(increase): { A0: 市值 (Market value), A1:44亿元(4.4 billion yuan)}	格力电器(Gree Electric)
温州市百盈房产营销策划有限公司是温州最大的外地房地产代理商。张和平是总经理。	Wenzhou Baiying Real Estate Marketing Planning Co., Ltd. is Wenzhou's largest foreign real estate agent. Zhang Heping is the general manager.	温州市百盈房产营销策划有限公司(Wenzhou Baiying Real Estate Marketing Planning Co., Ltd.)	是(is): A0: 张和平(Zhang Heping), A1:总经理(general manager)	温州市百盈房产营销策划有限公司(Wenzhou Baiying Real Estate Marketing Planning Co., Ltd.)

Table 2.7: Cases of successful MER prediction.

in FinRE). Further, in an effort to improve recall, we plan to upgrade the NER system to detect other classes of financial-related named entities, such as products and financial indices.

One possible drawback of the current back translation method is that it relies on the performance of translation models (e.g., MBART-50). With the improvement of the translation models, there could be more idiomatic translations. Thus it is likely that the back-translated sentences still contain idioms. In future work, we will discuss whether controllable natural language generation methods (e.g., prompt a model with *generate a non-idiom paraphrase of the sentence*) can also remove idioms while keeping the same semantic meaning.

Sentences	Translation	Entity list	CIFRE extraction	MER prediction
以头号重仓股大华股份为例，受益于"智慧城市"建设和安防行业景气提升，作为龙头公司的大华股份业绩大幅增长，2012年净利润同比增长85.2%，且2013年一季度净利润再度大幅预增。该股票2013年以来市价已经上涨约70%。	Take Dahua, the largest heavy-duty stock, as an example, benefiting from the construction of "smart cities" and the boom in the security industry, Dahua, the leading company, has seen substantial growth in its performance. Its net profit in 2012 increased by 85.2% year-on-year, and its net profit in the first quarter of 2013 Profits are again expected to increase substantially. The market price of the stock has risen by about 70% since 2013.	龙头公司 (Leading Company), 股份 (stock)	上涨(rise): { A0: 市价 (Market price), A1:70%, ARGM-TMP: 2013年以来(since 2013)}	股份 (stock)

Table 2.8: A failed case of MER. In this case the NER system did not recognize the correct entity name 大华股份 (Dahua Co. Ltd.).

Sentence	Translation	CIFRE extraction
美的集团旗下美的日电集团照明电气公司日前并购江西贵雅照明电器公司。	Midea Nippon Electric Group Lighting Electric Company, a subsidiary of Midea Group, recently acquired Jiangxi Guiya Lighting Electric Company.	并购(acquire):{A0: 日电集团照明电气公司(Nippon Electric Group Lighting Electric Company),A1:江西贵雅照明电器公司(Jiangxi Guiya Lighting Electric Company), ARGM-TMP: 日前(recently)}

Table 2.9: Example extractions captured by CIFRE but is labeled "unknown" in MG-lattice

Chapter 3

Self-Training in Controllable Text Generation

This chapter is a combination of two published papers at ACL 2023 [25] and IJCAI 2023 [24]. I was the main investigator of these two papers. and the works were mainly done during my internship at Microsoft Asia. Throughout the project, I led the process of defining project goals and key research questions, the model implementation, and designing and running the experiments. The works were done under the supervision of my intern mentor Xiaoyuan Yi and my supervisor Laks V.S. Lakshmanan.

3.1 Introduction

Recently, Pretrained Language Models (PLM) [18, 63, 83, 84] have shown superiority in Natural Language Processing (NLP). However, the ever-growing size of these models demands more training data, which destabilizes the fine-tuning of PLMs when labeled data is highly insufficient [134]. In this case, *Self-training (ST)* [31, 91, 127], a classical semi-supervised paradigm, has come to the fore again. As depicted in Fig. 3.1, ST produces pseudo labels for text using a classifier and then retrains the classifier with augmented data in an iterative process. By this means, ST utilizes massive unlabeled text to denoise the pseudo-annotated neighbors and improve the generalization on real data [112, 133], boosting various Natural Language Understanding (NLU) tasks [56, 73].

Nevertheless, how to apply ST to Natural Language Generation (NLG), especially the data-hungry attribute-controllable NLG, remains an open question. Different from typical NLU tasks like text classification, controllable NLG takes an attribute label as input to generate a textual sequence meeting the given attribute rather than predicting the label given input text.

This brings three new challenges for ST.

Challenge 1: Limited unlabeled data: since model inputs become discrete labels, there is no massive unlabeled data for the NLG model to

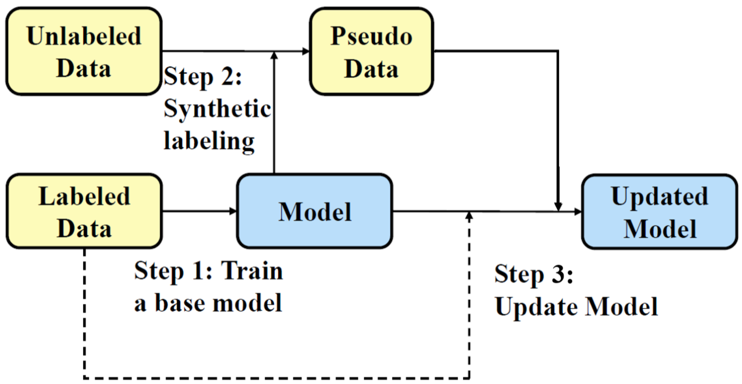


Figure 3.1: Classic Self-training. ST trains a base classification model on a small labeled data set. Then the model iteratively predicts pseudo labels for unlabeled data to augment the original set. Finally, we train the model using the augmented training set.

extend the learned distribution boundary. With limited unlabeled text, a potential approach to further improve ST performance is to leverage the generative ability of NLG models and produce synthetic (pseudo) text [90, 125] from given labels besides pseudo labels from text. This in turn raises two other major challenges.

Challenge 2: Over-exploitation. Augmented by self-generated text, NLG models are forced to repeatedly fit the already learned text distribution. This gradually homogenizes the generated pseudo text and causes a shrunken (collapsed) generalization boundary, resulting in decreased controllability and generation diversity.

Challenge 3: Training deceleration. We need to re-generate all pseudo text in each ST iteration with updated model parameters, which interrupts the parallelism of Transformer [101]-based models, severely decelerating training and impairing practicality.

To handle *Challenge 1* and *Challenge 2*, we propose a novel **Du**al **N**oisy **S**elf **T**raining (**DuNST**) method, for semi-supervised controllable NLG. DuNST jointly learns to generate text from given attribute labels and predict labels for text, characterizing these two directions as a dual process via a shared Variational AutoEncoder (VAE) [49]. Such duality allows our model to leverage not only generated pseudo text but also pseudo labels predicted for available unlabeled text. Both generation and classification would be augmented by the two kinds of pseudo data that will hence be gradually refined during the ST process, handling *Challenge 1*. Besides, DuNST corrupts the generated pseudo text by two kinds of noise, softmax temperature and soft pseudo text, to further disturb and escape from the text space learned at the previous ST iteration, addressing *Challenge 2*. Our method can be theoretically regarded as propagating local smoothness [10] and exploring a larger and potential space, which helps extend the generalization boundary and improve attribute coverage. Therefore, DuNST further boosts controllability with comparable generation fluency and diversity.

Though DuNST works well, it has the following limitations. First, DuNST still did not address *Challenge 3*. As with all other Self-training methods, DuNST also needs to reproduce pseudo labels and pseudo text at each ST iteration. Since the pseudo text (both hard and soft) is generated in an auto-regressive manner, it is impossible to be done in parallel and thus leads to longer training time. Besides, because of the dual VAE structure in DuNST, it involves multiple optimization objectives (classification, generation, and KL terms) as well as the weights for BOW loss and cyclical annealing. There are many hyperparamters to be tuned carefully. Thus, it is somewhat challenging to tune all the hyperparameters to get the best performance due to the large search space.

Further, to tackle *Challenge 2* and *Challenge 3* and address the limitation of DuNST, we propose another novel self-training framework, **K**ernel Distance Based **E**fficient **S**elf **T**raining (**KEST**), for improving semi-supervised controllable NLG. Instead of learning from generated pseudo textual sequences with traditional cross-entropy loss, KEST directly fits the approximated text distribution from the last iteration in the embedding space. Such an objective not only relaxes the constraint imposed by the previous ST iteration but also encourages diverse outputs of the current model, addressing *Challenge 2*. Besides, we design a non-autoregressive generation schema to produce soft representations of pseudo text (rather than hard strings) in parallel, greatly reducing time cost and handling *Challenge 3*. Furthermore, such a soft text is naturally a kind of noisy pseudo data [37, 117], which helps the model denoise errors and propagate local smoothness [10, 112].

In the remainder of this chapter, Sec. 3.2 talks about the related work. Sec 3.3 and Sec. 3.4 describe the methods and experimental results of DuNST and KEST, respectively. Sec. 3.3.3 introduces the datasets, evaluation metrics, and baselines used in the experiments of both DuNST and KEST.

3.2 Related Work

Controllable Language Generation Attribute-controllable language generation aims to generate high-quality text satisfying desired attributes, *e.g.*, sentiment, topic and style, which could facilitate diverse downstream applications, such as stylistic writing [26] and language detoxification [29]. In the era of PLM, an effective paradigm for controllable NLG lies in fine-tuning PLMs on datasets containing labeled text [34, 48]. However, as the scale of PLMs keeps increasing, insufficient labeled data becomes a new obstacle to fine-tuning [128, 134].

As a remedy, two lines of research have been developed. *Lightweight tuning* searches a trigger [92] or optimizes only a few parameters like adapter [87] or prefix [57, 80], requiring much less training data. *Plug-in control* manipulates the output generation probability of models to encourage attribute-related tokens. The manipulation is achieved broadly through two paradigms: updating cached hidden states [16] or reshaping the output distribution guided by off-the-shelf attribute classifiers [50, 124] or conditional PLMs [60] at inference time without fine-tuning. Despite no/weak dependence on labeled data, these two lines of work would cause limited control accuracy or decreased fluency.

Self-training Recently, Self-training has flourished again by iteratively generating pseudo labels and augmenting the tuning of data-hungry PLMs, showing great advantages in further enhancing NLU [4, 10, 21, 70, 102] and Neural Machine Translation (NMT) [37, 46] where massive unlabeled input text exists. Besides classical ST, diverse follow-up modifications have been developed for further improvement, which generally fall into two lines. The first line, *i.e.*, *sample selection*, selects only a part of unlabeled instances in terms of (1) model confidence to avoid over-noisy pseudo labels [4, 93], (2) prediction uncertainty to obtain informative instances and enhance performance on the hard ones [46, 73], or (3) label balance to benefit minority classes [111]. The other line is *noisy labeling* [37, 117], which injects synthetic noise into the pseudo data, *e.g.*, token shuffle or image distortion to propagate local smoothness and improve model robustness.

However, as mentioned in Sec.3.1, due to *Challenges 1&2*, it is difficult to directly apply ST (as well as the synthetic noise above) to attribute-controllable NLG.

VAE and Dual Learning VAE [49] has proven to be effective in generating diverse text when combined with PLMs due to the flexible semantic properties captured in the latent space [40, 52], which could further enhance the variety of pseudo text and thus is more suitable for ST. Dual Learning (DL) [36] has been traditionally proposed and applied in NMT and then extended to joint optimization of NLU-NLG tasks [116], which is promising for tackling Challenge 1. Tseng et al. [99] successfully combined DL with VAE for table-to-text and text-to-table generation, but their model cannot simultaneously optimize the two directions and share learned features, not compatible with our design for Challenge 1.

Non-Autoregressive Generation (NAG) Relevant to our work, NAG aims to simultaneously generate all target tokens rather than one by one to increase the inference speed. NAG was first proposed in NMT [33, 66] and then applied to broader scenarios like Text Summarization [62] and Text-to-Speech Synthesis [13]. All the tasks are learned with encoder-decoder architectures, relying on long input sequences (*e.g.*, source language) to provide rich initial context information. However, it is still challenging to leverage NAG for controllable generation since the inputs are only attribute labels and short prompts.

Unlike the aforementioned work, in DuNST (Sec 3.3) we revisit the challenges of incorporating Self-training with controllable generation and utilize the duality and flexible noise to handle these challenges, leading to a novel and practical ST framework. In KEST (Sec 3.4), we take a further step to investigate the challenges of incorporating ST with controllable NLG and propose a practical NAG method to generate soft pseudo text, which is then learned in a kernel space, leading to a more efficient ST framework.

3.3 DuNST

3.3.1 Contributions

The contributions of DuNST are as follows:

- To the best of our knowledge, we are the first to incorporate Self-training into semi-supervised controllable language generation and propose a novel and effective ST method.
- We demonstrate that DuNST explores a larger potential text space and extends the generalization boundary, providing a theoretical interpretation for our method.

- We conduct thorough experiments on three attribute-controllable generation tasks and manifest the superiority of DuNST in improving control accuracy with competitive quality of the generated text, further exploiting the capacity of powerful PLMs for NLG.

3.3.2 Methods

Formulation and Overview

Let $\mathbf{x}$ be the text, y be the attribute label, $D_L = \{\mathbf{x}_i, y_i\}$ be a labeled dataset with paired text and its corresponding label, and $D_U = \{\mathbf{x}_i\}$ be an unlabeled dataset from the same domain. We aim to learn an attribute-controllable generator $\mathcal{G} = P_\theta(\mathbf{x}|y)$ parameterized by θ (*e.g.*, a large PLM) to generate high-quality text $\mathbf{x} \sim P_\theta(\mathbf{x}|y)$ (in an auto-regressive manner) satisfying the given label y. We also endow our model with the ability to produce pseudo attribute labels for $\mathbf{x} \in D_U$ through jointly learning a text classifier $\mathcal{C} = P_\phi(y|\mathbf{x})$. We simultaneously model and optimize $\mathcal{G}$ and $\mathcal{C}$ with a shared PLM as a dual process (Sec. 3.3.2).

During the training of DuNST (Sec. 3.3.2), the pseudo labels predicted by $\mathcal{C}$ help cover more unseen samples and hence extend the learned distribution boundary (*tackling Challenge 1*), while the noisy pseudo text generated by $\mathcal{G}$ helps perturb the previously learned space, further improving generalization (*addressing Challenge 2*). Though we emphasize generation in this work, both $\mathcal{G}$ and $\mathcal{C}$ would be promoted and thus keep refining the augmentation data during ST, which acts as a *joint exploration and exploitation* process (Sec.17).

Dual Generation and Classification

We jointly learn the conditional distribution of text $P_\theta(\mathbf{x}|y)$ and label $P_\phi(y|\mathbf{x})$ to match the real ones. However, we don't directly optimize them with traditional cross-entropy loss but resort to the variational approaches [49]. In detail, we involve a latent variable $\mathbf{z}$ to capture the underlying semantics and hence have $P(\mathbf{x}|y) = \int P(\mathbf{x}, \mathbf{z}|y) d\mathbf{z}$. We could sample a generated text $\mathbf{x}$ by the decomposition $P(\mathbf{x}, \mathbf{z}|y) = P(\mathbf{x}|\mathbf{z}, y) * P(\mathbf{z}|y)$. To this goal, we minimize a generation loss as:

$$\begin{aligned}\mathcal{L}_g = &-\mathbb{E}_{Q_\psi(\mathbf{z}|\mathbf{x},y)}[\log P_\theta(\mathbf{x}|\mathbf{z}, y)] \\ &+ \mathrm{KL}[Q_\psi(\mathbf{z}|\mathbf{x}, y)||P_\theta(\mathbf{z}|y)],\end{aligned} \tag{3.1}$$

where $Q_\psi(\mathbf{z}|\mathbf{x},y)$ and $P_\theta(\mathbf{z}|y)$ are approximated posterior and prior distributions of $\mathbf{z}$ and KL is the Kullback–Leibler divergence, respectively. Optimizing this loss is equivalent to maximizing a lower bound of $P_\theta(\mathbf{x}|y)$.

The posterior $Q_\psi(\mathbf{z}|\mathbf{x},y)$ is typically assumed as a multivariate Gaussian $\mathbb{N}(\mu_{post},\sigma_{post})$ and approximated by $[\mu_{post},\log\sigma_{post}]=\mathrm{MLP}([\mathbf{h}_\mathbf{x},\mathbf{h}_y])$ with $\mathbf{h}_\mathbf{x}=\mathrm{Encoder}(\mathbf{x})$, where $\mathbf{h}_y$ is the label embedding of y. Encoder is a Transformer [101] encoder, and MLP is a multilayer perceptron. Similarly, we could build the prior $P_\theta(\mathbf{z}|y)\sim\mathbb{N}(\mu_{\text{gen-prior}},\sigma_{\text{gen-prior}})$ where $[\mu_{\text{gen-prior}},\log\sigma_{\text{gen-prior}}]=\mathrm{MLP}(\mathbf{h}_y)$.

Symmetrically, we could build the classification prior in a similar manner: $P_\phi(\mathbf{z}|\mathbf{x})\sim\mathbb{N}(\mu_{\text{cls-prior}},\sigma_{\text{cls-prior}})$ where $[\mu_{\text{cls-prior}},\log\sigma_{\text{cls-prior}}]=\mathrm{MLP}(\mathbf{h}_\mathbf{x})$. Then we optimize classification by:

$$\begin{aligned}\mathcal{L}_c = &-\mathbb{E}_{Q_\psi(\mathbf{z}|\mathbf{x},y)}[\log P_\phi(y|\mathbf{z},\mathbf{x})]\\ &+\mathrm{KL}[Q_\psi(\mathbf{z}|\mathbf{x},y)||P_\phi(\mathbf{z}|\mathbf{x})].\end{aligned} \tag{3.2}$$

The text is generated by an autoregressive Transformer decoder $\mathbf{x}=\mathrm{Decoder}(\mathbf{z})$ and the label is predicted by $y=\mathrm{MLP}(\mathbf{z})$ with $\mathbf{z}$ drawn from the posterior distribution in training and from the prior ones in testing. $\mathcal{G}$ and $\mathcal{C}$ share most parameters (*e.g.*, encoder), as well as the same posterior distribution $Q_\psi(\mathbf{z}|\mathbf{x},y)$, to enhance the connection of text and corresponding labels, and better utilize the knowledge learned via the two directions.

The final loss is computed as follows:

$$\mathcal{L}=\lambda_g\mathcal{L}_g+\lambda_c\mathcal{L}_c, \tag{3.3}$$

where λ_g and λ_c are hyper-parameters to balance the importance of classification and generation. We will show later that such variational dual learning further boosts controllability and text diversity (Sec. 3.3.4) and helps refine pseudo labels (Sec. 3.3.4).

Dual Noisy Self-training

As discussed in Sec. 3.1, augmented only by self-generated text, the model would increasingly enhance the exploitation of the previously learned space but fail to explore more, resulting in constrained attribute distributions and thus marginal improvement of control accuracy (*Challenge 2*, see Table 3.3), Table 3.5, and Table 3.4. Injecting noise into pseudo text is a practical way to facilitate exploration. However, the typical synthetic noise [37] (*e.g.*, randomly shuffle tokens in pseudo text) encourages isotropic exploration, which may diverge far from the valid space and get too noisy for NLG.

Algorithm 1: Training Process of DuNST

Input: Labeled set D_L, unlabeled set D_U, attribute set Y.

Jointly train base model $\mathcal{G}$, $\mathcal{C}$ on D_L by optimizing Eq.(3.3), store the best $\mathcal{G}_0$, $\mathcal{C}_0$.
for *epoch* $\leftarrow 1$ **to** *MaxEpoch* **do**
    for $\mathbf{x}_i$ ***in*** D_U **do**
        $\hat{y}_i = \mathcal{C}_{epoch-1}(\mathbf{x}_i)$
    end
    Build pseudo label set: $D_{PL} = \{\mathbf{x}_i, \hat{y}_i\}$
    for y_j ***in*** Y **do**
        Sample t priors: $\{z_k\}_{k=1}^{t} \sim P_\theta(\mathbf{z}|y_j)$
        for $k \leftarrow$ *0* ***to*** t **do**
            for $m \leftarrow$ *0* ***to*** *MaxLength* **do**
                Compute soft pseudo token $\mathbf{d}_k^m$ using $\mathcal{G}_{epoch-1}$ and Eq.(3.4). Set $y_k \leftarrow y_j$.
            end
        end
    end
    Build soft pseudo text: $D_{PT} = \{\mathbf{d}_k, y_k\}$
    Train $\mathcal{G}_{epoch-1}$, $\mathcal{C}_{epoch-1}$ on $\{D_{PT}, D_{PL}, D_L\}$ by optimizing Eq.(3.3) and Eq.(3.5), update the parameters to $\mathcal{G}_{epoch}$ and $\mathcal{C}_{epoch}$.
end

To address this problem, we propose two novel and effective types of soft noise to enable safer exploration, namely *High-temperature Generation* and *Soft Pseudo Text*, in what follows.

High-temperature Generation (HTG): We introduce temperature τ in the softmax layer:

$$\mathbf{d}^m = \sigma(\mathcal{G}(y, \hat{\mathbf{x}}_{<m}, \mathbf{z})/\tau), \tag{3.4}$$

where $\mathbf{d}^m$ is the output token distribution for the m-th token, $\hat{\mathbf{x}}_{<m}$ is the previously generated $m-1$ tokens and σ means softmax. Lower τ (*e.g.*, $\tau < 1$) leads to a sharper distribution and thus motivates more certain output (usually used in NMT). Differently, we choose $\tau > 1$ to encourage more diverse but semantically reasonable (high generation probability) tokens which could enhance local smoothness and help explore more potential directions. Besides, the degree of noise is easy to control by adjusting τ for a better trade-off.

Soft Pseudo Text (SPT): HTG improves the diversity of pseudo text, but also takes the risk of sampling invalid tokens and propagating errors in an autoregressive generation. Moreover, HTG produces discrete pseudo text (a point in text space) and thus requires numerous sampled pseudo text (points) to cover a small neighborhood (Fig. 3.7). Therefore, we further propose to generate soft pseudo text, where we directly store the output token distribution $\mathbf{d}$ and let $\mathcal{G}$ directly learn to reproduce $\mathbf{d}$. Then we replace Eq.(3.1) with:

$$\mathcal{L}_g^{'} = \begin{cases} -\log P_\theta(\mathbf{x}|\mathbf{z}, y) + \\ \mathrm{KL}[Q_\psi(\mathbf{z}|\mathbf{x}, y)||P_\theta(\mathbf{z}|y)], \mathbf{x}, y \in D_L, D_{PL} \\ \mathrm{KL}[\mathbf{d}||P_\theta(\mathbf{x}|\mathbf{z}, y)] + \\ \mathrm{KL}[Q_\psi(\mathbf{z}|\mathbf{x}, y)||P_\theta(\mathbf{z}|y)], \mathbf{x}, y \in D_{PT}. \end{cases} \tag{3.5}$$

Such SPT acts as a kind of Knowledge Distilling [38] in an iterative manner. In this way, we avoid losing relevant semantic information in $\mathbf{d}$ and reduce needed samples, further extending the generalization boundary (see Table 3.7).

The complete algorithm is described in Alg. 1.

Theoretical Analysis

To understand why DuNST could work well, we interpret its advantages with the following theorem:

Theorem 1. *Optimizing the training objective of DuNST is equivalent to approximately minimizing the upper bound of*

$$KL\left[Q||P_\theta\right] + KL\left[P_{\theta'}||P_\theta\right] + KL\left[U||P_\theta\right], \tag{3.6}$$

where Q is the real text distribution, P_θ and $P_{\theta'}$ are models estimated at the current and last ST iteration, respectively, and U is a noise distribution.

Proof. **Derivation of Dual VAE ELBO:** To optimize the attribute-controllable generation direction, we aim at learning the conditional distribution of text, namely $P(x|y)$, and derive the evidence lower bound (ELBO) as:

$$\begin{aligned}
&\log P(x|y)\\
=&\log \int P(x,z|y)\frac{Q(z|x,y)}{Q(z|x,y)}dz\\
=&\log \mathbb{E}_{Q(z|x,y)}[\frac{P(x,z|y)}{Q(z|x,y)}]\\
\geq&\mathbb{E}_{Q(z|x,y)}[\log \frac{P(x|z,y)P(z|y))}{Q(z|x,y)}]\\
=&\mathbb{E}_{Q(z|x,y)}[\log P(x|z,y)] - \mathrm{KL}[Q(z|x,y)||P(z|y)]\\
=&-\mathcal{L}_g,
\end{aligned}$$

where we approximate the true prior and posterior distributions $P(z|y)$, $Q(z|x,y)$ with a prior network and a posterior network (a.k.a. recognition network). The last but two lines is from Jensen's inequality. When we input a prompt c as in our experiments on IMDb, similarly we can get $\log P(x|y,c) \geq \mathbb{E}_{Q(z|x,y,c)}[\log P(x|z,y,c)] - \mathrm{KL}[Q(z|x,y,c)||P(z|y,c)]$.

For attribute label classification, we maximize $P(y|x)$ and get a ELBO symmetrically:

$$\begin{aligned}
&\log P(y|x)\\
=&\log \int P(y,z|x)\frac{Q(z|x,y)}{P(z|x,y)}dz\\
=&\log \mathbb{E}_{Q(z|x,y)}[\frac{P(y,z|x)}{Q(z|x,y)}]\\
\geq&\mathbb{E}_{Q(z|x,y)}[\log \frac{P(y|z,x)P(z|x))}{Q(z|x,y)}]\\
=&\mathbb{E}_{Q(z|x,y)}[\log P(y|z,x)] - \mathrm{KL}[Q(z|x,y)||P(z|x)]\\
=&-\mathcal{L}_c,
\end{aligned}$$

where we similarly approximate the true prior $P(z|x)$ with another prior network.

Please note that the two optimization directions shared most parameters, and utilize the same recognition network but incorporate different prior distributions.

Proof of Theorem 1 For brevity, we ignore the hyper-parameters λ. Define Q as the real data distribution while P as the estimated one. We assume we could approximate the real prior distribution of label, $Q(y)$, by statistics under the i.i.d. assumption, and assume our model also estimates the real text distribution, $Q(x)$, well enough with a large unlabeled dataset D_U. That is, $\mathrm{KL}[Q(x)||P(x)] < \epsilon$ and $\mathrm{KL}[Q(y)||P(y)] < \epsilon$. Then over the whole labeled dataset $Q(x, y)$ we have:

$$\begin{aligned}
&\mathcal{L}_g + \mathcal{L}_c \\
=&\mathbb{E}_{Q(x,y)}\{-\mathbb{E}_{Q(z|x,y)}[\log P(x|z,y) \\
&+\log P(y|z,x)] + \mathrm{KL}[Q(z|x,y)||P(z|y)] \\
&+\mathrm{KL}[Q(z|x,y)||P(z|x)]\} \\
=&\mathbb{E}_{Q(x,y)}\{\int Q(z|x,y)[\log \frac{Q(z|x,y)}{P(x|y,z)P(z|y)} \\
&+\log \frac{Q(z|x,y)}{P(y|x,z)P(z|x)}]dz\}.
\end{aligned}$$

Then we consider the left term of the above equation and have:

$$\begin{aligned}
&\mathbb{E}_{Q(x,y)}\{\int Q(z|x,y)[\log \frac{Q(z|x,y)}{P(x|y,z)P(z|y)}dz \\
=&\mathbb{E}_{Q(x,y,z)}\{\log \frac{Q(x,y,z)P(y,z)P(y)}{Q(x,y)P(x,y,z)P(y,z)}\} \\
=&\mathrm{KL}[Q(x,y,z)||P(x,y,z)] + \mathbb{E}_{Q(x,y)}[\log \frac{P(y)}{Q(x,y)}] \\
\approx&\mathrm{KL}[Q(x,y,z)||P(x,y,z)] + H_Q(x|y) \\
\geq&\mathrm{KL}[Q(x,y,z)||P(x,y,z)],
\end{aligned}$$

where the second last step is because by assumption we have $Q(y) \approx P(y)$. Similarly, for the left term, we have:

$$
\begin{aligned}
&\mathbb{E}_{Q(x,y)}\{\int Q(z|x,y)[\log \frac{Q(z|x,y)}{P(y|x,z)P(z|x)}dz \\
\approx&\mathrm{KL}[Q(x,y,z)||P(x,y,z)] + H_Q(y|x). \\
\geq&\mathrm{KL}[Q(x,y,z)||P(x,y,z)].
\end{aligned}
$$

Combining all the results above, we conclude:

$$
\mathcal{L}_g + \mathcal{L}_c \geq \mathrm{KL}[Q(x,y,z)||P(x,y,z)]. \tag{3.7}
$$

Then we consider the scenario of Self-training. Define the real distribution formed by the dataset as $Q(x,y,z)$, the estimated distribution at the last ST iteration as $P_\theta^{'}(x,y,z)$ which is formed by the generated pseudo labels and text, and the one at the current ST iteration as $P_\theta(x,y,z)$. As discussed in Sec. 3.3.2, we add noise to pseudo text to enhance exploration. Therefore, the previously learned $P_\theta^{'}(x,y,z)$ is disturbed and becomes $P_\theta^{'}(x,y,z)+U$ where U is the noise distribution. For brevity, we abbreviate these distributions as Q, $P_\theta^{'}$, P_θ and U, respectively. In Self-training, we are actually fitting P_θ to not only Q but also $P_\theta^{'}$ and U. Therefore, we are minimizing an upper bound of:

$$
\begin{aligned}
&\mathrm{KL}[Q+P_\theta^{'}+U||Q_\theta] \\
=&\int (Q+P_\theta^{'}+U)\log \frac{Q+P_\theta^{'}+U}{P_\theta}d.
\end{aligned}
$$

Consider the first term:

$$
\begin{aligned}
&\int Q \log \frac{Q+P_\theta^{'}+U}{P_\theta}d \\
=&\int Q \log \frac{Q}{P_\theta} * \frac{Q+P_\theta^{'}+U}{Q}d \\
=&\mathrm{KL}[Q||P_\theta] - \mathrm{KL}[Q||Q+P_\theta^{'}+U].
\end{aligned} \tag{3.8}
$$

Since Q, $P_\theta^{'}$ and P are all fixed at the current iteration, we can ignore the last term $\mathrm{KL}[Q||Q+P_\theta^{'}+U]$. Similarly, we have that minimizing $\mathrm{KL}[Q+P_\theta^{'}+u||P_\theta]$ equals to minimizing $\mathrm{KL}[Q||P_\theta]+\mathrm{KL}[P_\theta^{'}||P_\theta]+\mathrm{KL}[U||P_\theta]$, concluding the proof. □

In Theorem 1, the first KL term corresponds to the optimization of Eq.(3.3) that approximates the real distribution. The second term corresponds to classic Self-training, which works as a regularization. As depicted in Fig. 3.7, such regularization forces the model to fit the already learned space, causing over-exploitation. The last one is the noise to enhance exploration. Compared to the isotropic synthetic noise (too noisy) and the hard pseudo text (too sparse), DuNST with soft pseudo text could explore potential directions, cover larger space more smoothly, and thus further push the boundary.

3.3.3 Datasets and Baselines

Tasks

We conduct exhaustive experiments on three controllable generation tasks, described below:

Sentiment control with prompt: We evaluate the controllability of sentiment on the IMDb movie review dataset [67]. Following Dathathri et al. [16], we use their 15 manually created prompts and another 85 sampled from IMDb (100 in total) as model input and generate 10 samples for each prompt and each sentiment.

Topic control w/o prompt: We use the AGNews dataset [135] to evaluate topic controllability. We assess our model's ability to generate from scratch on this dataset and sample 300 generations for each topic.

Text detoxification: We use the Jigsaw Toxic Classification Dataset. Following Qian et al. [80], we use the 203 "challenging" prompts (toxicity < 0.5) from Gehman et al. [29], and generate 10 non-toxic sentences for each prompt.

Dataset Description

For IMDb[10] dataset [67], the authors claimed in their paper that *In the interest of providing a benchmark for future work in this area, we release this dataset to the public without claiming any further copyright.* For AGNews [11] dataset [135], it is claimed in the website that *You are encouraged to download this corpus for any non-commercial use.* For Jigsaw [12] dataset, the dataset is under CC0, with the underlying comment text being governed by Wikipedia's CC-SA-3.0. All datasets we used are open-sourced and are used for research only, which is consistent with their intended use.

For IMDb dataset and AGNews dataset, we leave 10% of the training set as validation data, and others as training data. For the AGNews dataset, we

[10] `https://huggingface.co/datasets/imdb`
[11] `https://www.kaggle.com/amananandrai/ag-news-classification-dataset`
[12] `https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/`

use the description for text generation and wrote a script to resolve HTML tags. For Jigsaw dataset, we apply a binary setting where we keep the "non-toxic" class unchanged and group all other classes into "toxic" class.

	labeled	Unlabeled	Dev	Test	Avarage Length
IMDb(5%)	1125	33750	2500	25000	270
AGNews(3%)	3240	97200	12000	7600	41
Jigsaw(3%)	4308	43080	15957	63978	73

Table 3.1: Description of datasets used in the experiment

	Acc. ↑	F1 ↑	AUC ↑
IMDb			
RoBERTa-large	96.15	96.20	99.22
BERT-base	88.40	88.62	95.21
AGNews			
RoBERTa-large	94.88	94.89	99.34
BERT-base	89.93	89.91	98.23

Table 3.2: Classifier performance of our evaluator RoBERTa-large and pseudo labeler BERT-base on the test set.

We sample 5% of IMDb training samples as labeled data and directly take their provided unlabeled set. Since there is no separate unlabeled text in AGNews dataset, 3% of training samples as labeled data and use the others as unlabeled ones. For a fair comparison, we keep the ratio of labeled/pseudo/unlabeled data to 1:1:30.

The details of final datasets that we used in the experiments are described in Table 3.1. IMDb and AGNews are balanced datasets, where there are equal numbers of labeled data for each class in the training set. For the Jigsaw dataset, there are only 414 toxic data (9.6%) in the Jigsaw dataset, which shows that Jigsaw is an extremely imbalanced dataset, bringing difficulty in detoxification.

Evaluation Metrics

In this chapter, we mainly focus on the controllable NLG side, considering the following four kinds of metrics. We set the minimum generation length to

10. For the maximum length, we set 490 for sentiment, 50 for detoxification, and 40 for topic. We evaluate generation quality on the following metrics:

Fluency: We evaluate generation fluency by the perplexity of generated text measured by GPT2-XL [83], *i.e.*, **Output PPL**.

Generalizability: We calculate the perplexity of each model on each testing set, *i.e.*, **Model PPL**, to evaluate the generalizability of the model. For VAE-based models, we can only obtain the lower bound of $\log P(x|y)$. Following [40, 52], We consider k latent variables $z_1, z_2, ..., z_k$ sampled from the posterior distribution $Q(z_i|x, y)$, and PPL based on these latents $P(x, z_i|y)$. Based on the fact that average importance weights are an unbiased estimator of $\log P(x)$ [7] and Jensen's Inequality, we have:

$$\begin{aligned} L_k =& \mathbb{E}\left[\log \frac{1}{k} \sum_{i=1}^{k} \frac{P(x, z_i|y)}{Q(z_i|x, y)}\right] \\ \leq& \log \mathbb{E}\left[\frac{1}{k} \sum_{i=1}^{k} \frac{P(x, z_i|y)}{Q(z_i|x, y)}\right] = \log P(x|y) \end{aligned} \tag{3.9}$$

Thus we use L_k to estimate the output PPL in VAE-like models.

Controllability: We evaluate the control accuracy through classification performance (accuracy (**Acc**) and Macro-F1(**F1**)) on the generated text by the two fine-tuned RoBERTa-large classifiers for sentiment and topic. Table 3.2 presents the performance of our evaluator RoBERTa-large. We find that RoBERTa-large has a satisfactory classification accuracy and F1 on these two tasks, and thus is able to act as a good evaluator of generation quality. For detoxification, we report the percentage of toxic sentences (**Toxic %**) using Google Perspective API. Perspective API is a free API for scoring the toxicity of text. Following Qian et al. [80] we also use this Perspective API for toxicity evaluation.

Diversity: To evaluate the diversity of generated text, we consider the following metrics: (1) **Dist-n** [54]: the percentage of distinct n-grams on generated samples. We evaluate on $n = 1, 2, 3, 4$ and compute the geometric mean as **Dist**. Higher **Dist** means the generation contains more diverse n-grams. (2) **Self-BLEU** [139]. Self-Bleu calculates the BLEU score on the generated samples, which averages the BLEU score of each generated sequence calculated with other generated ones as references. The BLEU score is computed as the geometric mean of BLEU-n ($n = 2, 3, 4$). This metric measures the diversity of a set of generated sequences. Lower Self-BLEU means these generated sequences are more distinguishable from each other. Note that Self-BLEU only compares n-grams between different generations with given prompts, while Dist also considers the diversity *inside* a piece of text.

Among all the above metrics, Accuracy, F1, AUC, Dist-n, and Self-BLEU are reported as 100 times their original value for convenience.

Baselines

We compare our model with three kinds of (supervised or semi-supervised) strong NLG baselines.

Finetune PLM: We finetune different powerful PLMs on each downstream dataset, including GPT2 [83], UniLM [18] and T5 [84]. We feed a prepend sentence as a control sentence. For sentiment-controlled generation, we use *This is a [positive/negative] review* as the control sentence. For topic-controlled generations, we use *The following is about [topic]*. For detoxification, we use *This is a [toxic/non-toxic] comment* as a control sentence. Since T5 acts in a sequence-to-sequence manner, we feed the control sentence to the encoder and training text to the decoder. We fine-tune all pre-trained LMs under learning rate 5e-5 for 10 epochs and warmup steps to be 1 epoch.

Lightweight fine-tuning methods:
(1) Prefix-tuning (PF) [57]: this method only tunes the prefix and freezes all parameters of the PLM, requiring fewer data. (2) Ctr-PF[80]: A contrastive version of PF. We follow the implementation details described in Qian et al. [80].

Self-training methods:
(1) PT: the classical Self-training [31], which generates pseudo text in each ST iteration and updates parameters with both real and pseudo text from the last iteration.
(2) PT(noise): Noisy Self-training [37], which brings synthetic noise (token drop, swap and mask) to the pseudo text for self-training. We use the same implementation of Noise Layer as He et al. [37]. We set the token drop rate and mask rate to 5%. Since GPT2 does not have a *Mask* token, we randomly substitute this token for another token. We set the parameter of word shuffle to 1.1.
(3) PT(noise)+PL: We combine PT(noise) and *pseudo labeling* to produce and utilize both pseudo text and pseudo labels, which are predicted from the real unlabeled text by a BERT-base [17] fine-tuned on our labeled data. The performance of pseudo labeler BERT-base-cased in Table 3.2.
(4) PT(select)+PL: PT(select) is a modified ST method with sample selection [73], which over-generates noisy pseudo text and selects high-quality ones by the classifier confidence and uncertainty. Specifically, the classi-

fication confidence s_{conf} is the softmax probability of the predicted label by the previously trained BERT-base-cased classifier. Uncertainty score $s_{uncertain}$ is Bayesian Active Learning by Disagreement (BALD) computed by Monte-Carlo Dropout [73]. A high BALD score means the model is highly confused. We want to select the sample with high confidence and low BALD score. Thus we select samples based on the following score:

$$s_{select} = s_{conf} + \frac{1e-5}{s_{uncertain}}$$

Besides, we also provide the evaluation results of Ground Truth as an upper bound of performance.

Human Evaluation Metrics

For Sentiment/Topic-controlled generation tasks, annotators are asked to score the generations on the following three criteria **Fluency**, **Novelty**, and **Attribute Relevance**. For the detoxification task, annotators are asked to score the generations on the following three criteria **Fluency**, **Novelty**, and **Toxicity**. We ask the annotators to evaluate the quality of generated text only after the prompt since the prompt itself might be toxic. We attach the instruction for detailed metrics to the annotators below.

Fluency: Whether the generated text is fluent/grammatically correct. A small number of HTML tags/special words (like ¡br¿) are permitted. The score ranges from 1 (least fluent) to 5 (most fluent).

Novelty: The extent of novelty INSIDE the generated text. E.g., distinct words/expressions, and repeated phrases. The score ranges from 1 (least novel, a sentence/word repeat till the end) to 5 (most novel, all sentence describes new stuff).

Attribute Relevance: Whether the generated text is relevant to its label. There might exist examples that contain both positive/negative expressions. 5=exactly relevant; 4=mostly relevant; 3=neutral (not contain sentiment or 50/50); 2=slightly relevant, but more similar to another class;1=totally opposite.

Toxicity: whether the provided textual sequences contain any toxic content, including but not limited to offensive text, abusive language, swearwords, hate speech, denigrating messages, microaggression, discrimination, sex, rude words, and hominem attack. The score ranges from 1 (most non-toxic) to 5 (most toxic).

	Sentiment						
	O-PPL ↓	M-PPL ↓	Acc ↑	F1 ↑	AUC ↑	Dist ↑	S-BLEU ↓
Test set	25.14	–	96.15	96.20	99.22	48.27	43.34
GPT2(raw)	13.20	38.39	55.9	68.5	61.37	35.91	58.79
Finetune LM							
GPT2	16.40	44.02	77.55	80.44	88.35	26.34	71.00
UniLM	25.20	54.33	76.45	75.35	85.18	31.05	66.97
T5	25.69	34.97	82.80	83.77	90.50	30.03	69.57
Lightweight method							
PF	13.02	37.09	67.55	75.05	81.84	29.48	65.10
Ctr-PF	13.01	37.12	71.00	77.33	86.51	29.63	**64.83**
Self-Training with GPT2							
+PT	14.62	68.04	76.10	79.57	87.92	30.58	65.22
+PT+noise	11.91	44.31	74.95	77.46	85.02	25.40	72.19
+PT(noise)+PL	11.26	33.85	87.60	88.47	95.59	27.26	70.90
+PT(select)+PL	**10.89**	33.89	88.32	88.75	96.24	27.17	71.41
Self-Training with UniLM							
+PT	26.62	58.37	72.2	70.27	80.37	31.17	66.69
+PT+noise	30.28	62.07	77.75	75.78	85.35	31.68	65.18
+PT(noise)+PL	18.92	**33.53**	89.95	89.73	96.38	30.94	66.84
+PT(select)+PL	18.40	33.56	90.08	90.06	96.66	31.27	67.61
Our Methods							
DuNST	21.67	42.82	**92.90**	**93.05**	**98.02**	**31.79**	65.80

Table 3.3: Results of DuNST on IMDb dataset.

	Topic						
	O-PPL ↓	M-PPL ↓	Acc ↑	F1 ↑	AUC ↑	Dist ↑	S-BLEU ↓
Test set	31.04	–	94.88	94.89	99.34	67.24	23.31
GPT2(raw)	16.94	74.41	55.75	52.17	83.28	46.88	45.55
Finetune LM							
GPT2	22.22	23.46	82.92	83.08	95.23	54.93	39.93
UniLM	55.79	36.28	87.67	87.70	96.30	54.76	43.77
T5	48.33	32.12	88.33	88.43	97.95	58.06	37.01
Lightweight method							
PF	**20.27**	32.35	68.67	68.44	87.14	59.17	32.73
Ctr-PF	20.41	33.90	83.25	83.21	95.47	**60.34**	**31.20**
Self-Training with GPT2							
+PT	23.74	27.88	83.50	83.55	95.49	57.89	36.02
+PT+noise	26.39	27.02	82.42	82.45	94.58	58.06	35.53
+PT(noise)+PL	30.62	**13.96**	87.83	87.48	97.42	47.11	56.67
+PT(select)+PL	31.34	14.07	87.92	87.54	97.46	46.71	57.33
Self-Training with UniLM							
+PT	57.40	40.95	86.42	86.36	96.69	52.35	46.41
+PT+noise	58.59	45.32	85.42	85.27	95.88	53.35	46.57
+PT(noise)+PL	32.36	16.64	89.67	89.70	98.11	53.79	47.95
+PT(select)+PL	33.23	16.66	90.5	90.52	98.31	53.71	47.69
Our Methods							
DuNST	34.73	33.58	**93.58**	**93.59**	**98.99**	59.42	37.02

Table 3.4: Results of DuNST on AGNews dataset.

3.3.4 Experiments

Experimental Settings

We use UniLM-base-cased [18] as the shared encoder and decoder of DuNST. We use the state of *[CLS]* token to obtain the representation in the encoder. The dimension of latent z is set to 128 for sentiment-controlled generation and detoxification(2-class) and 256 for topic-controlled generation (4-class). To fuze the latent z better with the Transformer decoder, we use a simplified fusion method of DELLA [40] where we concatenate z to the attention output of each token in each Transformer layer, and then add a linear layer to transfer the new attention output to the original shape of attention output. We did not use the low-rank tensor to compute layer-wise latent z to save the number of parameters. As a common practice [39], we use top-p with $p = 0.9$ sampling method for decoding. To stabilize training, we further incorporate BOW [104] and annealing [28] techniques. Following ST in NLU [73], we start ST from a base model tuned on D_L without any sample selection as in [102].

To avoid KL-vanishing, we utilize cyclical annealing tricks [28] to train DuNST and set the cycle length equal to training steps in each epoch. In each cycle, first the KL weight increases from 0 to 1 linearly for the first 80% steps in a cycle, and keeps to be 1 for the remaining 20% steps. KL annealing is activated for 5 epochs for classification KL-loss and 7 epochs for generation KL-loss. Besides, we use the KL thresholding scheme [52] to give up driving down KL for dimensions of z that are already beneath the target compression rate *KL-lambda*.

We tuned *KL-lambda* $\in \{0.01, 0.03, 0.05, 0.1\}$ (following Li et al. [52]), $\lambda_c \in \{1, 5, 10\}$, the ratio of Pseudo Texts (Fig. 3.6(b)), and softmax temperature $\tau \in \{0.2, 1, 5, 10\}$ (Fig. 3.4) to obtain the reported results. We set *KL-lambda* to be 0.05 for sentiment-controlled generation, 0.03 for detoxification, and 0.01 for topic-controlled generation. λ_c is 10 for sentiment-controlled generation and 1 for topic-controlled generation and detoxification. Softmax temperature τ is set to be 5 for all tasks. For other hyperparameters, λ_g is set to be 1, and weight for BOW loss [104] λ_{bow} is set to be 0.2 for all tasks. We use AdamW [65] as an optimizer. The training batch size is 8 and the learning rate is $5e-5$. We apply linear warmup to the optimizer and the number of warm-up steps is one epoch.

We implement DuNST and all other baselines based on Huggingface Transformers [115] library of v4.21.1 and use NVIDIA A100 to train our model. The total number of training GPU hours is around 8h for IMDb, 10h for Jigsaw, and 9h for AGNews. The number of parameters of our model is 134.56M for sentiment-controlled generation and text detoxification. For a topic-controlled generation, the number of parameters is 136.19M. In the generation phase, we use top-p sampling ($p = 0.9$) as the decoding method. Other configuration of the generator includes a length penalty to be 1.0, a repetition penalty to be 1.0, and a no-repeat-ngram-size to be 4 for all baselines. All experimental results are trained and tested in a single run.

Additional Settings for Detoxification Tasks

As mentioned in 3.3.3, the Jigsaw dataset suffers from severe imbalanced labels where toxic data only counts for 9.6% of training data. To alleviate this problem, we can tune the ratio of toxic and non-toxic data when generating pseudo texts and in conclusion balance the whole training set. We can obtain a less imbalanced dataset if we increase the ratio of toxic to non-toxic data in PT. We propose DuNST(pos) where all pseudo texts are generated from toxic attributes.

Similarly, in the baseline for detoxification tasks, we additionally tested a new variant for GPT2-based self-training methods. GPT2+PT(select, all toxic)+PL refers to all pseudo texts generated from toxic attributes, while GPT2+PT(select)+PL refers to generating 1:1 toxic/non-toxic pseudo texts.

Results

As shown in Table 3.3, Table 3.4, and Table 3.5, on all three tasks, our DuNST achieves significant improvement in controllability compared to fine-tuned PLMs and lightweight tuning and is comparable in fluency, generalizability, and diversity. Fine-tuned PLMs obtain limited F1 improvement but severely decreased diversity (+6.7 S-BLEU at most), indicating they are overfitted to these few labeled data points and fail to cover larger attribute spaces. PF and Ctr-PF only reduce required data but perform even worse than tuned PLMs. The unnatural O-PPL (much lower than that of ground truth) shows they lose the capacity of PLMs and cause degenerated results. In contrast, thanks to the duality, DuNST simultaneously refines pseudo labels and enhances the quality and diversity of pseudo text in an iterative manner, boosting controllability and diversity (*Challenge 1*).

We also have some interesting findings about existing self-training methods. 1) The classic ST method even hurts controllability and generalizability in the sense of *Challenge 2*. As discussed in Sec. 3.1, merely self-generated text over-stresses exploitation of the learned space and hinders exploration.

2) Traditional synthetic noise PT(noise) motivates isotropic exploration, which diverges from valid attribute distributions (poorer O/M-PPL). 3) Sample selection brings a marginal improvement but costs 50% more training time. Thus we did not apply such a selection in DuNST. 4) Additional pseudo-labels significantly improve performance. However, unlike our dual method, the fixed pseudo labels by PT(noise)+PL cannot evolve during ST. By comparison, DuNST utilizes high-temperature sampling and soft text to introduce flexible noise, encouraging safer exploration and better controllability and diversity while maintaining good quality.

As shown in Table 3.5, DuNST outputs the least toxic text while keeping a relatively high diversity. We find that generating all toxic pseudo texts performs better than generating 1:1 toxic/non-toxic pseudo texts for GPT2 and UniLM, which shows that adding pseudo text in self-training can tackle the issue of the imbalanced dataset. The Output PPL and Model PPL of DuNST are larger than the baselines. We explain the reason as follows. Since we are choosing toxic prompts marked as "challenging", it means that toxic sentences would be more likely to be generated and thus have a lower PPL score. Similarly, some non-toxic continuation might get a high PPL score from the GPT2-XL model, since it is rarer to be seen and is less natural from the challenging prompt. This does not mean that generation fluency is worse.

Human Evaluation

To better verify the effectiveness of DuNST, we also conduct a human evaluation. For each model, we generated 100 samples on each task. We invite 5 competent annotators to score these samples on three criteria – **Fluency**, **Novelty**, and **Attribute Relevance**. As shown in Table 3.6, DuNST consistently outperforms all other baselines on all three metrics, which indicates that DuNST not only has better controllability over attributes but also generates fluent and diverse texts. Human evaluation of detoxification tasks demonstrates that DuNST generation does not have a significant difference from UniLM generation in fluency and novelty. On the other hand, its toxicity level is significantly lower than the two baselines, which further demonstrates that DuNST can improve generation controllability.

Ablation Study

We conduct an ablation study on the IMDb dataset. As shown in Table 3.7, we can find: 1) variational learning further enhances control accuracy

	Detoxification				
	O-PPL ↓	M-PPL ↓	Toxic% ↓	Dist ↑	S-BLEU ↓
Test set	48.77	–	–	54.26	32.22
GPT2(raw)	25.06	10397.67	47.4	52.71	37.13
Finetune LM					
GPT2	32.79	66.61	43.94	51.62	42.05
UniLM	52.23	67.92	34.38	38.26	55.31
T5	**27.21**	42.04	22.81	39.83	63.49
Lightweight methods					
PF	28.67	52.73	38.37	49.68	**41.53**
Ctr-PF	29.28	57.39	31.53	49.47	46.70
Self-Training with GPT2					
+PT	36.29	71.47	41.03	**51.91**	42.15
+PT+noise	34.69	66.12	40.59	51.31	43.42
+PT+PL+noise	29.20	26.37	40.99	49.75	43.10
+PT(select)+PL	29.83	26.44	43.45	49.65	43.03
+PT(pos)+PL	29.49	**25.87**	40.00	49.52	43.22
Self-Training with UniLM					
+PT	46.78	74.71	34.68	36.82	55.89
+PT+noise	51.99	80.46	39.46	40.16	52.95
+PT+PL+noise	40.98	55.99	26.95	44.47	47.07
+PT(select)+PL	40.70	54.50	29.21	45.42	46.94
+PT(pos)+PL	45.09	55.87	25.13	45.91	46.70
Our Methods					
DuNST-PT	56.75	50.28	15.32	47.03	47.10
DuNST(pos)	74.74	63.75	**13.69**	50.37	42.62

Table 3.5: Results of DuNST on Jigsaw dataset. DuNST-PT refers to DuNST without pseudo text but only uses pseudo-labeled data.

	Model	Fluency ↑	Novelty ↑	Rel. ↑
	Ctr-PF	3.23**	3.38**	3.37**
Sentiment	ST	3.35	3.65	3.83**
	DuNST	**3.51**	**3.69**	**4.13**
	Ctr-PF	3.66**	4.16**	4.51*
Topic	ST	3.97	4.43	4.57*
	DuNST	**4.01**	**4.50**	**4.71**
	Model	Fluency ↑	Novelty ↑	Toxicity ↓
	Ctr-PF	3.57	**3.88**	2.12**
Detoxification	ST	3.55	3.72	2.40**
	DuNST	**3.58**	3.83	**1.64**

Table 3.6: Human evaluation results of DuNST on sentiment/topic-controlled generation and text detoxicification. ST refers to the best ST variant under automatic evaluation. We conduct Student t-test for statistical significance. Notation: **: p-value<0.01, *: p-value<0.05. The Cohen's kappa score is 0.63, indicating a satisfactory inter-annotator agreement.

	IMDb						
	O-PPL ↓	M-PPL ↓	Acc ↑	F1 ↑	AUC ↑	Dist ↑	S-BLEU ↓
DuNST	21.67	42.82	**92.9**	**93.05**	**98.02**	31.79	**65.80**
−Variational	19.67	38.56	92.11	92.12	97.85	31.39	66.21
−SPT	**18.53**	**36.53**	91.55	91.64	96.97	31.51	67.07
−PT	20.91	41.14	91.7	91.83	96.93	31.67	66.27
−PL	47.45	197.27	83.0	83.41	91.48	32.61	66.17
−PL-SPT	48.56	219.30	81.1	80.85	90.05	**32.86**	66.12
−PL−PT	42.12	147.14	82.7	82.89	91.46	29.75	68.91

Table 3.7: Ablation study on IMDb dataset. PT: pseudo text. SPT: soft pseudo text. PL: pseudo label. The symbol − means removing the settings from DuNST. −VAE reduces to jointly trained classifier and generator. −PL−PT reduces to naive dual variational learning.

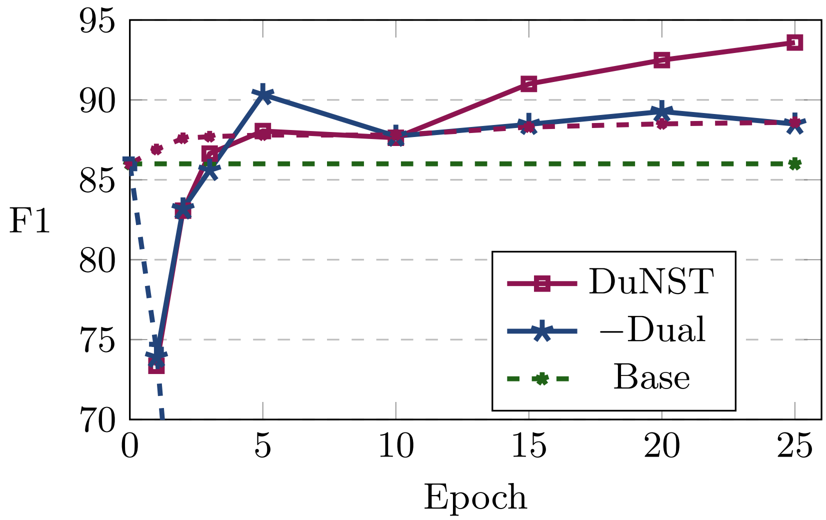


Figure 3.2: F1 score over the number of training epochs on topic. Solid lines indicate generation controllability, while dashed ones refers to classification. The green line is classification F1 of our base model at epoch 0.

and diversity with slight PPL loss, which is worthwhile since the generated text is already fluent enough (close to ground truth PPL). 2) pseudo labels lead to a significant improvement. 3) soft pseudo text outperforms the hard one on controllability and diversity but with marginal fluency loss. Solely hard pseudo text in ST limits model coverage, while the soft one brings a smoother noise and helps push the learned boundary.

Analysis

Effect of Duality: We compare our model with a variant (−**Dual**) where we annotate pseudo labels in advance fix them and cut off classification losses through training. All other settings are retained to be the same. As depicted in Fig. 3.2, since classification and generation share parameters, without optimizing the classifier and pseudo labels, the latent posterior would gradually shift such that the classification performance greatly drops. As a result, generation F1 reaches its maximum soon and stops increasing. On the other hand, thanks to the simultaneously optimized classifier, DuNST keeps improving classification, further distinguishing the posterior and thus enhancing controllability.

Table 3.8 shows the comparison result on the topic generation task. We

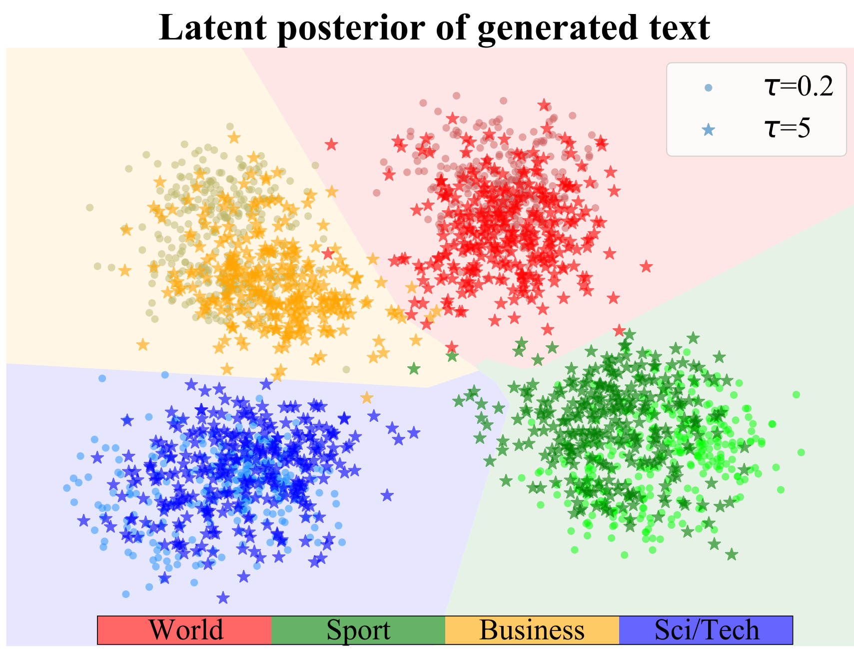


Figure 3.3: VAE posterior of generated texts based on different temperature for topic-controlled generation.

can see that without duality the generation performance drops significantly.

	Output PPL ↓	F1 ↑	Dist ↑
DuNST	**34.73**	**93.59**	**59.42**
−Dual	50.26	90.33	55.83

Table 3.8: Comparison about duality on topic generation.

Effect of Noise (temperature): To illustrate why noise encourages exploration and improves control, we plot the posterior of generations in different temperatures and visualize the simulated decision bound based on training data in Fig. 3.3. We find that higher noise pushes the latent space towards the decision bound and hence leads to more challenging pseudo text. Such data enables the model to learn to better distinguish latent representations under different attributes and push the generalization boundary, thus potentially improving generation controllability. Besides, the

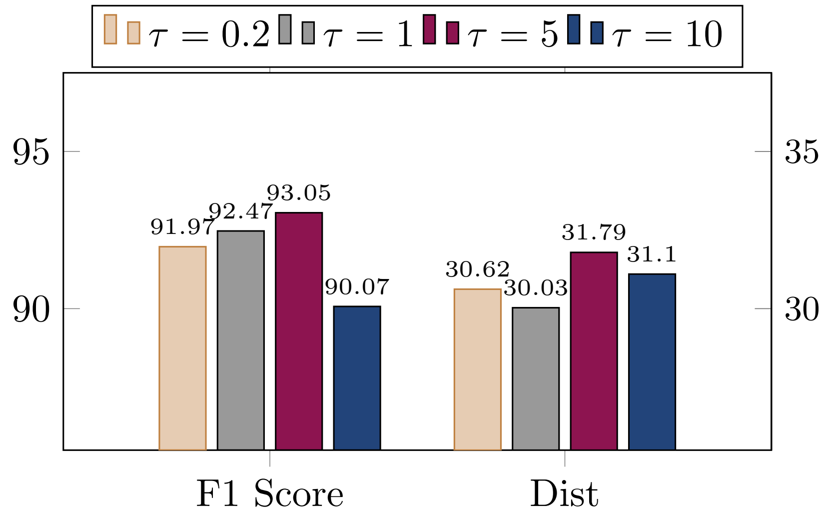


Figure 3.4: esults of DuNST using different level of softmax temperature on IMDb dataset.

noisy pseudo data also helps improve exploration and attribute coverage. Fig. 3.5 depicts the distribution of BERT-large embedding of training data and DuNST-generated data in different temperatures under *World* topic. Here we use the *[CLS]* embedding of the BERT-large model to represent sentence embedding. We find that larger generation temperature leads to more diverse sentence representation, which demonstrates that high-temperature generation of pseudo data could improve generation diversity.

Fig. 3.4 shows the generation performance of DuNST with different temperatures. We achieve the best controllability and diversity when the temperature is 5 in our setting, which indicates a suitable temperature is necessary to balance exploration and exploitation.

Number of labeled data: We show the effectiveness of different models over changing sizes of training data in Fig. 3.6(a) (unlabeled data size remains the same). We find that all models obtain improved generation performance with more labeled data, but our DuNST performs similarly when using only 5% of labeled data compared to 100%. Such results demonstrate the superiority of DuNST which additionally learns from unlabeled and pseudo data through self-training.

Number of pseudo text: We also evaluate DuNST on varying numbers of pseudo text, keeping all the other settings unaltered. As shown in Fig. 3.6(b), DuNST performs best with equal size of pseudo text and labeled data.

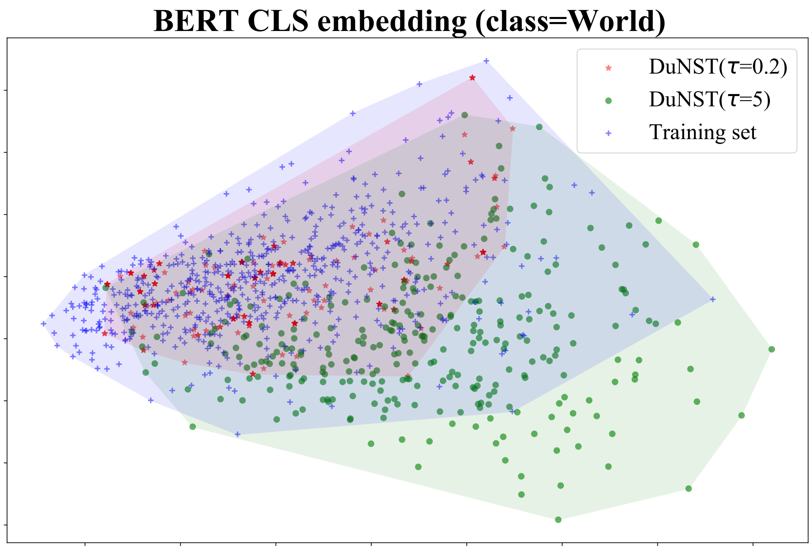


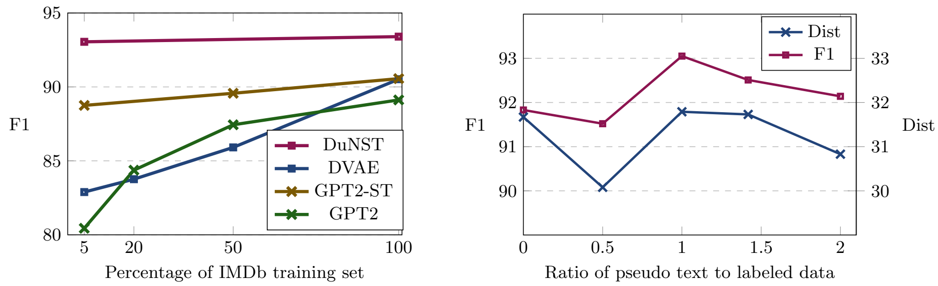
Figure 3.5: BERT *[CLS]* embedding of training texts and generated texts from DuNST model under different temperatures.

More pseudo text brings more noise which hurts generation quality as the model learns more noise than semantics. Too less pseudo text makes the model lose exploration ability and thus fail to extend the learned distribution boundary. Therefore, we should find a suitable noise level to balance exploration and fluency.

Case Study: In order to verify the generation quality and attribute relevance, we present some cases sampled from different models in Table 3.9. We can see that Ctr-PF and GPT2-ST suffer from repeating expressions (e.g., "poor quality" and "got bored"). In contrast, DuNST produces more diverse and fluent expressions, which are also more faithful to the given negative attribute.

3.3.5 Summary

We propose a novel DuNST method to apply Self-training to semi-supervised controllable NLG. DuNST (1) jointly optimizes generation and classification via a dual variational learning framework to leverage both pseudo text and pseudo labels, and (2) incorporates two kinds of soft noise into ST, better exploring larger potential text space and extending the attribute distribution



(a) Number of Labeled Data

(b) Number of Pseudo Text

Figure 3.6: Generation F1 on different number of (1) labeled data and (2) pseudo text on IMDb dataset.

Model	Generation
Ctr-PF	*At 20 years old*, ichorous and unrepentant, ichanarito is a well-liked celebrity, but a poor quality actor, and a poor quality movie...
GPT2-ST	*At 20 years old*, I grew up with it and had no expectations. I was a big fan of the series, I thought it was great, and I wanted to see more of it. But soon enough, I got bored and just got bored watching it.
DuNST	*At 20 years old*, the poorest hour of my life had me scratching my head. This was an insult to the intelligence of mankind, and a complete disregard for anyone else's intelligence. Do not watch this movie. It is a disgrace to any religion.

Table 3.9: Sampled generations with specified negative sentiment and input prompt *At 20 years old*. Words in blue/red are positive/negative indicators, respectively.

boundary. Theoretical analysis demonstrates that DuNST acts as a combination of regularization-like exploitation and attribute boundary exploration, which makes a better balance of the two requirements, significantly improving control accuracy with satisfactory generation fluency and diversity.

3.4 KEST

3.4.1 Contributions

The contributions of KEST are as follows:

- We dig into the over-exploitation problem of applying self-training to controllable NLG and propose a novel kernel-based ST framework to address this problem.

- We design a non-autoregressive generation schema to reduce the time cost of producing pseudo text for self-training, making ST more practical for real scenarios.
- We theoretically show that KEST could explore a larger potential text space and demonstrate through exhaustive experiments that our model significantly improves controllability with competitive generation diversity and quality, further exploring the capacity frontier of PLMs.

3.4.2 Method

Formulation and Overview

Let $\mathbf{x}_i$ denote a textual sequence and y an attribute label. Assume we have a labeled dataset $D_l = \{\mathbf{x}_i, y_i\}_{i=1}^{N_l}$, and an unlabeled in-domain set $D_u = \{\mathbf{x}_i\}_{i=1}^{N_u}$ where $N_u \gg N_l$. Our goal is to learn an attribute-controllable generator $\mathcal{G}_{ag}(y) = P_\theta(\mathbf{x}|y)$ (parameterized by θ) to generate high-quality text $\mathbf{x}$, matching the given label y. In addition, we endow the generator with the ability of multi-task generation. Concretely, the model is reused and jointly trained to generate (a) pseudo text $\hat{\mathbf{x}}$ in a non-autoregressive manner, depicted as $\mathcal{G}_{nag}(y)$, for further augmenting self-training, and (b) pseudo labels $\hat{y}$ for $\mathbf{x} \in D_u$, namely, a classifier $\mathcal{C} = P_\theta(y|\mathbf{x})$.

During the self-training phase, besides the pseudo label pairs $(\mathbf{x}, \hat{y})$, KEST also learns the pseudo text pairs $(\hat{\mathbf{x}}, y)$ in the kernel space to simultaneously cover more unseen instances and extend the previously fitted distribution, *handling Challenge 2*. All the pseudo text $\hat{\mathbf{x}}$ is produced through NAG efficiently, *handling Challenge 3*.

Multi-task Generator

To further enhance the performance and efficiency of our model, we design a multi-task generator to produce the desired text $\mathbf{x}$, *Pseudo Label (PL)* $\hat{y}$, and *Pseudo Text (PT)* $\hat{\mathbf{x}}$ jointly based on a shared PLM.

Autoregressive Text Generation. To obtain high-quality attribute-specified generated text $\mathbf{x}$, we optimize the generator $\mathcal{G}_{ag}$ in an autoregressive manner as follows:

$$\mathcal{L}_{ag} = -\frac{1}{N} \sum_{(\mathbf{x},y)\in D} [\sum_{j=1}^{L} \log P_\theta(\mathbf{x}^j|\mathbf{x}^{<j}, y)], \tag{3.10}$$

where $\mathbf{x}^j$ means the j-th token in $\mathbf{x}$, L is the length of $\mathbf{x}$, D is the training set with N samples. We will show later how to construct D for different training phases.

Pseudo Label Generation. We also make our model simultaneously learn a classifier $\mathcal{C}$ by minimizing:

$$\mathcal{L}_c = -\frac{1}{N} \sum_{(\mathbf{x},y)\in D} \log P_\theta(y|\mathbf{x}). \tag{3.11}$$

Eq. (3.11) enables our model to make full use of available unlabled text $\mathbf{x} \in D_u$ to produce pseudo labels by $\hat{y} = \mathrm{MLP}\,(\mathrm{Encoder}(\mathbf{x}))$, helping regularize the training and improve the generalization bound [112].

Non-autoregressive Pseudo Text Generation. With insufficient unlabeled text, we could produce pseudo text for further improvement and then speed up the repetitive PT generation via NAG. However, as shown in Sec. 3.2, an input consisting of just y is too uninformative to guide the generation, hampering convergence and causing extremely noisy PT.

To mitigate this problem, we resort to the Masked Language Model (MLM) [17] to train the NAG generator $\mathcal{G}_{nag}$ and conduct generation. Define $\mathbf{m} \sim \mathcal{B}(L, p_m)$ as a mask indicator vector, where $\mathcal{B}$ is the Bernoulli distribution. Given a text $\mathbf{x}$, we replace part of the tokens in it with the MASK symbol and get the masked one $\mathbf{x}^{\backslash\mathbf{m}} = [\mathbf{x}^1, \cdots, \mathrm{MASK}, \cdots, \mathbf{x}^L]$, where $\mathbf{x}^j = \mathrm{MASK}$ iff $\mathbf{m}^j = 1$. Then we optimize the following loss for NAG:

$$\mathcal{L}_{nag} = -\frac{1}{N} \sum_{(\mathbf{x},y)\in D} [\sum_{j=1}^{L} \mathbb{I}(\mathbf{m}_j = 1) \log P_\theta(\mathbf{x}^j|\mathbf{x}^{\backslash\mathbf{m}}, y)], \tag{3.12}$$

where $\mathbb{I}$ is the indicator function and the masking probability p_m can be adjusted as the noise level.

In this way, our model only needs to predict partial tokens according to the rich context $\mathbf{x}^{\backslash\mathbf{m}}$, which is easier to learn, reducing the time complexity of PT generation from $\mathcal{O}(L)$ to $\mathcal{O}(1)$ (see Fig. 3.8). Besides, the pseudo text $\hat{\mathbf{x}} = \mathcal{G}_{nag}(\mathbf{x}^{\backslash\mathbf{m}}, y)$ naturally introduces moderate noise in terms of re-predicted tokens while maintaining satisfactory fluency due to the unaltered high-quality ones. Such a flexible corruption acts as a kind of weak augmentation [10] which enhances the exploitation and outperforms typical synthetic noise (*e.g.*, token dropout) [37].

The final loss is computed as follows:

$$\mathcal{L} = \lambda_c\mathcal{L}_c + \lambda_{ag}\mathcal{L}_{ag} + \lambda_{nag}\mathcal{L}_{nag} \tag{3.13}$$

where λ_c, λ_{ag}, and λ_{nag} are hyper-parameters.

Kernel-based Learning

As we discussed in Sec. 3.1, learning from self-generated pseudo text $\hat{\mathbf{x}}$ with standard cross-entropy loss would force the current model P_θ over-exploit and be shackled to the previously learned one $P_{\theta'}$ (Sec. 2), resulting in a shrunken generalization boundary and decreased controllability.

To break such constraints, we make the current model P_θ directly fit the previous one $P_{\theta'}$. For this goal, we make use of *Maximum Mean Discrepancy (MMD)* [32], a well-known kernel-based probability measure, and minimize the following empirical loss for all generated pseudo text:

$$\begin{aligned}\mathcal{L}_{ker} = & \frac{1}{N(N-1)} \sum_{\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j \in D_o, i \neq j} k(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \\ & - \frac{2}{N^2} \sum_{\tilde{\mathbf{x}}_i \in D_o, \hat{\mathbf{x}}_j \in D_{pt}} k(\tilde{\mathbf{x}}_i, \hat{\mathbf{x}}_j),\end{aligned} \tag{3.14}$$

where $D_{pt} = \{\hat{\mathbf{x}}_i\}_{i=1}^N$ is set of pseudo text, $D_o = \{\tilde{\mathbf{x}}_i\}_{i=1}^N$ is set of text generated by $\mathcal{G}_{ag}(\hat{\mathbf{x}}_i, y)$ (or $\mathcal{G}_{nag}(\hat{\mathbf{x}}_i^{\backslash \mathbf{m}}, y)$) in the self-training phase. k is the kernel function, for which we take the RBF kernel here, that is, $k(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) = \exp\left(\frac{-\|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j\|^2}{2\sigma^2}\right)$ and σ is the bandwidth.

This MMD loss is an unbiased U-statistic estimator, which can be used in conjunction with stochastic gradient descent (SGD) methods. We will demonstrate in Sec. 2 that such an objective could relax the constraint imposed by the previous model $P_{\theta'}$ and encourage more diverse outputs.

Soft Pseudo Text (SPT) When optimizing Eq. (3.14), we need to calculate the l_2-distance between two text $\mathbf{x}_i$ and $\mathbf{x}_j$. Simply using hard text (one-hot representations) has two drawbacks. First, the signal would be too sparse since most dimensions are zeros in the vector. Second, the sampled discrete $\mathbf{x}_i$ (a point in the text space) causes information loss and forces us to sample numerous points to cover a small neighborhood region in the space.

Therefore, we further propose to generate soft pseudo text. We use the feature representation of the text $\mathbf{x}$, $e(\mathbf{x}) = P(\mathbf{x}) \times \mathbf{E} \in \mathbb{R}^{L \times d}$, where $P(\mathbf{x}) \in \mathbb{R}^{L \times V}$ are the generation probabilities of each token $\mathbf{x}^i$ on the vocabulary, and $\mathbf{E} \in \mathbb{R}^{V \times d}$ is the word embedding matrix. V and d are vocabulary and embedding sizes, respectively. Then we change Eq.(3.10) and Eq. (3.12) to:

$$\begin{aligned}\mathcal{L}_{ag}^{'} &= \mathcal{L}_{ker} \text{ if } \mathbf{x} \in D_{pt} \text{ else } \mathcal{L}_{ag} \\ \mathcal{L}_{nag}^{'} &= \mathcal{L}_{ker} \text{ if } \mathbf{x} \in D_{pt} \text{ else } \mathcal{L}_{nag}.\end{aligned} \tag{3.15}$$

Algorithm 2: Training Process of KEST

Input: Labeled set D_l, unlabeled set D_u
Output: The trained model P_θ

Jointly train base model $\mathcal{G}_{ag}$, $\mathcal{G}_{nag}$, $\mathcal{C}$ on D_l by optimizing Eq.(3.13), store the best $\mathcal{G}^0_{ag}$, $\mathcal{G}^0_{nag}$, $\mathcal{C}^0$.
for *epoch* $\leftarrow 1$ to $MaxEpoch$ **do**
  for $\mathbf{x}_i$ **in** D_u **do**
    $\hat{y}_i = \mathcal{C}^{epoch-1}(\mathbf{x}_i)$
  end for
  Build pseudo labeled dataset $D_{pl} = \{\mathbf{x}_i, \hat{y}_i\}$
  Sample a subset D_{pseudo} from $D_l \cup D_{pl}$
  for $(\mathbf{x}_i, y_i)$ **in** D_{pseudo} **do**
    Sample mask vector $\mathbf{m}$.
    $\hat{\mathbf{x}}_i = \mathcal{G}^{epoch-1}_{nag}(\mathbf{x}_i^{\setminus \mathbf{m}}, y_i)$
  end for
  Build pseudo text dataset: $D_{pt} = \{\hat{\mathbf{x}}_i, y_i\}$
  Train $\mathcal{G}^{epoch-1}_{ag}$, $\mathcal{G}^{epoch-1}_{nag}$, and $\mathcal{C}_{epoch-1}$ on $\{D_{pt}, D_{pl}, D_l\}$ by optimizing Eq.(3.13) and Eq.(3.15), update the parameter to $\mathcal{G}^{epoch}_{ag}$, $\mathcal{G}^{epoch}_{ag}$, and $\mathcal{C}^{epoch}$.
end for

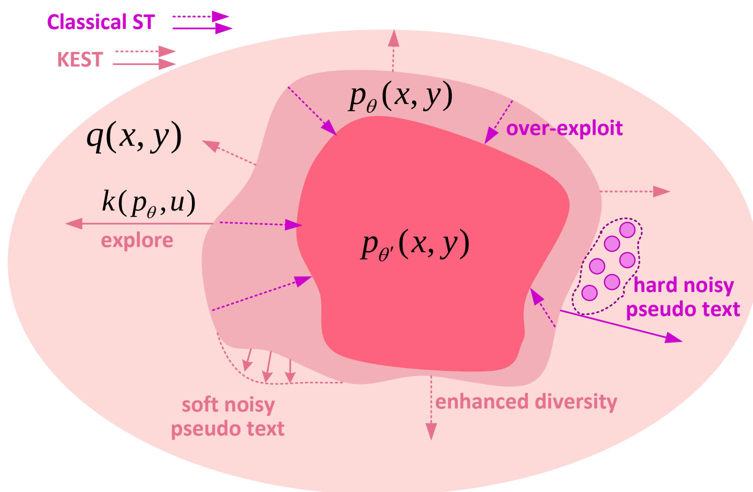


Figure 3.7: The illustration of KEST advantages.

In this way, we avoid losing relevant semantics information in the pseudo text, make the model fit a smoother distribution and further extend the generalization boundary (see Table 3.14).

Following the practice of self-training in NLU [102], we start ST from a strong base model tuned on D_l and use the full unlabeled D_u to produce pseudo labels, rather than select part of the data with certain criteria as in [46, 73]. The PLM word embedding $\mathbf{E}$ is frozen during self-training. The complete KEST process is described in Alg. 2.

Further Analysis of KEST

To better understand the advantages of KEST, we provide the following two results.

Lemma 1. *The optimization of classical self-training is equivalent to minimizing* $(1-\alpha) * KL[Q(x,y)||P_\theta(x,y)] + \alpha * KL[P_{\theta'}(x,y)||P_\theta(x,y)]$, *where* Q *is the real joint distribution of text and label,* P_θ *and* $P_{\theta'}$ *are models estimated at the current and last ST iteration, respectively,* KL *is the Kullback–Leibler divergence, and* α *is the ratio of pseudo text.*

Proof of Lemma 1: At each self-training iteration, define the set of real labeled data with N samples as D, and that of generated pseudo data with M samples as $\hat{D}$, the current model as $P_\theta(\mathbf{x})$. For brevity, we omit the label y here. For classical self-training, we minimize the following objective:

$$
\begin{aligned}
&\text{minimize} - \frac{1}{N+M} \sum_{\mathbf{x} \in D \cup \hat{D}} \log P_\theta(\mathbf{x}) \\
&= -\frac{1}{N+M} \left[\sum_{\mathbf{x} \in D} \log P_\theta(\mathbf{x}) + \sum_{\mathbf{x} \in \hat{D}} \log P_\theta(\mathbf{x}) \right] \\
&= -\frac{N}{N+M} \sum_{\mathbf{x} \in D} Q(\mathbf{x}) \log P_\theta(\mathbf{x}) - \frac{M}{N+M} \sum_{\mathbf{x} \in \hat{D}} P_{\theta'}(\mathbf{x}) \log P_\theta(\mathbf{x}) \\
&\approx \frac{N}{N+M} H[Q(\mathbf{x}), P_\theta(\mathbf{x})] + \frac{M}{N+M} H[P_{\theta'}(\mathbf{x}), P_\theta(\mathbf{x})] \\
&= \frac{N}{N+M} H[Q(\mathbf{x}), P_\theta(\mathbf{x})] + \frac{M}{N+M} H[P_{\theta'}(\mathbf{x}), P_\theta(\mathbf{x})] \\
&- \frac{N}{N+M} H[Q(\mathbf{x})] - \frac{M}{N+M} H[P_{\theta'}(\mathbf{x})] \\
&= (1-\alpha) * \mathrm{KL}[Q(\mathbf{x}), P_\theta(\mathbf{x})] + \alpha * \mathrm{KL}[P_{\theta'}(\mathbf{x}), P_\theta(\mathbf{x})] + C,
\end{aligned}
\tag{3.16}
$$

where $Q(\mathbf{x}) = \frac{1}{N}\mathbb{I}(\mathbf{x} \in D)$ is the empirical distribution of real text and $P_{\theta'}(\mathbf{x}) = \frac{1}{M}\mathbb{I}(\mathbf{x} \in \hat{D})$ is the empirical distribution of the model learned at the last self-training iteration, formed by previously generated pseudo samples. H is the entropy $H[Q(x), P(x)] = -\int Q(x) \log P(x) dx$, and we have $H[P_{\theta'}(\mathbf{x})] = H[Q(\mathbf{x})] = 0$. $\alpha = \frac{M}{N+M}$ is the ratio of pseudo data, concluding the proof. □

From Lemma 1, we can see that classical ST approximates the text distribution and fits the current model into the previously learned one. Since $KL[P_{\theta'}(x,y)||P_\theta(x,y)] = \iint P_{\theta'}(x,y) \log \frac{P_{\theta'}(x,y)}{P_\theta(x,y)} dxdy$, failing to assign enough probability mass to a point (x,y) in $P_{\theta'}$ will bring extremely larger loss. Consequently, P_θ is more inclined to cover $P_{\theta'}$ rather than explore Q, causing over-exploitation.

In contrast, we give a theorem of our KEST:

Theorem 2. *Minimizing the training objective of KEST is equivalent to minimizing the following:*

$$
\begin{aligned}
&KL\left[Q(x,y)||P_\theta(x,y)\right] \\
&+MMD^2\left[P_{\theta'}(x,y)||P_\theta(x,y)\right] \\
&-2 * \mathbb{E}_{P_\theta U}\left[k(x,u)\right],
\end{aligned}
\tag{3.17}
$$

where U is a noise distribution.

Proof of Theorem 2: From Lemma 1, we can see that learning the real text $\mathbf{x} \sim Q(\mathbf{x})$ only involves the first KL term. Thus, we focus on the second term, $\mathrm{KL}[P_{\theta'}(\mathbf{x})||P_\theta(\mathbf{x})]$ here, which is replaced by Eq.(3.14). We further rewrite Eq.(3.14) as:

$$\begin{aligned}&\frac{1}{N(N-1)}\sum_{\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j\in D_o, i\neq j} k(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j) - \frac{2}{N^2}\sum_{\tilde{\mathbf{x}}_i\in D_o,\hat{\mathbf{x}}_j\in D_{pt}} k(\tilde{\mathbf{x}}_i,\hat{\mathbf{x}}_j)\\ &\approx \mathbb{E}_{P_\theta(\tilde{\mathbf{x}})}[k(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j)] - 2\mathbb{E}_{P_\theta(\tilde{\mathbf{x}}),P_{\theta'}(\hat{\mathbf{x}})}[k(\tilde{\mathbf{x}}_i,\hat{\mathbf{x}}_j)].\end{aligned} \tag{3.18}$$

Since the previously learned model $P_{\theta'}$ is fixed in this iteration, minimizing Eq.(3.14) is equal to minimizing:

$$\begin{aligned}&\mathbb{E}_{P_\theta(\tilde{\mathbf{x}})}[k(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j)] + \mathbb{E}_{P_{\theta'}(\hat{\mathbf{x}})}[k(\hat{\mathbf{x}}_i,\hat{\mathbf{x}}_j)] - 2\mathbb{E}_{P_\theta(\tilde{\mathbf{x}})P_{\theta'}(\hat{\mathbf{x}})}[k(\tilde{\mathbf{x}}_i,\hat{\mathbf{x}}_j)]\\ =&\mathbb{E}_{P_\theta(\tilde{\mathbf{x}})}[<\varphi(\tilde{\mathbf{x}}_i),\varphi(\tilde{\mathbf{x}}_j)>_{\mathcal{H}}] + \mathbb{E}_{P_{\theta'}(\hat{\mathbf{x}})}[<\varphi(\hat{\mathbf{x}}_i),\varphi(\hat{\mathbf{x}}_j)>_{\mathcal{H}}]\\ &-2\mathbb{E}_{P_{\theta'}(\hat{\mathbf{x}}),P_\theta(\tilde{\mathbf{x}})}[<\varphi(\hat{\mathbf{x}}_i),\varphi(\tilde{\mathbf{x}}_j)>_{\mathcal{H}}]\\ =&<\mu_{P_\theta},\mu_{P_\theta}>_{\mathcal{H}} + <\mu_{P_{\theta'}},\mu_{P_{\theta'}}>_{\mathcal{H}} -2<\mu_{P_\theta},\mu_{P_{\theta'}}>_{\mathcal{H}}\\ =&\mathrm{MMD}^2(P_{\theta'},P_\theta),\end{aligned} \tag{3.19}$$

where $\varphi(\cdot)\in\mathcal{H}$ is that feature map, $<\cdot,\cdot>_{\mathcal{H}}$ is the dot product in the reproducing kernel Hilbert space $\mathcal{H}$, and $\mu_{P_\theta} = \mathbb{E}_{P_\theta}[\varphi(\tilde{\mathbf{x}})]$.

However, in our method KEST, we don't use the exact $P_{\theta'}$, but generate the noisy pseudo $\hat{\mathbf{x}}$ by NAG. Therefore, we could consider the previously learned distribution as a noisy one $P_{\theta'}+U$ by incorporating a noise distribution U, and get:

$$\begin{aligned}&\mathrm{MMD}^2(P_{\theta'}+U,P_\theta)\\ =&||\mu_{P_{\theta'}}+\mu_U-\mu_{P_\theta}||^2_{\mathcal{H}}\\ =&||\mu_{P_\theta}||^2_{\mathcal{H}}+||\mu_{P_{\theta'}}||^2_{\mathcal{H}}+||\mu_U||^2_{\mathcal{H}}-2<\mu_{P_{\theta'}},\mu_{P_\theta}>_{\mathcal{H}}\\ &-2<\mu_U,\mu_{P_\theta}>_{\mathcal{H}}+2<\mu_{P_{\theta'}},\mu_U>_{\mathcal{H}}\\ =&\mathrm{MMD}^2(P_\theta,P_{\theta'})+||\mu_U||^2_{\mathcal{H}}+2<\mu_{P_{\theta'}},\mu_U>_{\mathcal{H}}-2<\mu_U,\mu_{P_\theta}>_{\mathcal{H}}.\end{aligned} \tag{3.20}$$

Again, as $P_{\theta'}$ and U are fixed now, we can omit corresponding terms. Combining Lemma 1, optimizing the objective of KEST is equivalent to minimizing:

$$\begin{aligned}&\mathrm{KL}[P_{\theta'}||P_\theta]+\mathrm{MMD}^2(P_{\theta'},P_\theta)-2<\mu_U,\mu_{P_\theta}>_{\mathcal{H}}\\ =&\mathrm{KL}[P_{\theta'}||P_\theta]+\mathrm{MMD}^2(P_{\theta'},P_\theta)-2\mathbb{E}_{P_\theta,U}[k(\tilde{\mathbf{x}},u)],\end{aligned} \tag{3.21}$$

concluding the proof. □

In Theorem 1, our KEST fits the true distribution Q by KL divergence to cover the real space as large as possible while fitting $P_{\theta'}$ with MMD (See Fig. 3.7). Considering Eq.(3.14), we can see this loss not only regularizes P_θ by $P_{\theta'}$, but also diversifies P_θ via increasing the l_2-distance of generated text $\|\mathbf{x}_i - \mathbf{x}_j\|^2$, and enhances exploration through fitting a noise distribution and disturbing P_θ, further pushing the generalization boundary.

	Sentiment						
	O-PPL ↓	M-PPL ↓	Acc ↑	F1 ↑	AUC ↑	Dist ↑	S-BLEU ↓
GPT2(raw)	13.20	38.39	55.9	68.5	61.37	35.91	58.79
Finetune LM							
GPT2	**16.40**	44.02	77.55	80.44	88.35	26.34	71.00
UniLM	25.20	54.33	76.45	75.35	85.18	31.05	66.97
T5	25.69	34.97	82.80	83.77	90.50	30.03	69.57
Self-Training with UniLM							
PT	26.62	58.37	72.20	70.27	80.37	31.17	66.69
PT(noise)	30.28	62.07	77.75	75.78	85.35	31.68	**65.18**
PT(noise)+PL	18.92	**33.53**	89.95	89.73	96.38	30.94	66.84
PT(select)+PL	18.40	33.56	90.08	90.06	96.66	31.27	67.61
Our Methods							
KEST	20.65	38.15	**92.10**	**91.77**	**97.06**	**31.70**	66.60

Table 3.10: Automatic evaluation results of KEST on IMDb dataset (sentiment)

3.4.3 Experiments

The tasks, datasets, baselines, and evaluation metrics are described in Sec. 3.3.3.

Experimental Settings

We use pre-trained UniLM-base-cased [18] as the encoder and decoder of our KEST model since UniLM shares the parameter of transformer blocks in the encoder and decoder, more suitable for our joint classification and generation schema. The label embedding dimension is set to 128. To fuse the label embedding better with the Transformer decoder, we concatenate the label embedding to the attention output of each token in each Transformer

	Topic						
	O-PPL ↓	M-PPL ↓	Acc ↑	F1 ↑	AUC ↑	Dist ↑	S-BLEU ↓
GPT2(raw)	16.94	74.41	55.75	52.17	83.28	46.88	45.55
Finetune LM							
GPT2	**22.22**	23.46	82.92	83.08	95.23	54.93	39.93
UniLM	55.79	36.28	87.67	87.70	96.30	54.76	43.77
T5	48.33	32.12	88.33	88.43	97.95	**58.06**	**37.01**
Self-Training with UniLM							
PT	57.40	40.95	86.42	86.36	96.69	52.35	46.41
PT(noise)	58.59	45.32	85.42	85.27	95.88	53.35	46.57
PT(noise)+PL	32.36	**16.64**	89.67	89.70	98.11	53.79	47.95
PT(select)+PL	33.23	16.66	90.50	90.52	98.31	53.71	47.69
Our Methods							
KEST	31.19	20.46	**91.92**	**91.94**	**98.34**	56.16	42.10

Table 3.11: Automatic Evaluation of KEST on AGNews dataset (topic).

	Detoxification				
	O-PPL ↓	M-PPL ↓	Toxic% ↓	Dist ↑	S-BLEU ↓
GPT2(raw)	25.06	10397.67	47.40	52.71	37.13
Finetune LM					
GPT2	32.79	66.61	43.94	51.62	42.05
UniLM	52.23	67.92	34.38	38.26	55.31
T5	**27.21**	42.04	22.81	39.83	63.49
Self-Training with UniLM					
PT	46.78	74.71	34.68	36.82	55.89
PT(noise)	51.99	80.46	39.46	40.16	52.95
PT(noise)+PL	40.98	55.99	26.95	44.47	47.07
PT(select)+PL	40.70	54.50	29.21	45.42	46.94
Our Methods					
KEST	66.74	53.42	**18.37**	51.17	**40.71**

Table 3.12: Automatic evaluation results of KEST on Jigsaw dataset.

layer and then add a linear layer to transfer the new attention output to the original shape of the attention output.

We tuned $\lambda_c \in \{1, 5, 10\}$, $\lambda_{nag} \in \{0.5, 1\}$, $p_m \in \{0.3, 0.5, 0.7\}$ to obtain the reported results. Finally we set $\lambda_c = \lambda_{ag} = \lambda_{nag} = 1$ in Eq. (3.13) and $p_m = 0.7$ for all tasks. We use AdamW [65] as an optimizer. The training batch size is 8, and the learning rate is $5e-5$. We apply linear warmup to the optimizer, and the number of warm-up steps is one epoch. For the MMD kernel, we use the median heuristic, where σ is chosen from $(2^a H_N)_{a=-M}^{M}$. Here $H_N = \frac{1}{N(N-1)} \sum_{\tilde{\mathbf{x}}_i \in D_o, \hat{\mathbf{x}}_j \in D_{pt}} \|\tilde{\mathbf{x}}_i - \hat{\mathbf{x}}_j\|_{L_2}^2$ is the median heuristic. In our experiment, M is set to 2.

We implement KEST and all other baselines based on Huggingface Transformers [115] library of v4.21.1 and use one NVIDIA RTX 3090 node to train our model. The total number of training GPU hours is around 19.12h for IMDb, 10.18h for Jigsaw, and 9.34h for AGNews. The number of parameters of our model is 132.65M. In the generation phase, we use top-p sampling (p=0.9) as the decoding method. Other configuration of the generator includes a length penalty to be 1.0, a repetition penalty to be 1.0, and a no-repeat-ngram-size to be 4 for all baselines. All experimental results are trained and tested in a single run with fixed random seeds.

Results

As shown in Table 3.10 and Table 3.11, and Table 3.12, in all of the three tasks, our KEST achieves significant improvement in control accuracy (+8.0 F1 at most) compared to fine-tuned PLMs. The generally much higher PPL (for UniLM and T5), limited F1 improvement, and severely decreased diversity (for GPT2) indicate these PLMs either fail to be adapted to new domains (*e.g.*, positive movie reviews) or overfit with inadequate labeled data, as analyzed in [134]. On the contrary, thanks to the self-augmented data, KEST notably enhances controllability as well as fluency and diversity, especially compared to the backbone UniLM.

We also observed some interesting results considering existing self-training methods. 1) The naive self-training with PT performed poorly in controllability and diversity, even worse than tuned PLMs, due to the over-exploitation and shrunken distributions as interpreted in Sec. 3.4.2. 2) The traditional synthetic noise (PT(noise)) slightly boosts control accuracy and diversity, which verifies the effectiveness of noise [37] again, but greatly hurts fluency and generalizability (+3.7 O-PPL at most). This is because such hard corruption is too noisy and makes the model diverge far from valid attribute distributions. In contrast, KEST utilizes a NAG generator to produce flexible noise, improving local smoothness. 3) Additional pseudo-labels bring significant improvement, especially on PPL. However, with a fixed number of unlabeled data, the performance of these methods is limited. Besides, KEST utilizes the multi-task generator to produce soft pseudo text in feature space, which helps cover a larger attribute space and obtain further improvement.

Human Evaluation

	Model	Fluency ↑	Novelty ↑	Rel. ↑
Sentiment	UniLM-PT(select)+PL	3.60	3.40	3.62**
	KEST	**3.67**	**3.48**	**3.87**
Topic	UniLM-PT(select)+PL	3.97**	4.43	4.57
	KEST	**4.11**	**4.54**	**4.62**
	Model	Fluency ↑	Novelty ↑	Toxicity ↓
Detoxification	UniLM-PT(select)+PL	3.38	3.93	2.75**
	KEST	**3.40**	**3.97**	**2.33**

Table 3.13: Human evaluation results of KEST on sentiment/topic-controlled generation and text detoxification. We conduct the Student t-test to evaluate statistical significance (**: p-value< 0.01). The overall Cohen's kappa score is 0.62, showing a satisfactory inter-annotator agreement.

To better verify the effectiveness of KEST, we also conduct a human evaluation. For each model, we generated 100 samples on each task. We invite 6 competent annotators to score these samples on three criteria – **Fluency**, **Novelty**, and **Attribute Relevance** in a blind review manner. As shown in Table 3.13, KEST consistently outperforms the best baseline (UniLM-PT(select)+PL) on all three metrics, which indicates that KEST not only has better controllability over attributes but also generates fluent and diverse texts.

Ablation Study

We conduct an ablation study on the AGNews dataset and compare different KEST variants. As shown in Table 3.14, we can find: 1) Soft pseudo text prominently improves PPL and diversity, outperforming the hard one. As discussed in Sec. 3.4.2, such soft PT could bring smoother noise and help further push the learned distribution boundary. 2) Kernel-based learning

	AGNews						
	O-PPL ↓	M-PPL ↓	Acc ↑	F1 ↑	AUC ↑	Dist ↑	S-BLEU ↓
KEST	**31.19**	**20.46**	**91.92**	**91.94**	**98.34**	**56.16**	**42.10**
−Soft	38.04	29.07	90.92	90.96	98.07	55.40	44.09
$-\mathcal{L}_{ker}-$Soft	38.98	28.77	90.72	90.81	97.98	54.97	45.02
$-\mathcal{L}_{nag}-\mathcal{L}_{ker}-$Soft	39.73	28.58	90.33	90.42	97.74	55.07	44.73
−PT	38.09	28.77	90.92	90.97	98.11	55.48	44.13
−PL−PT	37.24	256.66	87.41	87.45	96.18	43.18	69.30

Table 3.14: Ablation study of KEST on AGNews dataset. The symbol − means removing the settings from KEST. −Soft: using sampled hard tokens instead of the soft $e(\mathbf{x})$. $-\mathcal{L}_{ker}$: using the cross-entropy loss instead of Eq.(3.14). $-\mathcal{L}_{nag}$: using $\mathcal{G}_{ag}$ to generate pseudo text instead of $\mathcal{G}_{nag}$. −PT/−PL: do not use pseudo text/labels.

alleviates the over-exploitation problem of the traditional cross-entropy loss and further enhances inner-group diversity (-0.93 S-BLEU), empirically supporting Theorem 1. 3) KEST's NAG ability not only reduces time complexity but also slightly boosts fluency (-0.75 O-PPL). However, the diversity improvement attributed to NAG is correlated to the noisy level. Only with an appropriate masking probability p_m could NAG facilitate more diverse text (see Fig. 3.10). Besides, pseudo text notably promotes all metrics, verifying our claim in Sec. 3.4.1 that such synthetic pseudo text leads to further improvement beyond pseudo labels.

Analysis

Time Consumption: Fig. 3.8 shows the decoding time of our NAG and AG generators for generating pseudo text with different text lengths. We found that the time costs of the AG module increase almost linearly w.r.t. the text length. In comparison, our NAG generator $\mathcal{G}_{nag}$ greatly accelerates the generation of pseudo text, especially when the sequence length is long. Furthermore, we compare the training time of KEST using $\mathcal{G}_{ag}$ and $\mathcal{G}_{nag}$, respectively. We observe that the latter achieves *1.2*× and *1.3*× speedup on IMDb and AGNews, respectively, which could be further improved with a larger ratio of pseudo text, making self-training more practical.

Effect of Kernel-based Learning: To analyze the effect of our kernel distance loss $\mathcal{L}_{ker}$ in Eq. (3.14), we train two models for 5 epochs with only pseudo text given the same prompt and starting checkpoint using the kernel loss and the traditional cross-entropy loss, respectively. We then

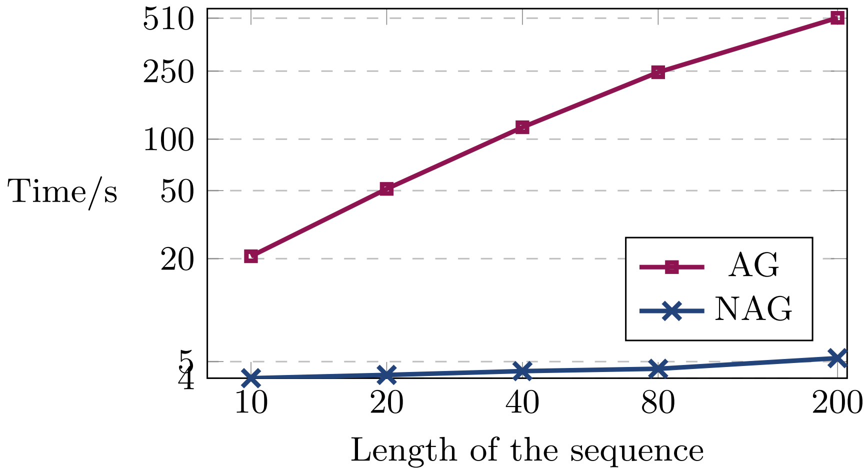


Figure 3.8: Comparison of decoding time of NAG and AG for 100 pseudo text batches (batch size=8) with different text lengths.

visualize the text generated with given prompts from the two models by using corresponding BERT-large *[CLS]* embedding as text representations and plot them. As depicted in Fig. 3.9. We can find that the model trained with cross-entropy loss collapses in a smaller space than the training data space. In contrast, the one with kernel loss successfully extends the learned distribution, which helps explore a larger potential space towards the real one, corroborating our claim and theoretical analysis.

Effect of random mask ratio: The random mask ratio p_m is a hyper-parameter that can control the noise level of generated pseudo text. Fig. 3.10 shows the generation performance of KEST with different mask ratios in the AGNews dataset. We find that a higher ratio leads to a more noisy and diverse generation. A moderately higher ratio also generally improves controllability. However, an extremely high p_m brings too much noise and hence obstructs learning. We achieve the best controllability with $p_m = 0.7$, indicating a suitable mask ratio is necessary to balance exploration and exploitation.

Number of pseudo text: We evaluate KEST on varying numbers of pseudo text, keeping all the other settings unchanged. As shown in Fig. 3.11, KEST performs the best with equal size of pseudo text and labeled data (Ratio = 1). More pseudo text brings more noise which hurts generation quality as the model captures more meaningless noise than semantics. Too little pseudo text makes the model lose exploration ability and thus fail to

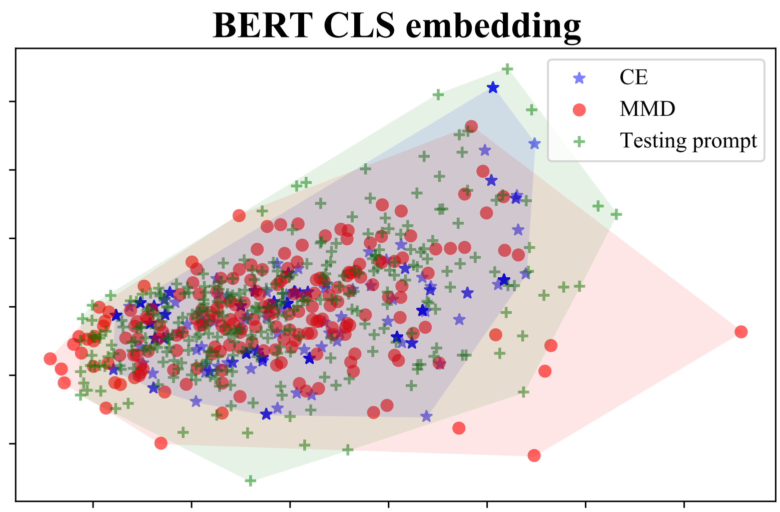


Figure 3.9: BERT *[CLS]* embedding of generated texts from KEST using cross-entropy (CE) and our MMD loss $\mathcal{L}_{ker}$ respectively.

extend the learned distribution boundary, causing poor control accuracy and diversity. Therefore, a suitable ratio is essential to balance exploration and fluency.

Case Study: In order to verify the generation quality and attribute relevance, we present some cases sampled from different models in Table 3.15. We can see that traditional ST methods (UniLM+PT(select)+PL) suffer from repeating phrases (e.g., “love story” and “not like”), exhibiting poor generation diversity and novelty. In contrast, KEST produces more diverse expressions thanks to kernel-based learning and smoother soft pseudo text while staying faithful to the given positive attribute.

3.4.4 Summary

We propose a novel KEST method to incorporate Self-training into semi-supervised controllable NLG. KEST (1) applies a practical multi-task generator to generate soft pseudo text in parallel, significantly reducing decoding time while injecting soft noise to the text; (2) uses soft kernel-based loss to encourage exploration of the learned distribution and increase control accuracy and generation diversity. Theoretical analysis and empirical experiments demonstrate that KEST acts as a combination of regularization-like exploitation and attribute boundary exploration, improving control accuracy

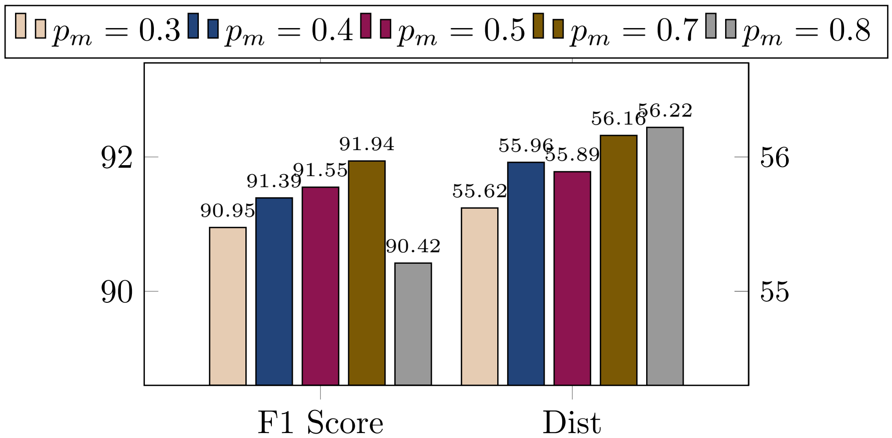


Figure 3.10: Results of KEST with different levels of mask ratio in AGNews Dataset.

with satisfactory generation fluency, diversity, and accelerated training.

3.5 Example of Generation

We sample some generated texts based on Ctr-PF, GPT2-ST, UniLM-ST, DuNST, and KEST and place them on Table 3.16, 3.17, and Table 3.18. Due to the offensive content generated by these models in the detoxification task, we do not include their examples in detoxification experiments in this thesis. From these examples we can find out that DuNST and KEST significantly outperforms the baselines in generation controllability and diversity.

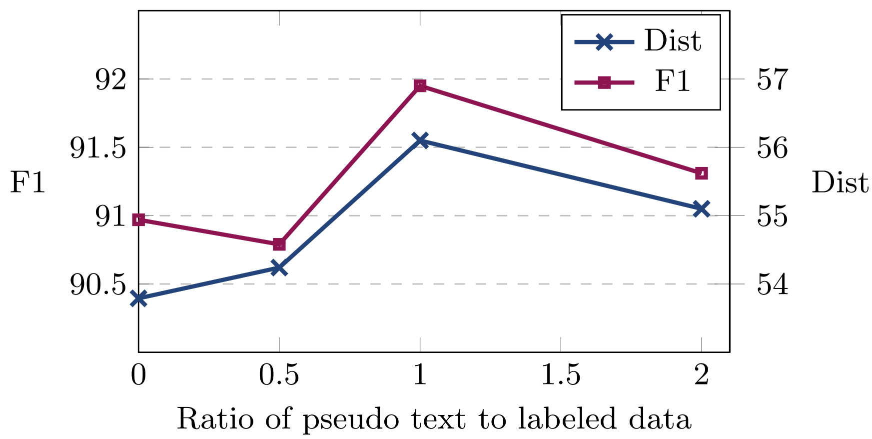


Figure 3.11: Generation controllability (F1) on a different number of pseudo text on AGNews dataset.

Model	Generation
	Sentiment: *Negative*; **Prompt**:*At 20 years old*
Ctr-PF	*At 20 years old*, ichorous and unrepentant, ichanarito is a well-liked celebrity, but a poor quality actor, and a poor quality movie...
GPT2-ST	*At 20 years old*, I grew up with it and had no expectations. I was a big fan of the series, I thought it was great, and I wanted to see more of it. But soon enough, I got bored and just got bored watching it...
UniLM-ST	*At 20 years old*, I didn't expect much when I heard this movie was being made. My initial reaction was that it was an art - house movie (a very low budget movie at that), but when I saw how they did it, it just got boring and boring...
DuNST	*At 20 years old*, the poorest hour of my life had me scratching my head. This was an insult to the intelligence of mankind, and a complete disregard for anyone else's intelligence. Do not watch this movie. It is a disgrace to any religion...

KEST	*At 20 years old*, I remember seeing this film in an old building. The film is a little weird. One of the main characters, an aged man who plays guitar plays strange music and the other is a suave young man who lives with his mother in the same building. By the end of the film, I was a bit confused because I don't remember seeing any of the main actors or their playing the roles...
	Sentiment: *Positive*; **Prompt**:*At 20 years old*
Ctr-PF	*At 20 years old*, ixxoxx was a fan favorite and loved his work and movies like it was his own and his own. A hard worker who enjoyed getting his hands dirty and trying to figure out the best ways to make a buck, ixixx made his name on film and television...
GPT2-ST	*At 20 years old*, ixTch was my first exposure to the French film industry. I immediately knew the "circles" of French cinema, where each character had a particular expression in their mind, but the film itself was only a very short one. I was not interested in the story or the plot, just in the characters...
UniLM-ST	*At 20 years old*, it is still one of my favorite films and the music still rings true. I have seen it countless times and each time I hear it, it gets better and better...
DuNST	*At 20 years old*, I was 13 when this came out. I remember it as being a typical 1950's soap opera. The writing was excellent. It was realistic, the characters were believable. The story was accurate, but the real problems the family had with adult life in particular were out to bother with...
KEST	*At 20 years old*, I have been watching this movie every week, and this is it. I watched it with my younger brother and we enjoyed it. We had always wanted to see a version of the " Little Red Riding Hood " story, but no one had ever heard of it. When I heard about the movie, we both looked to the theater and found ourselves falling in love with it. We both agreed that it was the perfect movie to watch with our younger brothers.

Table 3.16: Example text for sentiment-controlled generation. Words in blue/red are positive/negative indicators, respectively.

Model	Generation
UniLM + PT (select) + PL	1) *Well, some people might* think that this film is a masterpiece. They are right. The film is not just a love story, but a love story. What I like about this film is that it shows a different side of women... 2) *Well, some people might* not like this film, but some people might. Well, most people would not like this movie. But the main reason I like it so much is that it has a lot of humor...
KEST	1) *Well, some people might* think it's a little over the top and the story is really predictable, but as I saw on TV in the early 90's I wasn't disappointed in this movie! While the plot is kind of predictable and the main character is supposed to be a guy, the whole thing has been made into a very cool and entertaining film... 2) *Well, some people might* think that this was a lot like "Jaws", or "Alien", or something like that. Sadly, it is not. I was lucky enough to see it. It's a very clever, intelligent and entertaining film with good performances...

Table 3.15: Samples generated with specified positive sentiment and input prompt '*Well, some people might*'. Words in blue/red are positive/negative indicators, respectively.

Model	Generation
	Sentiment: *Negative*; **Prompt**: *The book*
Ctr-PF	*The book* I loved a lot about The Legend of Zelda: Breath of the Wild. Its protagonist, Zelda, is an ordinary woman who grows up playing on a farm. She gets to choose a young boy in her village who can only dream about Zelda. The village gets more and more out of her. She tries to save the villagers by making them happy by getting them into a happy life. The game takes this into an extreme (it's not like Zelda is a good actress, just a poor one)...
GPT2-ST	*The book*, "Fatal Attraction" by R.D.Hilbert, was a great book and even if the book had been better, they probably would have given it a 10 out of 10. I was disappointed by this movie. I do not believe that a movie can be this boring, with actors playing their roles and acting like the movie would give it a realistic look...
UniLM-ST	*The book* was wonderful. The film is a mess. I could watch this and then get frustrated. It is very much like an extended stage play. The actors do what they can with little money and they do a wonderful job in acting. I will not make this up.

DuNST	*The book* was a rather confusing and disappointing adaptation of a classic. It may very well have been a good book if it had been adapted to a movie, but this movie could have been much better. It's impossible to pinpoint exactly what exactly was the point of this film...
KEST	*The book* was terrible. I can't believe it, and I'm not lying. It's awful! There were a lot of mistakes in the book, but it was okay in the movie...
	Sentiment: *Positive*; **Prompt**: *The book*
Ctr-PF	*The book* has an overall feel to it and is quite good. There are plenty of laughs and a decent ending to the book. However, there are some other issues with the author, who has some questionable editing that can ruin the book, making the movie and television series too good to be true...
GPT2-ST	*The book* is a perfect example of why not to make a movie. The dialogue, the actors, the sets, everything. That said, the book is perfect, despite the fact that it has very little plot, and that is good. The acting is good and is also very funny. The story is even more funny than the book...
UniLM-ST	*The book* was excellent. And it was not only about an alien invasion of Earth, but a series of events in the near future. I was a bit reluctant to read the book as it was written after the global collapse of the european market. But then again, it would still be hard to buy the movie More or less, the book is just brilliant, with wonderful characters and good animation...
DuNST	*The book* (which I'm in a "Truly Hip") is a fantastic comic book. The movie adaptation of the book is a superb one. It's faithful to the spirit of the book and looks to the footage and characters. I highly recommend this film if you like and appreciate the story...
KEST	*The book*, by Graham Greene, is one of the finest psychological novels out of all - time. It's a compelling tale about a young woman who travels to England in search of ẗhe book. Äs a reader, I found the book to be faithful to the book as it unfolds...

Table 3.17: Example text for sentiment-controlled generation continued. Words in blue/red are positive/negative indicators, respectively.

Model	Generation
	Topic: *World*
Ctr-PF	(1)On Wednesday, the Iraqi security forces stormed the main Islamic State office, killing nine members, killing 12 people, but did not withdraw from the building and taking over two other buildings. (2)He said this could not be confirmed, but was quoted by Reuters. Hide Caption : A woman reacts after her car was blown up near an Islamist group rally in Dhaka.
GPT2-ST	(1)LOS ANGELES (Reuters) - A former crematory operator agreed on Wednesday to plead guilty to dumping bodies and ashes at the same crematory site where he was born, officials said (2)GAZA (Reuters) - A rocket killed two Israeli soldiers in Gaza on Wednesday, the first time the army in occupied territory in more than a year that Hamas militants have launched a fierce...
UniLM-ST	(1) The Israeli Army has suspended a company commander accused of emptying an ammunition clip into a 13 - year - old Palestinian girl. (2) BAGHDAD, Iraq - A roadside bomb killed two American soldiers and wounded three others in Iraq, the U. S. command said Friday, as insurgents hit Baghdad targets with rocket and rocket bombs...
DuNST	(1) AP - An Italian aid worker walked free from the southern Philippines on Sunday, a day after he was abducted at gunpoint on the streets of Real Aires. (2) AFP - The United States and South Korea failed to hammer out a deal over a timetable for the planned reduction of US forces in Iraq, with Seoul asking for more troops to join another group.
KEST	(1) MOSCOW (Reuters) - At least one Russian ministry has signed letters agreeing to Moscow's approval of the Kyoto Protocol, a spokesman said on Friday. (2) Reuters - Former Peruvian President Alberto Fujimori on Saturday called for the world's highest military ruler to be re-elected, a move that would improve relations between the former foes.
	Topic: *Sport*

Ctr-PF	(1)This past weekend, when the Los Angeles Lakers drafted Michael Jordan, he looked like a real contender to play the role of mentor at the small forward spot. (2)Houston is now playing "The Voice of America" at Madison Square Garden. The Knicks are 0-5 and facing a 10-point Los Angeles Lakers team that, if they win tonight...
GPT2-ST	(1) SEATTLE (Reuters) - Olympic chiefs may have to reconsider their decision to stage a one-day event in Atlanta after protests from marchers in the southern city... (2)NEW YORK (Reuters) - Tommy Haas looked as though he had the flu, as he sat in his BMW 712 at the World Championship in Akron, Ohio, on Friday.
UniLM-ST	(1) England coach Sven - Goran Eriksson says striker Michael Owen must prove in training Monday that he deserves to face Wales in a World Cup qualifier. (2) ATHENS - - The tears were from the Brazilian women's soccer team, who had just won their first Olympic gold medal in women's tennis, beating Australia 4 - 2 in the final...
DuNST	(1) South Carolina assistant Skip Holtz left the game with an injured tailback Ciatrick Fason. Freshman Adrian Peterson rushed for 140 yards and two touchdowns and Ronnie Brown added 127 yards. (2) AP - The New York Yankees wasted little time getting down to business, and their starting pitcher, Tony Womack, was allowed to sit out Saturday night after missing two games because of an elbow injury...
KEST	(1) MINNEAPOLIS - - Minnesota Timberwolves guard Latrell Sprewell was suspended one game without pay by the NBA on Tuesday for directing obscenities. (2) CHELSEA manager Jose Mourinho has dismissed the challenge of keeping the money - losing outman out of Chelsea for the rest of the season.
	Topic: *Business*
Ctr-PF	(1)President Barack Obama has said his administration is "very concerned about Iran's nuclear program and concerns about the growing threat from terrorist groups in Iran. (2) A number of firms have taken steps to make their online business more efficient and more efficient. New York-based Gartner says that new companies such as AT&T, Bell, IBM...

GPT2-ST	(1)NEW YORK (Reuters) - U.S. blue chips sank on Thursday after Ford Motor Co. (2)SINGAPORE (Reuters) - Asian stock markets opened lower on Thursday, helped by poor weather forecasts and gains by technology firms, but some oil-related stocks remained higher.
UniLM-ST	(1) TORONTO (CP) - Stock markets were poised for an early rally Thursday as crude oil prices reached record highs and energy stocks surged on easing supply fears. (2) In the latest move by the US Justice Department, The Washington Post has announced that it will pay $ 60 million cash to buy the parent company of CBS MarketWatch.
DuNST	(1) Tokyo stocks plunged Monday morning as investors took profits from recent gains. The US dollar was up against the Japanese yen. The Nikkei Stock Average of 225 issues was up 36. (2) NEW YORK, Aug 18 (Reuters) - Rupert Murdoch's News Corp. Ltd. has agreed to sell its stake in Sky Latin America to DirecTV Group D
KEST	(1) Pfizer said it would sponsor a major clinical trial of Celebrex, its prescription arthritis pain medication, to assess the cardiovascular benefits of the drug. (2) US consumer confidence rose in August following a sharp drop in the previous month, the Conference Board reported on Tuesday.
	Topic: *Sci/Tech*
Ctr-PF	(1)$1,000 for 'Millionaire's 'Rape Crisis' Victim Fund By Michael S. Osterholm - 6/9/17 07:08:04: (2) US government has approved a $10 million loan to provide medical equipment to the Palestinian Authority for 'humanitarian and medical equipment' on the West Bank. The agreement provides $3 million for a...
GPT2-ST	(1)NEW YORK (Reuters) - The U.S. Securities and Exchange Commission has voted 5-0 to recommend that Internet advertising services stop soliciting fees from Web sites, according to a... (2)LOS ANGELES (Reuters) - "Cell " phones offer fast data rates, low prices and no worries about getting fat in the long run, says a survey by analysts at research firm...

UniLM-ST	(1) Sony Corp.'s music unit is abandoning its CDs that use built - in technology that limits copying them, after pushing the program for two years. (2) The future of the internet could be in doubt in around two years'time, according to two leading internet watchers, who outlined a series of steps they hope will turn the internet into a business...
DuNST	(1) Toshiba has announced a new transmission system for routers and switches that will improve automatic transmission rates. The 6500 Super_GSM/GPRS system will feature high clock speed... (2) At a press conference this week, Bill Murray, Microsoft's CEO, expressed doubt that the software giant's strategy for regaining PC identity is considerable, but heard little reason to believe it
KEST	(1) The Internet Corporation for Assigned Names and Numbers (ICANN) has modified its proposal to include some domain names in the name of the Internet Corporation for assigned Names and numbers. (2) This holiday season, Apple Computers Inc. plans to open its first European retail store later this month in the capital city of Munich.

Table 3.18: Example text for topic-controlled generation.

3.6 Conclusion

In this chapter, We propose two novel methods (DuNST and KEST) to apply Self-training to semi-supervised controllable NLG. DuNST (1) jointly optimizes generation and classification via a dual variational learning framework to leverage both pseudo text and pseudo labels, and (2) incorporates two kinds of soft noise into ST, better exploring larger potential text space and extending the attribute distribution boundary, solving the challenges of *Limited unlabeled data* and *over-exploitation*. KEST (1) applies a practical multi-task generator to generate soft pseudo text in parallel, significantly reducing decoding time while injecting soft noise to the text; (2) uses soft kernel-based loss to encourage exploration of the learned distribution and increase control accuracy and generation diversity, addressing the challenges of *over-exploitation* and *training deceleration*. As a comparison, DuNST can achieve better generation controllability at the cost of longer training time.

Besides, DuNST applies a VAE-like structure, which needs more effort in tuning the hyper-parameters to achieve satisfactory results. On the other hand, KEST provided a solution with faster training speed but sub-optimal performance.

Chapter 4

Dual Contrastive Self-Training for Document-level Relation Extraction

A version of this chapter has been submitted to ACL Rolling Review in 2023 December and will be published at the main conference of NAACL 2024. I was the main investigator. Throughout the project, I led the process of defining project goals and key research questions, the model implementation, and designing and running the experiments. The work was done under the supervision of Laks V.S. Lakshmanan.

4.1 Introduction

Relation Extraction (RE) from unstructured data sources is a key component of building large-scale knowledge graphs (KG) [51, 75]. Among all the RE tasks, Document-level RE [138] extracts subject-relation-object triples from documents, which remains daunting due to the significant challenges in modeling long text spans and obtaining high-quality supervision signals. Current document-level relation extraction methods [89, 138] can discover the semantic relation that holds between two entities under supervised learning. However, these methods typically require lots of manually labeled data for model training, which could be labor-intensive to obtain.

On the other hand, since a large amount of in-domain text is usually accessible, we can tackle document-level RE using semi-supervised learning [8]. There has been substantial work on exploring how to alleviate the amount of human supervision required for RE. Mintz et al. [71] makes use of distant supervision which leverages external knowledge bases to obtain annotated triples. Since distant supervision makes a strong assumption that the relation between entity pairs should not depend on the context, it usually leads to context-agnostic label noises and sparse matching results.

Alternatively, self-training (ST) [91, 127], a classic semi-supervised learning paradigm, has been proposed in relation extraction [42, 96, 129]. ST minimizes the prohibitively expensive human labeling by iteratively pseudo-annotating unlabeled data with a classifier which is then retrained with the augmented labels. In this way, ST benefits from a vast number of unlabeled instances and extends the generalization bound [112, 133].

A significant challenge of ST is inadequate training data for long-tail relations. As shown in Fig. 1.4, current document-level RE systems [138] do not perform well on long-tail relations, which hardly appear in the training data. For example, the F1-score for class *located in* is 83.02 while *ethnic group* is only 6.45. The reason could be that the amount of training data is vastly different (20k vs. 155). Assuming that training data and unlabeled data have the same distribution, we cannot expect these long-tail relations to appear sufficiently often in the unlabeled text corpus. To address this, Tan et al. [96] propose to re-sample training set and to assign more weight to the classes that have high precision and low recall. However, this method does not bring new information to the relation classifier. As a result, these self-training methods might not be able to improve the RE performance on these rare relations.

In order to solve the above issue of long-tail relations, we propose a novel method – Dual contrastive self training for semi-supervised Relation Extraction (DuRE). Unlike previous ST methods [42, 96], we simultaneously train a controllable text generator, generating diverse outputs given specific relation triples. To improve the controllability of the generator, we leverage the signal of the trained RE classifier to label positive and negative generated sequences, and then apply a ranking calibration loss [137] to contrast the positive and negative sequences to improve generation quality. In addition, we propose a self-adaptive way to sample pseudo text from different relation classes. We add noise by increasing generation temperature for relations with higher precision, which introduces diversity to the training set and helps reduce overfitting. Besides, we sample more examples from relations with lower recall. Since long-tail relations usually have a low recall (Fig. 1.4), they are more likely to be sampled, and thus their recall can be increased through training.

The contributions of this work are as follows:

- We dig into the problem of document-level extraction of long-tail relations and propose to simultaneously train a controllable text generator to address the limitation of previous self-training methods [96] that only leverage pseudo-labeling.

- We propose a contrastive loss to control the quality of generated pseudo text, improving the generation quality and thus helping to enhance the classification performance of the relation classifier.

- Comprehensive experiments show that our model significantly improves F1-score in different RE benchmarks on general and biomedical domains, especially on long-tail relations.

4.2 Related Work

Relation Extraction: Deep neural models have proven to be successful in sentence-level and document-level relation extraction. Zhang et al. [136] proposed position-aware attention to improve sentence-level RE and published TACRED, which became a widely used RE dataset. However, most relations in real-world data can only be extracted based on inter-sentence information. To extract relations across sentence boundaries, recent studies began to explore document-level RE. As previously mentioned, Yao et al. [126] proposed the popular benchmark dataset DocRED for document-level RE. Zeng et al. [130] leveraged a double-graph network to model the entities and relations within a document. To address the multilabel problem of Document-level RE, Zhou et al. [138] proposed using adaptive thresholds to extract all relations of a given entity pair. Zhang et al. [132] developed the DocUNET model to reformulate document-level RE as a semantic segmentation task and used a U-shaped network architecture to improve the performance of DocRE. Tan et al. [94] proposed the use of knowledge distillation and focal loss to denoise the distantly supervised data for DocRE. Wang et al. [109] proposed a positive-unlabeled learning algorithm under incomplete annotation scenario. However, *the methods above were not designed to tackle the challenge of long-tail relations.*

Self-training: Recently, Self-training has flourished again by iteratively generating pseudo labels and augmenting the tuning of data-hungry language models, showing great advantages in further enhancing NLU [4, 10, 21, 70, 102] and Relation Extraction (RE) [42, 119, 129], where massive unlabeled input text exists. To tackle the issue of confirmation bias in self-training, Wei et al. [112] re-samples pseudo-labels based on the frequencies of training examples. Tan et al. [96] samples different numbers of pseudo-labeled data based on the development set performance. However, it ignores the relations where both precision and recall are low. *All the above ST methods for RE apply ST only in generating pseudo-labels.* Feng et al. [25] were the first to propose dual self-training to improve controllable text generation by introducing two kinds of noise. However, they do not study the effectiveness of dual self-training in classification problems (such as relation extraction), and in fact the strategy of adding noise sometimes harms the classification performance.

Unlike all the above ST methods, we are the first to apply dual self-training on document-level relation extraction, generating both **pseudo-labeled data** using RE classifier and **pseudo texts** given specific relation triples using a text generator. The design of contrastive loss and self-adaptive generation for different relation classes further improves the performance, especially on long-tail relations.

4.3 Method

4.3.1 Problem Formulation

Document-level relation extraction Given a document x and a set of entities $\mathbf{e} = \{e_j\}_{j=1}^{m}$, the task of document-level relation extraction is to predict a subset of relations from $\mathcal{R} \cup \{N_A\}$ between entity pairs $(e_h, e_t)_{h,t=1...m,h \neq t}$, where $\mathcal{R}$ is a pre-defined set of relations, N_A represents no relations between given entities, and e_h, e_t are identified as head and tail entities, respectively. At the test time, the model needs to predict the labels of all entity pairs in document x.

Semi-supervised document-level relation extraction Let $\mathbf{x}$ be the text, $\mathbf{e}$ be the entities mentioned in $\mathbf{x}$, and $\mathbf{y} = \{e_{h_j}, r_j, e_{t_j}\}_{j=0}^{n_x}$ be the existing relations in $\mathbf{x}$, $D_L = \{\mathbf{x}_i, \mathbf{y}_i, \mathbf{e}_i\}$ be a labeled dataset with paired text and its corresponding relation sets, and $D_U = \{\mathbf{x}_i, \mathbf{e}_i\}$ be an unlabeled dataset from the same domain. In reality, we do not obtain $\mathbf{e}_i$ for unlabeled text corpus. However, we can use tools of name entity recognition (NER) and coreference resolution (CR) to get the entity list in advance. Since we focus on relation extraction only, we assume we have already obtained the entity list for simplicity.

Controllable text generation given relation triples Given relation triples $\mathbf{y} = \{e_{h_j}, r_j, e_{t_j}\}_{j=1}^{n_x}$, where n_x is the number of relations, the task is to generate document x that contains these relation triples.

4.3.2 Methodology

We aim to jointly learn an attribute-controllable generator $\mathcal{G} = P_\theta(\mathbf{x}|\mathbf{y})$ parameterized by θ (*e.g.*, a large PLM) to generate, in an auto-regressive manner, high-quality text $\mathbf{x} \sim P_\theta(\mathbf{x}|\mathbf{y})$ containing the given relations $\mathbf{y}$. We also endow our model with the ability to produce pseudo extractions for $\{\mathbf{x}_i, \mathbf{e}_i\} \in D_U$ through jointly learning a Document-RE classifier $\mathcal{C} = P_\phi(\mathbf{y}|\mathbf{x}, \mathbf{e})$. We simultaneously model and optimize $\mathcal{G}$ and $\mathcal{C}$ with a shared PLM as a dual process.

We train our relation classifier using Adaptive Thresholding loss [138]:

$$\begin{aligned}
\mathcal{L}_P &= -\sum_{r \in \mathcal{P}_T} \log \left(\frac{e^{\text{logit}_r}}{\sum_{r' \in \mathcal{P}_T \cup \{\text{TH}\}} e^{\text{logit}_{r'}}} \right), \\
\mathcal{L}_N &= -\log \left(\frac{e^{\text{logit}_{\text{TH}}}}{\sum_{r' \in \mathcal{N}_T \cup \{\text{TH}\}} e^{\text{logit}_{r'}}} \right), \\
\mathcal{L}_C &= \mathcal{L}_P + \mathcal{L}_N
\end{aligned} \tag{4.1}$$

where positive classes $\mathcal{P}_T \subseteq \mathcal{R}$ are the relations that exist between the entities in T, negative classes $\mathcal{N}_T \subseteq \mathcal{R}$ are the relations that do not exist between the entities, logit_r and logit_{TH} are the predicted logits for class r or threshold TH by classifier $\mathcal{C}$.

For generation side, we use cross-entropy loss for auto-regressive generation.

$$\mathcal{L}_G = -\frac{1}{N} \sum_{(\mathbf{x},y) \in D} [\sum_{j=1}^{L} \log P_\theta(\mathbf{x}^j | \mathbf{x}^{<j}, y)], \tag{4.2}$$

where $\mathbf{x}^j$ means the j-th token in $\mathbf{x}$, L is the length of $\mathbf{x}$, D is the training set with N samples. We will show later how to construct D for different training phases. Finally, we compute a weighted sum of classification and generation loss, where λ_C and λ_G are tunable hyperparameters.

$$\mathcal{L} = \lambda_C \mathcal{L}_C + \lambda_G \mathcal{L}_G. \tag{4.3}$$

4.3.3 Self-adaptive pseudo text generation

The full DuRE method is described in Alg. 3. Following the practice of self-training in NLU [102], we start ST from a strong base model tuned on D_L and use the full unlabeled D_U to produce pseudo labels, rather than select part of the data with certain criteria as in [96]. In addition to pseudo-labeled data, we also use our generator to generate pseudo-text given entity-relation triples. To better improve the RE performance, we propose to use Contrastive Loss (CL) and Self-Adaptive Generation (SAG) methods, which are illustrated below.

Algorithm 3: Training Process of DuRE

Input: Labeled set D_L, unlabeled set D_U, relation set $\mathcal{R}$.

Jointly train base model $\mathcal{G}$, $\mathcal{C}$ on D_L by optimizing Eq.(4.3), store the best $\mathcal{G}_0$, $\mathcal{C}_0$.
for *epoch* $\leftarrow 1$ **to** *MaxEpoch* **do**
    for $\mathbf{x}_i, \mathbf{e}_i$ ***in*** D_U **do**
        $\hat{y}_i = \mathcal{C}_{epoch-1}(\mathbf{x}_i, \mathbf{e}_i)$
    end
    Build pseudo label set: $D_{PL} = \{\mathbf{x}_i, \mathbf{e}_i, \hat{y}_i\}$
    for r_j ***in*** $\mathcal{R}$ **do**
        Sample n_j triples $\{\mathbf{y}\}^{n_j} \subset D_L$ with relation r_j following Eq.(4.5).
        for $k \leftarrow$ *0* ***to*** n_j **do**
            Generate m pseudo texts: $\{\mathbf{x}^k\}^m = \{\mathcal{G}_{epoch-1}(\mathbf{y}^k)\}^m$
            Select the entities $\mathbf{e}^k$ by parsing $\{\mathbf{x}^k\}$
            Compute pseudo labels with $\mathcal{C}_{epoch-1}(\mathbf{x}^k, \mathbf{e}^k)$
            Select positive example $\mathbf{x}^+, \mathbf{y}^+, \mathbf{e}^+$ and negative example $\mathbf{x}^-, \mathbf{y}^-$
        end
    end
    Build pseudo text: $D_{PT} = \{\mathbf{x}^+, \mathbf{y}^+, \mathbf{e}^+, \mathbf{x}^-\}$
    Train $\mathcal{G}_{epoch-1}$, $\mathcal{C}_{epoch-1}$ on $\{D_{PT}, D_{PL}, D_L\}$ by optimizing Eq.(4.3) and Eq.(4.4), update the parameters to $\mathcal{G}_{epoch}$ and $\mathcal{C}_{epoch}$.
end

Contrastive Loss (CL) Following Zhao et al. [137], we use Sequence Likelihood Calibration (SLiC) to align a language model's sequence likelihood, $P_\theta(x|y)$, over decoded sequences according to their similarity to reference sequences. Given $\mathbf{y}_j = \{e_{h_j}, r_j, e_{t_j}\}$, we generate a batch of m generations $x_j^k = \mathcal{G}(\mathbf{y}_j)$. Then the ranking calibration loss contrasts a positive sequence x^+ and a negative sequence x^-, encouraging the model to assign more probability mass to positive compared to negative sequences, and thus enhancing generation quality. Thanks to the dual learning framework, we can use the classifier $\mathcal{C}$ to evaluate the confidence of whether generation x contains relation triple $\{e\}$. More specifically, $x^+ = \text{argmax}_i logit^i_{r_j}$, and $x^- = \text{argmin}_i logit^i_{r_j}$. Thus the contrastive loss is computed as follows:

$$\mathcal{L}_G = \max\big(0, \beta - \log P_\theta(x^+|y) + \log P_\theta(x^-|y)\big) \tag{4.4}$$

where $\beta > 0$ is a hyperparameter to control the margin between positive and negative examples.

Self-Adaptive Generation (SAG) Another advantage of using pseudo text is that we can tune the distribution of relations and generate more pseudo texts for long-tail relations. Compared to traditional methods like bootstrapping [22] and class-rebalanced self-training [112] where duplicated samples are chosen for a specific class, the generator can produce more diverse training examples.

Here we propose a self-adaptive pseudo-text generation strategy. For different relation classes with different development set performances, we generate pseudo text accordingly. For a relation class r, if its recall R_r is low, then we should improve the recall by generating more pseudo examples. If its precision P_r is high, we believe this class is well-predicted. Then the purpose of pseudo-text generation is to augment the training set with noisy data to enhance generalization. A good way to add noise is to generate data with higher temperature [25]. Otherwise, if P_r is low, we use a lower temperature to sample a more certain output to ensure the generation quality. In summary, we define the sampling probability ϕ_r and generation temperature temp_r for relation r as follows.

$$\begin{aligned} \phi_r &\sim (1 - R_r), \\ \text{temp}_r &= \alpha + P_r. \end{aligned} \tag{4.5}$$

where $\alpha > 0$ is a hyperparameter to control the noise level of generation.

4.4 Experiments

4.4.1 Datasets

We experimented with our method on two datasets. For general-domain document level RE, we use Re-DocRED [95], a high quality dataset. We

Dataset	Train	Dev	Test	Unlabeled
Re-DocRED				
# Documents	3053	500	500	108000
Avg. # Entities	19.4	19.4	19.6	19.4
Avg. # Triples	28.1	34.6	34.9	-
Avg. # Sentences	7.9	8.2	7.9	8.0
# NA rate	94.3%	93.1%	93.1%	-
CDG (5%)				
# Documents	3847	1480	523	72697
Avg. # Words	196.9	236.5	235.6	197.0
Avg. # Entities	7.4	8.8	10.0	7.6
Avg. # Triples	2.1	2.2	2.6	-
Avg. # Sentences	12.6	14.0	13.2	12.6
# NA rate	96.8%	97.7%	93.8%	-

Table 4.1: Dataset statistics.

use the distantly-labeled set of Re-DocRED as unlabeled set, only keeping named entities information and coreference information but ignoring the distant labels. As a second dataset, we tested our method on the biomedical document-level RE dataset ChemDisGene (CDG) [131]. Since there is no unlabeled set in CDG, we only leveraged part (e.g., 5%) of training set as labeled data and kept aside the rest as unlabeled data. Our models are evaluated on the test sets of Re-DocRED and CDG. Both of the test sets are human-annotated and have high quality. The statistics of the datasets can be found in Table 4.1.

4.4.2 Experimental Settings

We use ATLOP [138] as our base RE classifier and an encoder-decoder framework as our pseudo text generator. Following Feng et al. [25], we share the encoder part of RE classifier and pseudo text generator to save the total number of training parameters. For Re-DocRED dataset, we use Flan-T5-base [14] as the base encoder-decoder model. For CDG dataset, we use a version of Flan-T5-base that is pretrained on Pubmed dataset[13]. We use AdamW [65] with learning rate = 5e-5, warm-up rate = 0.06, $\lambda_g = 1$, $\lambda_c = 5$, $\alpha = 0.7$, $\beta = 0.3$, and batch size = 8 for optimization across all

[13]The model can be found at `https://huggingface.co/gubartz/ssc-flan-t5-base-pubmed`.

Model	P	R	F1	Ign_F1	Freq_F1	LT_F1
Bert-base-cased						
ATLOP†	**86.70**	62.46	72.61	71.86	75.92	67.46
NS†	77.63	69.17	73.16	72.92	77.28	67.59
VST†	72.77	**75.55**	74.14	72.48	78.47	68.13
SSR-PU	76.78	71.46	74.33	72.91	78.41	68.32
CREST†	75.94	72.47	74.17	72.77	77.93	68.68
CAST†	76.59	72.84	74.67	73.32	78.53	69.34
Flan-T5-base						
ATLOP-Flan	86.40	61.78	72.05	71.32	76.15	65.45
ATLOP-Dual	85.17	61.93	71.72	70.95	76.07	64.70
DuRE	79.01	73.84	**76.84**	**75.32**	**79.81**	**72.88**

Table 4.2: Relation classification results on Re-DoCRED dataset. †Results are obtained from Tan et al. [96].

tasks. As is common practice [39], we use the top-p sampling method with $p=0.9$ for decoding. For the generation task, we add a prompt sentence at the beginning of the input: *Generate text given the following relation triples.* The number of pseudo text is the same as the number of labeled data, i.e., 3053 for Re-DocRED and 3847 for CDG.

We tuned all hyperparameters only on the held-out development set. In self-training phase, we tuned $\lambda_c \in \{1, 5, 10\}$, $\alpha \in \{0.5, 0.7, 1.0\}$, and $\beta \in \{0.1, 0.3, 0.5\}$ in Re-DocRED dataset to obtain the reported results. Finally, we set $\lambda_c = 5$ and $\lambda_g = 1$ in base-model training phase, while $\lambda_c = 5$, $\lambda_g = 1$, $\alpha = 0.7$, and $\beta = 0.3$ in the self-training phase. We tuned the hyperparameters in Re-DocRED dataset and applied them to all tasks. We use AdamW [65] as an optimizer. The training batch size is 8, and the learning rate is $5e-5$. We apply linear warmup to the optimizer, and the warmup ratio is 0.06.

We implement DuRE on Huggingface Transformers [115] library of v4.31.1 and use four NVIDIA Tesla V100 nodes to train our model. The total number of training hours is around 39.62h for Re-DocRED and 34.54h for CDG(5%). The number of parameters of our model is 242.91M. In the generation phase, we use top-p sampling ($p=0.9$) as the decoding method. Other configuration of the generator includes a length penalty to be 1.0 and a repetition penalty to be 1.0 for all baselines. All experimental results are trained and tested in a single run with fixed random seeds.

For the number of pseudo text generated, following [25], we choose the same number as the labeled training set. The prompt for generation is *Generate text given the following relation triples*. An example of generation could be *Generate text given the following relation triples. * The Invisible Man*, lyrics by, * Roger Taylor**. For each input triple, we sample 8 sentences and apply the RE classifier to select positive and negative pseudo texts among these samples.

Model	**P**	**R**	**F1**
PubMedBERT on CDG (100%)			
ATLOP †	**76.17**	29.70	42.73
NS †	71.54	35.52	47.47
SSR-PU	54.27	43.93	48.56
CREST †	59.42	42.12	49.28
CAST †	66.68	45.48	54.03
Flan-T5-base on CDG (5%)			
ATLOP-Dual	46.40	21.78	32.05
DuRE	52.01	48.84	50.38
Flan-T5-base on CDG (50%)			
ATLOP-Dual	47.67	51.24	49.39
DuRE	54.91	**59.43**	**57.08**

Table 4.3: Results on CDG dataset. †Results are obtained from Tan et al. [96].

4.4.3 Evaluation Metrics

Following Tan et al. [96], we used micro-averaged F1 score as the evaluation metric. We also evaluate the F1 score for frequent classes and long-tail classes, denoted as Freq_F1 and LT_F1, respectively. For the Re-DocRED dataset, the frequent classes include the top 10 most popular relation types in the label space; the rest of the classes are categorized as long-tail classes. We also use an additional metric Ign_F1 on the DocRE task. This metric represents the F1 score, calculated for the triples that do not appear in the training data.

4.4.4 Baselines

We compare our model with the following strong document-level RE baselines, including both supervised and semi-supervised.

Supervised Approaches (1) **ATLOP** [138] A vanilla baseline model for document-level RE. (2) **ATLOP-Flan** Only use Flan-T5 [14] Encoder to train the RE classifier without generator. (3) **ATLOP-Dual** Simultaneously train RE classifier given input documents and text generator given relation triples but without self-training. (4) **Negative Sampling (NS)** [58]: randomly select partial negative samples in training to alleviate the detrimental effect of the false negative problem.

Semi-supervised Approaches All the baselines below leverage ATLOP as their backbone. (1) **Vanilla Self-Training (VST)** [47, 78]: a variant of simple self-training where models are trained with N folds, and all pseudo-labels are directly combined with the original labels. (2) **SSR Positive Unlabeled Learning (SSR-PU)** [109]: SSR-PU utilizes positive unlabeled learning and a shift-and-squared ranking (SSR) loss to accommodate the distribution shifts for the unlabeled examples. (3) **Class Re-balancing Self-Training (CREST)**[112]: This algorithm re-samples the pseudo-labels generated by models based on the frequencies of the training samples. (4) **Class-Adaptive Self-Training (CAST)** [96]: this method calculates the precision and recall scores of each class on the development set and uses the calculated scores to compute the sampling probability of each class to alleviate confirmation bias caused by erroneous pseudo labels.

4.4.5 Results

The experimental results on the test set of Re-DocRED (Table 4.2) demonstrate that our DuRE achieves consistent performance improvements in terms of F1 scores over all baselines. The F1 difference between the best baseline CAST and our DuRE is 2.17 (76.84 vs. 74.67). We also found that simply adding an additional RE-controlled generation task does not improve the relation classification performance (ATLOP-Flan vs. ATLOP-Dual), where F1 scores decreased slightly (72.05 vs. 71.72). In addition, training on an encoder-decoder framework (Flan-T5-base) does not outperform a single encoder framework (Bert-base) in document-level RE tasks. However, we notice that the gap between CAST and ATLOP(bert) is 2.06, while the gap between ATLOP-DuRE and ATLOP-Dual is 5.12. This indicates that our proposed dual self-training method improves the RE classification quality significantly more compared to the backbone model. We also notice a considerable improvement (+3.54 F1) especially in long-tail relations, showing that self-training on a rebalanced pseudo text is better than simply doing bootstrapping in the existing training set. The reason is that our trained generator can generate more diverse examples, which helps reduce overfitting on training data.

Table 4.3 presents the experiments on biomedical RE (CDG dataset). Our DuRE model achieves a performance comparable to the baselines with only 5% of the training data. When trained with 50% of the training data, we get the best performance, outperforming the CAST baseline with +3.05 F1. Based on the results of RE experiments in general and biomedical domains, self-training-based methods aim to improve recall and consistently improve overall performance. However, our DuRE maintains a better balance between increasing recall and maintaining high precision.

4.4.6 Ablation Study

We conduct an ablation study on the Re-DocRED dataset. As shown in Table 4.6, we can see that (i) Self-Adaptive Generation benefits the F1 score for both frequent and long-tail relations; (ii) Contrastive loss further enhances all F1 scores in relation extraction; and (iii) Self-training on pseudo-labeled text leads to an improvement in recall but has relatively low precision, which shows that self-training is able to balance precision and recall. This observation is also consistent with Tan et al. [96]. Full version of Table 4.6 is included in the Appendix.

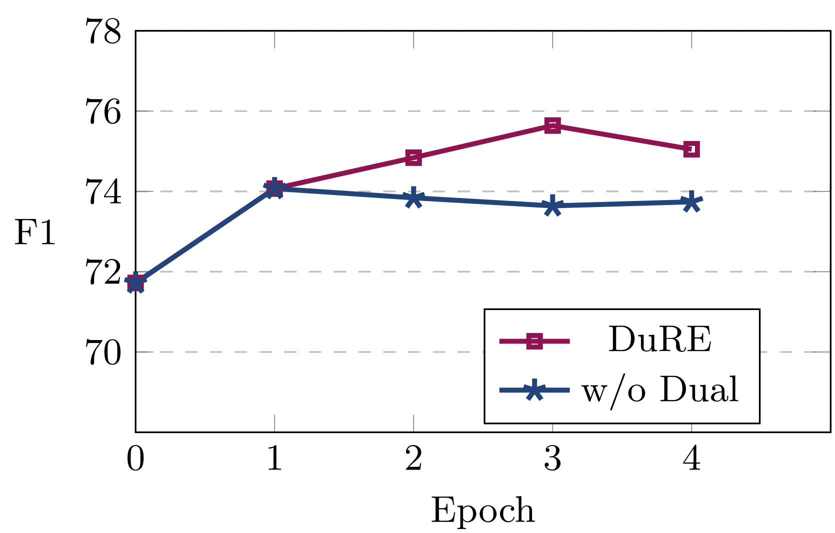


Figure 4.1: F1 score over the number of training epochs on Re-DocRED.

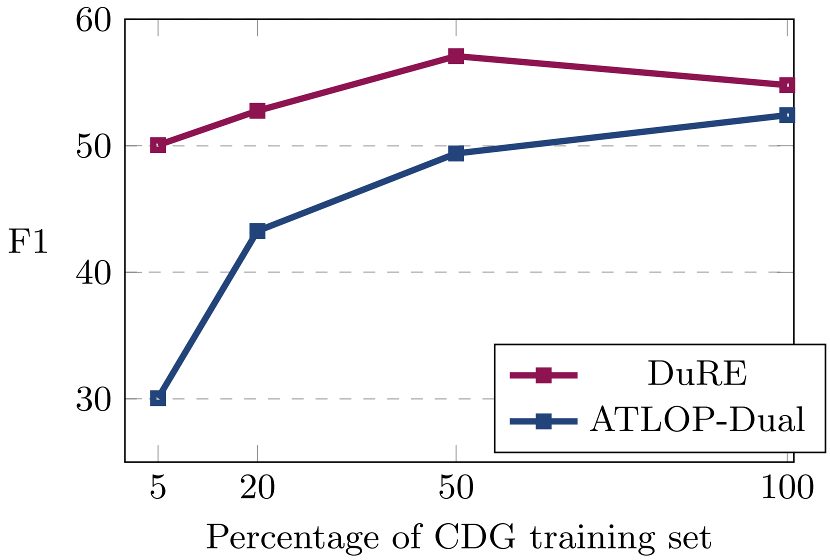
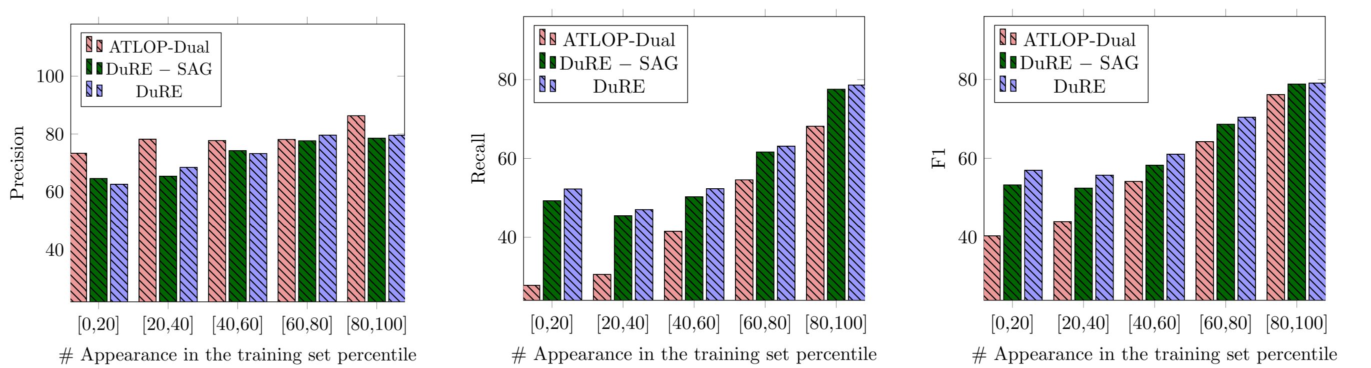


Figure 4.2: Test F1 score with different numbers of labeled data on CDG dataset. All methods use FLAN-T5 as the base model.

4.4.7 Analysis

Effect of Self-Training We compare our model with a variant (−**Dual**) where we use the base generation model $\mathcal{G}_0$ to generate pseudo text and do not update it through self-training. As depicted in Fig. 4.1, classification F1 reaches its maximum quickly and then stops increasing. On the other hand, thanks to the simultaneously optimized generator, DuRE keeps improving generation quality and refining pseudo labels, and thus iteratively improves the quality of the relation classifier.

Number of labeled data We also assess our model with varying numbers of labeled training instances, with the remaining instances treated as unlabeled data. We observe consistent superiority of DuRE to ATLOP-Dual model. Indeed, even DuRE (50%) outperforms ATLOP-Dual (100%), showing that our method can work with scenarios with fewer training data. However, we notice a drop in DuRE (100%). We conjecture that the reason is that the training set of CDG itself is distantly labeled, which is noisy and incomplete. Through self-training on pseudo-labeled data, the model can figure out incompletely labeled relations. However, as for training on the full distantly-labeled dataset, DuRE only benefits from generated pseudo texts compared to ATLOP-Dual, which limits the improvement of the classification model on pseudo-labeled data. Notice that the generated pesudo texts might still have higher quality than simply distantly labeled data.

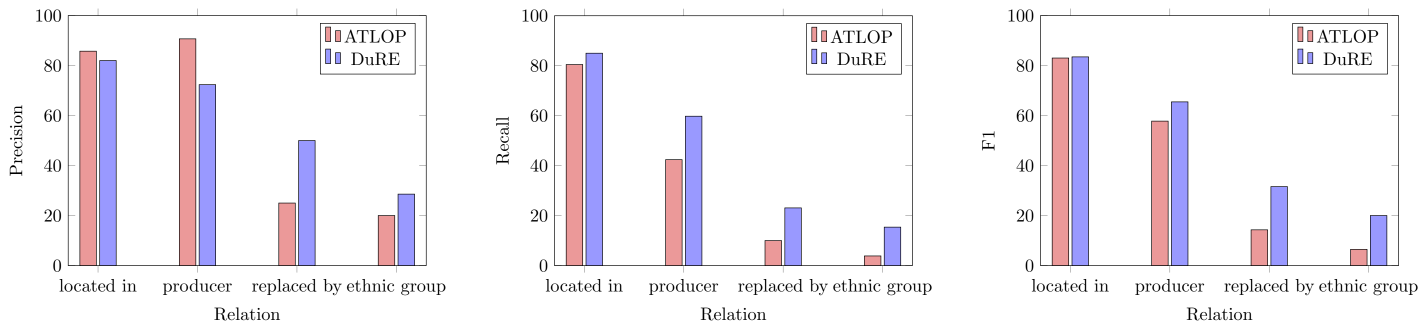


(a) Precision (b) Recall (c) F1

Figure 4.3: Precision, recall, and F1 score on the development set over different percentile of appearances in the Re-DocRED training set.

Input	* The Invisible Man*, lyrics by, * Roger Taylor*.
Positive	“ * The Invisible Man* ” is a rock song originally **written by** * Roger Taylor* and * Mark Smith* in * 1942*. The song was sung at the * Royal Academy of Music* (Royal Academy of Music) in * 1952*. The lyrics and compositions of ” The Invisible Man ” were originally a single with * Taylor*'s voice in a voodoo voice, a version of * Ralph Ellis*'s ” * The Man Who Walks Away* ”. The medley is a ballad, and is not a traditional ballad. ” * The Invisible Man* ” is often performed in * England* and * France*, but it was originally recorded as a concert song for the * Royal Academy of Music*, in * 1936*, and in * Japan* (including the final song).
Negative	* The Invisible Man* is the sixth **album by** * Roger Taylor*. It was released in October 1966. The album features a different version of ” * The Man In the Sun* ” and a different theme. *** The Man In the Sun*** is set in a mystical (and heavily symbolic) future. **It** was **composed by** * Roger Taylor*. The album is very similar to the * The Black Swan* album. It is not necessarily a remake of the album, but rather a remastered version of the earlier album. The album contained a new riff and a new orchestration which was intended to provide additional background music for the final section of the album.

Table 4.4: Example of generation on Re-DocRED dataset.



(a) Precision (b) Recall (c) F1

Figure 4.4: Comparison of test set performance of baseline ATLOP and DuRE trained on Re-DocRED dataset.

Performance of relations with different frequency We sorted the relations by their frequency of appearance in the training set from low to high and grouped them into five groups by their percentile. Fig. 4.3 plots the development set performance of relations with different frequencies. We can see that original ATLOP tends to have much higher precision than recall, while our DuRE consistently improves recall and F1 scores over relations across all frequency groups, at the price of a modest drop in precision. We also notice a tendency that our method improves F1 score more for less frequent relations thanks to the self-adaptive generation, indicating the effectiveness of our methods on long-tail relations.

Improvement on rare relations We plot the precision and recall scores of DuRE and CAST for different relation classes in Fig. **??**, where the experimental results are obtained by training with the Re-DocRED dataset. We found that DuRE significantly improves the recall scores of long-tail classes (*producer*, *replaced by*, and *ethnic group*) and thus improves F1 scores, while maintaining the F1 scores of frequent classes (*located in*). We also notice that the improvements in recall scores are accompanied by a decline in precision scores for frequent classes due to our strategy of learning from noisy pseudo texts for frequent classes. Learning from these augmented noisy texts would decrease the threshold and thus improve recall but decrease precision.

Further analysis of contrastive loss Table 4.4 illustrates a case study of how contrastive loss improves generation quality and thus improves relation classification. Both positive and negative examples are generated under the same decoding strategy. We can see that the negative example does not entail the relation (*The Invisible Man, lyrics by, Roger Taylor*), given that there are similar relations (*The Invisible Man, album by, Roger Taylor*) and (*The Man In the Sun, composed by, Roger Taylor*). This example shows that our learned text generator could sometimes generate text that does not entail the given prompt. Thanks to the contrastive loss, the learned text generator could learn from the signal from the RE classifier, maximize the margin of positive and negative samples, and be more faithful to the given prompt.

Performance of Generation Model To evaluate the quality of generated pseudo text, we measure generation fluency (perplexity, PPL), faithfulness (accuracy of generated text followed by given prompts, Acc), and diversity (number of distinct n-grams, DIST). We measured different types of samples:

Method	Samples	PPL ↓	Acc ↑	DIST ↑
	Testing set	17.80	-	53.42
	random	17.31	71.94	49.84
ATLOP-Dual	positive	17.04	80.96	45.09
	negative	18.98	60.81	50.82
	random	17.03	79.54	50.24
DuRE	positive	16.93	82.28	49.92
	negative	18.32	67.43	52.88

Table 4.5: Performance of learned generator on Re-DocRED dataset.

Model	**P**	**R**	**F1**	**Ign_F1**	**Freq_F1**	**LT_F1**
DuRE	80.01	73.84	76.84	75.32	79.81	72.88
DuRE −SAG	79.51	73.24	76.24	75.02	79.71	70.68
DuRE −CL	80.44	71.43	75.67	74.04	79.24	69.90
DuRE −PT	80.04	70.35	74.88	73.72	79.14	67.88
DuRE −PL	82.71	66.53	73.75	72.79	77.81	67.20
DuRE −PL−PT	85.17	61.93	71.72	70.95	76.07	64.70

Table 4.6: Ablation study on Re-DocRED dataset. Here − means removing components from DuRE. −SAG: remove the self-adaptive generation strategy and sample pseudo texts for different classes in the same setting. −CL: do not generate positive/negative examples and sample random examples. −PT/−PL: do not use pseudo text/labels.

random, positive, and negative. The result is shown in Table 4.5. We found that our generator improves all metrics through dual self-training. Also, the generation fluency is comparable to the testing set. Besides, the diversity of DuRE also improves thanks to the use of more unlabeled text. However, the diversity of generation is still worse than the testing set due to the restricted model size.

4.5 More Examples of Generation

We sample some generated texts given long-tail relations in Re-DocRED dataset. Table 4.7 reports the positive and negative examples generated by our DuRE model. We can find out that even if the relations are long-tail and might not be recognized by our RE classifier, positive examples still contain the relation triples, while some negative examples no longer contain the given relation triples. Learning the contrastive loss helps the generator learn more faithful examples, and thus improves the quality of pseudo training data for RE classifier. This can explain why generated pseudo text for long-tail relations can still be helpful in improving the F1 scores.

Input	* Christianity*, seperated from, * Judaism*.
Positive	The * Holy Spirit* (* Synchrony of the Holy Spirit*) is the most effective means of establishing and strengthening the religious adherence to * Judaism*. As * Christianity* continues to be a branch of * Judaism*, it is crucial that the doctrines of the * Holy Spirit* and of * Judaism* remain with us. It is the aim of the * Holy Spirit* to establish and strengthen the character of the religious adherence of * Judaism* through systematic means of reforming the system of religious practices, the decentralization of religious ceremonies, and the re-establishment of unity among the * Judaic* and * Christian* communities. The most successful form of this is the systematic observance of the * Bible*, in which the primary source of the faith is the knowledge of * Allah* (" Father, Lord, High, Father, High)... In modern terms, the * Holy Spirit* focuses primarily on the advancement of religious adherence, which has been known as the most effective form of a religious system in * Judaism*. The * Holy Spirit* emphasizes the observance of the * Bible*, which has its roots in the * Bible*. The * Holy Spirit* is focused on the expansion of * Judaism*, which has a tendency to create a sense of unity among all * Judaism*. Its success in teaching and interpreting * Judaism* is demonstrated by its use of simple language and the presentation of the * Bible* as a unified unit.
Negative	* Christianity* and * Judaism* are two forms of the * Western Church* which originated around * 1500*. * Christianity* and * Judaism* had a strong influence on the * Eastern Church* and in * Europe* during the second half of the 19th century, with significant influence over * Germany* and * Poland*, which had a relatively neo - western influence. * Christianity* was characterized by a strong influence on the * Eastern Church* and in * European* countries. It is not clear when these influences actually began to manifest in the * Western Church*. However, * Judaism*'influence spread to other areas of the world such as * Africa*, * Asia*, and * the Americas*, and was influenced by this influence.

Input	* Chipilo*, sister city, * Segusino*
Positive	The * Segusino* (or * Segusino* in * Japanese*) was a city in the western part of * Asia*. It was a part of the * Western Asian region* that was unified by the * East Asian region* and * North American region*. It was originally the location of the * Pacific* coast resort town of * Chipilo*, and then the * Mid - East* area. It was a " hamlet ", also called a " sarket ", a kind of rubber cement used in transport. In * Japan*, the town of * Chipilo*, located north of * Tokyo*, was a hamlet that was the home of * Segusino*'s samurai army and the * Japanese* colonel's troops.
Negative	* Segusino* is the second largest city in the island of * Chipilo*. The city has a population of around * 1.27 million* as of the year * 2010*. Its urban population has declined in recent years due to urban development. It is home to the * Segusino Opera House*, a branch of the * Segusino Music Theatre*. The capital of the city is * Segusino*, and the city is surrounded by a number of historic sites, such as the * Segusino Cathedral* and the * Rialto Hall*. A major site dedicated to historic architecture is * Seiji*'s city hall. The first major buildings were built on * Segusino*'s main mountain, * Yatsuda*. The second large building was built on * Seiji*'s main mountain, * Yatsuda*. It was built by * Yatsuda* in * 1955*, but was demolished in * 1964*.

Input	* Russian*, seperated from, * Soviet Union*
Positive	The * Russian* military in the late 18th and early 19th centuries was characterized by systematic, largely self - destructive attacks on civilians, in which large numbers of soldiers and civilians were transported to other parts of the world, the " concentration camp ", or at least the most isolated concentration camp in the world. These, however, were mostly civilians who were not included in the camp's security measures. Most of them were forced to abandon their lives because of armed and destructive attacks, or because they did not comply with the orders of a military commander. The * Russian* military system consisted of a government known as the " * General Army of the Russian People* " (¡unk¿, * Military Provisional Army of the Soviet People*), which was created for the defense of civilians and was separated by the * Soviet Union* at the beginning of the nineteenth century. * Soviet* military operations began in the late * 19th century*, after the dissolution of the * Soviet Union*. The * General Army of the Soviet People* was, under the command of * Alexander Nevsky*, one of the most successful units in the * Soviet* army, in the early * 20th century*.
Negative	The * Russian* and * German* * Civil War* of * World War I*, fought between * 1940* and * 1941*, was a period of political uncertainty in the post - Communal era. Initially, the crisis erupted in * World War II* (and possibly the reunification of the * German* and * Polish* - * Soviet Union*) when the * German* and * Russian* military governments were faced with the imposition of a series of massive counter - Communal invasions and repressions. The conflict also marked the end of the * German* war, when * France* and * Germany* were both forced to declare war. In the aftermath of the war, * German* and * Russian* nationalists became the major force involved, often with the help of * German* troops. This increased tension, as the war progressed, the use of armies, and a series of armed rebellions.

Table 4.7: More examples of generation on Re-DocRED dataset on long-tail relations.

4.6 Conclusion and Future Work

We propose a novel DuRE method to apply Self-training to semi-supervised document-level relation extraction. DuRE (1) jointly optimizes generation and classification via a dual learning framework to leverage both pseudo text and pseudo labels, (2) incorporates contrastive loss to improve the quality of pseudo texts, and (3) applies self-adaptive generation to reduce overfitting of well-predicted relation classes and to improve the performance of long-tail relations. Given that the pseudo data is generated in an auto-regressive manner, which takes longer training time, we plan to explore ways to accelerate the self-training process in the future.

Chapter 5

Conclusion and Future Work

5.1 Summary of the Thesis

There has been a growing interest in automatically constructing knowledge graphs (KGs) in recent years, particularly for applications such as fact-checking, question-answering, semantic search, and recommendations. The typical pipeline for constructing KGs involves various components, including Relation Extraction (RE) from unstructured data sources. RE is a crucial step in building large-scale KGs, and it involves identifying and extracting relation triples from unstructured data. While previous research has focused on sentence-level attribute classification, where relation triples are extracted from individual sentences, there has been a growing interest in document-level relation extraction, which involves extracting relations from entire documents. However, this task poses significant challenges, such as modeling long text spans and obtaining high-quality supervision signals, which have made it relatively underexplored. Another significant challenge of document-level RE is the cost of annotation. Since a document consists of multiple sentences, annotating labeled data for document-level RE is also much more expensive than sentence-level RE.

This thesis aims to reduce the need for human labor in document-level relation extraction by exploring unsupervised and semi-supervised methods that require minimal or no labeled training data. A typical semi-supervised method Self-Training (ST) is discussed in this thesis. As outlined in Chapter 1, we have proposed three research questions (Figure 5.1:

- **RQ1 (Unsupervised RE)**: Can pipeline-based approaches extend to other domains and languages? Further, can we improve the recall of pipeline-based approaches while keeping high precision?
- **RQ2 (Text Generation)**: How can we improve the controllability and diversity of text generation? Specifically, how do we overcome the three challenges (inadequate unlabeled data, over-exploitation, and training deceleration) for self-training in controllable NLG?

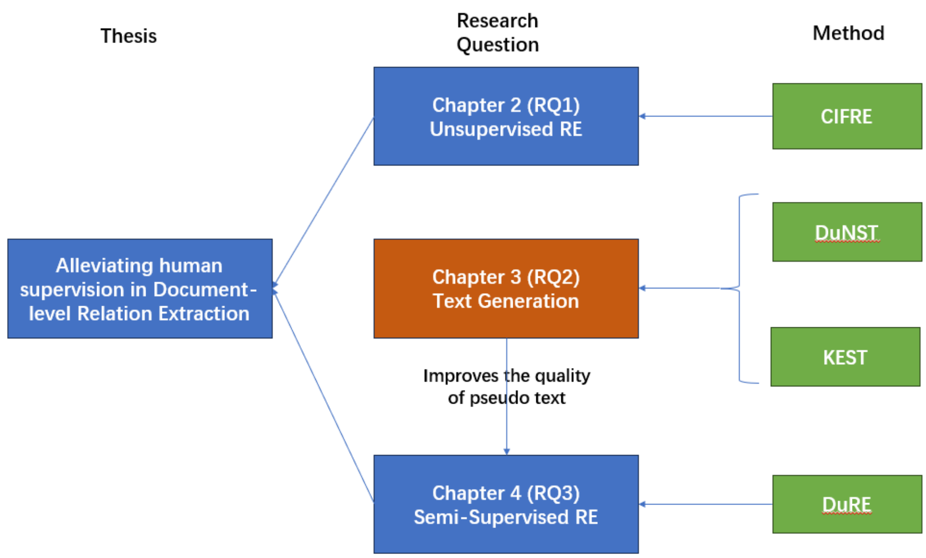


Figure 5.1: The overview of this thesis with corresponding concepts. The left column is the main theme of the thesis, which is supported by the research questions in the middle. The middle column is organized by chapters for each research question. The right column describes the methods designed to answer these research questions.

- **RQ3 (Semi-supervised RE)**: How can we improve the performance (e.g., F1-score) of semi-supervised RE, especially for long-tail relations?

Chapter 2 aims to answer RQ1: *the pipeline-based approach can be potentially extended to other languages. We can improve the recall of pipeline-based approach with missing entity recognition and idiom resolution.* We have developed a system called ChInese Financial Relation Extraction (CIFRE) which uses a modular pipeline to construct large-scale high-quality knowledge graphs. We are the first to use semantic role labeling (SRL) for Chinese relation extraction, and we have created a dictionary of financially relevant and semantically constrained predicates to ensure high-quality extractions. We have also designed patterns specific to Chinese, leveraging the language's unique grammatical structures, for pattern-based extraction. Additionally, we have improved pattern-based extraction by considering coordinated relations, and we have developed methods to detect and complete extractions with missing entities and resolve Chinese idioms via back-translation. Finally, we have demonstrated the performance of our pipeline on two corpora of Chinese financial news, compared to a suite of baselines.

Chapter 3 aims to answer RQ2: *Self-Training with dual objective (both classification and generation side) is able to improve the controllability and diversity of text generation, with specific tricks to address the issue of over-exploitation (adding noise or kernel-based loss) and training deceleration (non-autoregressive generation).* To handle the challenge of inadequate unlabeled data and over-exploitation, we propose a novel **Du**al **N**oisy **S**elf **T**raining (**DuNST**) (Section 3.3), for semi-supervised controllable NLG. DuNST is a model that learns to generate text from attribute labels and predict labels for text, simultaneously. This dual approach allows the model to improve both text generation and classification by leveraging both pseudo text and pseudo labels, addressing the challenge of inadequate unlabeled data for training. Additionally, DuNST introduces two new types of noise to further improve the model's performance, softmax temperature and soft pseudo text. These noisy elements help the model escape the previously learned text space, leading to better robustness and generalization and addressing the challenge of over-exploitation. Theoretically, DuNST can be seen as exploring a larger potential space, which enables it to achieve a better balance between exploration and exploitation, resulting in improved controllability, generation fluency, and diversity. Further, to handle the challenge of over-exploitation and training deceleration, we propose another novel self-training framework, **K**ernel Distance Based **E**fficient **S**elf **T**raining (**KEST**) (Section 3.4), for improving semi-supervised controllable NLG. KEST improves upon traditional learning by skipping the use of cross-entropy loss and instead fitting the approximated text distribution from the last iteration directly in the embedding space. This approach relaxes constraints and encourages the model to produce diverse outputs, addressing the challenge of over-exploitation. Additionally, KEST uses a non-autoregressive generation schema to produce soft representations of pseudo text in parallel, rather than hard strings, reducing the time cost and addressing the challenge of training deceleration. The use of soft text also helps the model denoise errors and propagate local smoothness.

Chapter 4 aims to answer RQ3: *Dual Self-training on both classification and generation can improve the F1 score in semi-supervised document level RE, especially for long-tail relations.* We propose a novel method – Dual contrastive self-training for semi-supervised Relation Extraction (DuRE). Our approach differs from previous self-training (ST) methods by simultaneously training a controllable text generator and a relation extraction (RE) classifier. To enhance the generator's controllability, we utilize the signal from the trained RE classifier to label positive and negative generated sequences and apply a ranking calibration loss to improve generation quality. Additionally, we propose a self-adaptive method for sampling pseudo text from different relation classes. We introduce diversity to the training set by adding noise to the generation process, which helps reduce overfitting. We also sample more examples from relations with lower recall, particularly long-tail relations, to increase their recall through training.

Throughout the thesis, we have explored various ways to improve the quality of document-level relation extraction in unsupervised and semi-supervised settings. We apply self-training in both classification and generation directions, improving the classification quality in semi-supervised document-level RE, especially for long-tail relations. We overcome specific challenges of self-training and apply it in semi-supervised relation extraction. Comprehensive experiments are conducted to demonstrate the effectiveness of our proposed methods against state-of-the-art baselines.

5.2 Future Work

We have shown that our principled approach can benefit document-level relation extraction research and the understanding of self-training. We propose some future extensions of our work presented in this thesis.

Document-level Relation Extraction in the era of Large Language Models Recently large language models (LLMs) like ChatGPT [14] or GPT4 [76] have raised people's attention with their emergent ability [113] to learn from a few examples in the context, which is so-called in-context learning (ICL). Wadhwa et al. [103], Li et al. [53], and Xu et al. [120] evaluated the performance of LLMs in *sentence-level* relation extraction tasks. However, research on applying LLMs on *document-level* relation extraction tasks are still limited and cannot outperform supervised-finetuning on smaller encoder-based models (e.g., BERT[17] and RoBERTa [63])[77].

Although large language models like ChatGPT are powerful, there are still weaknesses. First, it can be hard to fine-tune these models with more than 10B parameters. Second, it suffers from an even more serious hallucination problem than previous relation extraction models [53]. Third, due to the model's large number of parameters and not open-source, it is even less interpretable.

There are some possible directions to improve the performance of LLMs on document-level relation extraction. 1) Improve the general ability in the foundation model, such as the quality of question answering and understanding extremely long documents. 2) Explore more effective prompts to improve the task-specific zero-shot/few-shot ability. 3) Large Language models can be used to improve the modules in the pipeline proposed in Chapter 2. The rules proposed in the pipeline can be used to regulate the output of LLMs, and the LLMs can be used in each module (such as named entity recognition and semantic role labeling), improving the recall and precision of zero-shot RE.

Relation Extraction for long documents In document-level RE, the distance between head and tail entity is usually much longer than in sentence-level RE, which brings more difficulty in modeling. Long distance relations are the relations whose head and tail entities are separated by a long text span. Tackling long distance relations is orthogonal to tackle long-tail relations proposed in the thesis. According to Ru et al. [89], around 20% of the relation triples consist of entity pairs between which the distance is more than 100 tokens in the DocRED dataset. As a result, document-level RE systems tend to perform worse on the relations whose head and tail entities are far from each other[89, 138]. Figure 5.2 indicates that the relations with longer distances perform worse in micro-F1 score than relations with closer distances. How to address this issue for extremely long documents is still an

[14] `https://openai.com/blog/chatgpt`

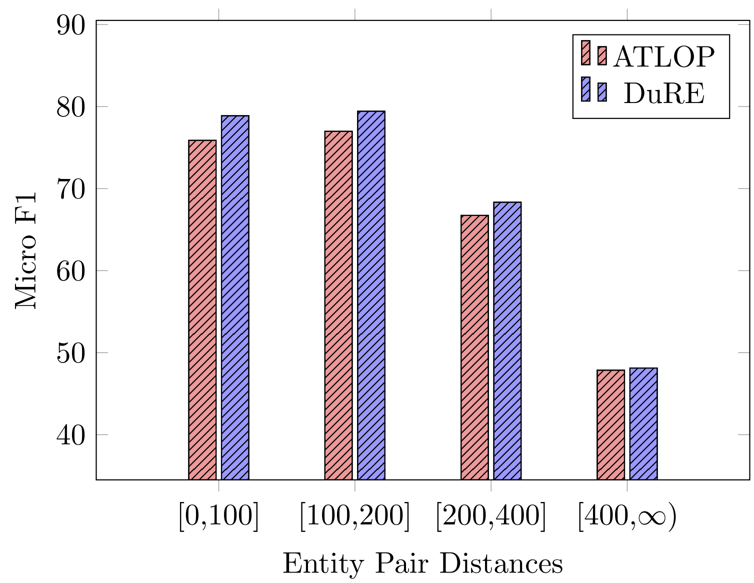


Figure 5.2: Micro F1 score on the development set of Re-DocRED over different distances of entity pairs. The distance is defined as the number of tokens between the head and tail entity.

open problem.

There are possible solutions to overcome this issue. 1) Improving LM's ability to understand long documents. This relates to the advancement in basic abilities of large language models. 2) Combination of rule learning and knowledge graph completion (KGC) [11]. Another way might be segmenting the long documents into pieces, running RE classifiers on such pieces, aggregating the results in graph-like structure, and annotating missing edges in the knowledge graphs by rule learning [89] or other knowledge graph completion methods [110]. Our proposed self-training approaches might potentially improve the quality of knowledge graph completion by annotating and learning from pseudo labels in KGC.

Acceleration of Self-Training In self-training-like methods, we need to reproduce pseudo labels and pseudo text at each ST iteration. In DuNST (Section 3.3) and DuRE (Chapter 4), the pseudo text is generated in an auto-regressive manner, which is hard to be done in parallel and takes longer training time. In KEST (Section 3.4) we propose to use a non-autoregressive method to generate pseudo text to accelerate the training speed. However, the current KEST method only leverages simple and standard masked language models, which yields suboptimal performance. Recently more advanced non-autoregressive generation methods have been proposed. Some of them use early exit [55] to accelerate the inference speed of non-autoregressive generation and alleviate the issue of repetition. With the emergence of faster and better non-autoregressive algorithms, the training speed and quality of self-training can be potentially improved.

Scalability of Self-Training As we analyzed in Sec. 3.3.2, ST actually acts as a kind of regularization and smoothing. How to apply this paradigm to super large PLMs (*e.g.*, GPT4 and Llama-2[97]), where the supervision signals from limited labeled data become extremely weak, is also an open question. Some recent works [44] demonstrate that a large language model can self-improve by taking datasets without ground truth outputs by leveraging chain-of-thought(CoT) reasoning [114] and self-consistency [107]. They leverage self-training in generating CoT questions and answers, and prove the effectiveness of self-training in achieving higher accuracy results on out-of-domain tasks, showing that the overall reasoning ability of the language model is improved. This is evidence that the method of self-training can be scaled to large language models. In the future, we want to check whether self-training can also be applied in the *downstream alignment* of large language models, such as document-level relation extraction and controllable natural language generation.

Bibliography

[1] Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. Leveraging linguistic structure for open domain information extraction. In Chengqing Zong and Michael Strube, editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1034. URL `https://aclanthology.org/P15-1034`. 2

[2] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In *IJCAI*, pages 2670–2676, 2007. 2, 3, 21

[3] Jan R Benetka, Krisztian Balog, and Kjetil Norvag. Towards building a knowledge base of monetary transactions from a news collection. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–10. IEEE, 2017. 3, 8, 21

[4] Meghana Moorthy Bhat, Alessandro Sordoni, and Subhabrata Mukherjee. Self-training with few-shot rationalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10702–10712, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. 6, 44, 96

[5] Qin Bing, Liu An-an, and Liu Ting. Unsupervised chinese open entity relation extraction. *Journal of Computer Research and Development*, 52:1029, 2015. 8, 19, 22

[6] Robert Bossy, Louise Deléger, Estelle Chaix, Mouhamadou Ba, and Claire Nédellec. Bacteria biotope at bionlp open shared tasks 2019. In *Proceedings of the 5th workshop on BioNLP open shared tasks*, pages 121–131, 2019. xii, 16

[7] Yuri Burda, Roger Baker Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *CoRR*, abs/1509.00519, 2016. 55

[8] Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. Semi-supervised learning. *IEEE Transactions on Neural Networks*, 20, 2006. 94

[9] Wanxiang Che, Zhenghua Li, and T. Liu. Ltp: A chinese language technology platform. In *COLING*, 2010. 24, 32

[10] Yiming Chen, Yan Zhang, Chen Zhang, Grandee Lee, Ran Cheng, and Haizhou Li. Revisiting self-training for few-shot learning of language model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9125–9135, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. 6, 12, 43, 44, 71, 96

[11] Zhe Chen, Yuehan Wang, Bin Zhao, Jing Cheng, Xin Zhao, and Zongtao Duan. Knowledge graph completion: A review. *IEEE Access*, 8:192435–192456, 2020. doi: 10.1109/ACCESS.2020.3030076. 123

[12] Z. Cheng, Xu Wu, Xiaqing Xie, and Jingchen Wu. Chinese open relation extraction with pointer-generator networks. *2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC)*, pages 307–311, 2020. 22

[13] Chung-Ming Chien and Hung-yi Lee. Hierarchical prosody modeling for non-autoregressive speech synthesis. In *SLT 2021*, pages 446–453. IEEE, 2021. 45

[14] Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416, 2022. URL `https://api.semanticscholar.org/CorpusID:253018554`. 101, 104

[15] G. Ciampaglia, Prashant Shiralkar, Luis Mateus Rocha, J. Bollen, F. Menczer, and A. Flammini. Computational fact checking from knowledge networks. *PLoS ONE*, 10, 2015. 1, 19

[16] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=H1edEyBKDS`. 44, 53

[17] 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. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`. 6, 22, 56, 71, 122

[18] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, December 2019. URL `https://www.microsoft.com/en-us/research/publication/unified-language-model-pre-training-for-natural-language-understanding`
41, 56, 60, 77

[19] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610, 2014. 1, 19

[20] Zhendong Dong, Qiang Dong, and Changling Hao. HowNet and its computation of meaning. In *Coling 2010: Demonstrations*, pages 53–56, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL `https://aclanthology.org/C10-3014`. 20, 25, 31

[21] Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. Self-

training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.426. URL `https://aclanthology.org/2021.naacl-main.426`. 6, 10, 44, 96

[22] Georges Dupret and Masato Koda. Bootstrap re-sampling for unbalanced data in supervised learning. *European Journal of Operational Research*, 134(1):141–156, 2001. 100

[23] Sarah Elhammadi, Laks V.S. Lakshmanan, Raymond Ng, Michael Simpson, Baoxing Huai, Zhefeng Wang, and Lanjun Wang. A high precision pipeline for financial knowledge graph construction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 967–977, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.84. URL `https://www.aclweb.org/anthology/2020.coling-main.84`. 3, 8, 22, 23, 24, 26, 29

[24] Yuxi Feng, Xiaoyuan Yi, Laks V.S. Lakshmanan, and Xing Xie. Kest: Kernel distance based efficient self-training for improving controllable text generation. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 5049–5057. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/561. URL `https://doi.org/10.24963/ijcai.2023/561`. Main Track. iv, 41

[25] Yuxi Feng, Xiaoyuan Yi, Xiting Wang, Laks Lakshmanan, V.S., and Xing Xie. DuNST: Dual noisy self training for semi-supervised controllable text generation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8760–8785, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.488. URL `https://aclanthology.org/2023.acl-long.488`. iv, 41, 97, 100, 101, 103

[26] Jessica Ficler and Yoav Goldberg. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on*

Stylistic Variation, pages 94–104, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. 44

[27] Spencer Frei, Difan Zou, Zixiang Chen, and Quanquan Gu. Self-training converts weak learners to strong learners in mixture models. In *International Conference on Artificial Intelligence and Statistics*, 2021. URL https://api.semanticscholar.org/CorpusID:235652115. 7

[28] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. 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 240–250, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1021. URL https://aclanthology.org/N19-1021. 60

[29] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301. 44, 53

[30] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, page 512–520, USA, 2000. Association for Computational Linguistics. doi: 10.3115/1075218.1075283. URL https://doi.org/10.3115/1075218.1075283. 3, 8, 19

[31] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004. 1, 41, 56

[32] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. 72

[33] Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. Non-autoregressive neural machine translation. In *ICLR*, 2018. 45

[34] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL `https://aclanthology.org/2020.acl-main.740`. 44

[35] Chunhui He, Zhen Tan, Haoran Wang, Chong Zhang, Yanli Hu, and Bin Ge. Open domain chinese triples hierarchical extraction method. *Applied Sciences*, 10:4819, 2020. 3, 21, 22

[36] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. *Advances in neural information processing systems*, 29, 2016. 45

[37] Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. Revisiting self-training for neural sequence generation. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SJgdnAVKDH`. 6, 7, 9, 12, 43, 44, 47, 56, 71, 79

[38] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL `http://arxiv.org/abs/1503.02531`. 49

[39] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2019. 60, 102

[40] Jinyi Hu, Xiaoyuan Yi, Wenhao Li, Maosong Sun, and Xing Xie. Fuse it more deeply! a variational transformer with layer-wise latent variable inference for text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 697–716, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.51. URL `https://aclanthology.org/2022.naacl-main.51`. 45, 55, 60

[41] Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and Philip Yu. SelfORE: Self-supervised relational feature learning for open relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3673–3682, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.299. URL `https://aclanthology.org/2020.emnlp-main.299`. 1, 3, 7, 21

[42] Xuming Hu, Chenwei Zhang, Fukun Ma, Chenyao Liu, Lijie Wen, and Philip S. Yu. Semi-supervised relation extraction via incremental meta self-training. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 487–496, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.44. URL `https://aclanthology.org/2021.findings-emnlp.44`. 2, 4, 8, 9, 10, 12, 95, 96

[43] Xuming Hu, Chenwei Zhang, Yawen Yang, Xiaohe Li, Li Lin, Lijie Wen, and Philip S. Yu. Gradient imitation reinforcement learning for low resource relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2737–2746, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.216. URL `https://aclanthology.org/2021.emnlp-main.216`. 1, 3

[44] Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.67. URL `https://aclanthology.org/2023.emnlp-main.67`. 124

[45] Shengbin Jia, E. Shijia, M. Li, and Yang Xiang. Chinese open relation extraction and knowledge base establishment. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17:1 – 22, 2018. 8, 19, 22, 24, 27, 31, 37

[46] Wenxiang Jiao, Xing Wang, Zhaopeng Tu, Shuming Shi, Michael Lyu, and Irwin King. Self-training sampling with monolingual data uncertainty for neural machine translation. In *Proceedings of the 59th Annual*

Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2840–2850, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.221. URL `https://aclanthology.org/2021.acl-long.221`. 6, 9, 44, 74

[47] Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. Better modeling of incomplete annotations for named entity recognition. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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 729–734, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1079. URL `https://aclanthology.org/N19-1079`. 104

[48] Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *ArXiv*, abs/1909.05858, 2019. 44

[49] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations*, 2014. 42, 44, 46

[50] Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. Gedi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, 2021. 44

[51] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2): 167–195, 2015. 1, 19, 94

[52] Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4678–4699, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.378. URL

https://aclanthology.org/2020.emnlp-main.378. 45, 55, 60

[53] Guozheng Li, Peng Wang, and Wenjun Ke. Revisiting large language models as zero-shot relation extractors. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6877–6892, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-emnlp.459. 5, 122

[54] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, June 2016. Association for Computational Linguistics. 55

[55] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. ELMER: A non-autoregressive pre-trained language model for efficient and effective text generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1044–1058, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.68. URL https://aclanthology.org/2022.emnlp-main.68. 124

[56] Shiyang Li, Semih Yavuz, Wenhu Chen, and Xifeng Yan. Task-adaptive pre-training and self-training are complementary for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1006–1015, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. 41

[57] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353. 44, 56

[58] Yangming Li, lemao liu, and Shuming Shi. Empirical analysis of unlabeled entity problem in named entity recognition. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=5jRVa89sZk`. 104

[59] Zhiyun Li, Ning Ding, Z. Liu, H. Zheng, and Ying Shen. Chinese relation extraction with multi-grained information and external linguistic knowledge. In *ACL*, 2019. 22, 37

[60] Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.522. URL `https://aclanthology.org/2021.acl-long.522`. 44

[61] Fangchao Liu, Lingyong Yan, Hongyu Lin, Xianpei Han, and Le Sun. Element intervention for open relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4683–4693, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.361. URL `https://aclanthology.org/2021.acl-long.361`. 3, 21

[62] Puyuan Liu, Chenyang Huang, and Lili Mou. Learning non-autoregressive models from search for unsupervised sentence summarization. In *ACL 2022*, pages 7916–7929, 2022. 45

[63] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019. 5, 6, 41, 122

[64] Yinhan Liu, Jiatao Gu, Naman Goyal, X. Li, Sergey Edunov, Marjan Ghazvininejad, M. Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020. 31, 32, 35

[65] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 60, 79, 101, 102

[66] Xuezhe Ma, Chunting Zhou, Xian Li, Graham Neubig, and Eduard Hovy. Flowseq: Non-autoregressive conditional sequence generation with generative flow. In *EMNLP-IJCNLP*, pages 4282–4292, 2019. 45

[67] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `https://aclanthology.org/P11-1015`. 53

[68] Christopher D. Manning, M. Surdeanu, John Bauer, J. Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *ACL*, 2014. 36

[69] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL `http://www.aclweb.org/anthology/P/P14/P14-5010`. 21

[70] Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. Text classification using label names only: A language model self-training approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9006–9017, Online, November 2020. Association for Computational Linguistics. 6, 44, 96

[71] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL `https://aclanthology.org/P09-1113`. 4, 94

[72] Bhavana Dalvi Mishra, Niket Tandon, and Peter Clark. Domain-targeted, high precision knowledge extraction. *Transactions of the Association for Computational Linguistics*, 5:233–246, 2017. 3, 21

[73] Subhabrata Mukherjee and Ahmed Hassan Awadallah. Uncertainty-aware self-training for few-shot text classification. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, Online, 2020. URL `https://papers.nips.cc/paper/2020/file/f23d125da1e29e34c552f448610ff25f-Paper.pdf`. 6, 41, 44, 56, 57, 60, 74

[74] Tapas Nayak, Navonil Majumder, and Soujanya Poria. Improving distantly supervised relation extraction with self-ensemble noise filtering. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1031–1039, Held Online, September 2021. INCOMA Ltd. URL `https://aclanthology.org/2021.ranlp-1.116`. 4

[75] Natasha Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. Industry-scale knowledge graphs: Lessons and challenges. *Communications of the ACM*, 62 (8):36–43, 2019. URL `https://cacm.acm.org/magazines/2019/8/238342-industry-scale-knowledge-graphs/fulltext`. 1, 19, 94

[76] OpenAI. Gpt-4 technical report, 2023. 5, 122

[77] Yilmazcan Ozyurt, Stefan Feuerriegel, and Ce Zhang. Document-level in-context few-shot relation extraction via pre-trained language models, 2024. 5, 122

[78] Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. Distantly supervised named entity recognition using positive-unlabeled learning. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2409–2419, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1231. URL `https://aclanthology.org/P19-1231`. 104

[79] Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Qiang Dong, Maosong Sun, and Zhendong Dong. Openhownet: An open sememe-based lexical knowledge base. *arXiv preprint arXiv:1901.09957*, 2019. 25, 32

[80] Jing Qian, Li Dong, Yelong Shen, Furu Wei, and Weizhu Chen. Controllable natural language generation with contrastive prefixes. In *Findings of the Association for Computational Linguistics: ACL*

2022, pages 2912–2924, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.229. URL https://aclanthology.org/2022.findings-acl.229. 44, 53, 55, 56

[81] Likun Qiu and Yue Zhang. ZORE: A syntax-based system for Chinese open relation extraction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1870–1880, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1201. URL https://www.aclweb.org/anthology/D14-1201. 22

[82] Yunqi Qiu, Yuanzhuo Wang, Xiaolong Jin, and K. Zhang. Stepwise reasoning for multi-relation question answering over knowledge graph with weak supervision. *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020. 1, 19

[83] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 41, 55, 56

[84] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html. 41, 56

[85] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Neural Information Processing Systems*, 2015. URL https://api.semanticscholar.org/CorpusID:5855183. 7

[86] Roi Reichart and Ari Rappoport. Self-training for enhancement and domain adaptation of statistical parsers trained on small datasets. In Annie Zaenen and Antal van den Bosch, editors, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 616–623, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/P07-1078. 6

[87] Leonardo F. R. Ribeiro, Yue Zhang, and Iryna Gurevych. Structural adapters in pretrained language models for AMR-to-Text generation. In *Proceedings of the 2021 Conference on Empirical Methods*

in Natural Language Processing, pages 4269–4282, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.351. URL `https://aclanthology.org/2021.emnlp-main.351`. 44

[88] Youngbin Ro, Yukyung Lee, and Pilsung Kang. Multiˆ2OIE: Multilingual open information extraction based on multi-head attention with BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1107–1117, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.99. URL `https://aclanthology.org/2020.findings-emnlp.99`. 3, 21

[89] Dongyu Ru, Changzhi Sun, Jiangtao Feng, Lin Qiu, Hao Zhou, Weinan Zhang, Yong Yu, and Lei Li. Learning logic rules for document-level relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1239–1250, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.emnlp-main.95`. 1, 3, 94, 122, 123

[90] Timo Schick and Hinrich Schütze. Generating datasets with pretrained language models. In *EMNLP 2021*, pages 6943–6951, 2021. 10, 42

[91] Henry Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965. 1, 6, 41, 95

[92] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, Online, November 2020. Association for Computational Linguistics. 44

[93] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin Dogus Cubuk, Alexey Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *ArXiv*, abs/2001.07685, 2020. 6, 44

[94] Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. Document-level relation extraction with adaptive focal loss and knowledge distillation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio,

editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.132. URL https://aclanthology.org/2022.findings-acl.132. 3, 96

[95] Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. Revisiting DocRED - addressing the false negative problem in relation extraction. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.580. URL https://aclanthology.org/2022.emnlp-main.580. xii, 7, 9, 17, 100

[96] Qingyu Tan, Lu Xu, Lidong Bing, and Hwee Tou Ng. Class-adaptive self-training for relation extraction with incompletely annotated training data. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8630–8643, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.549. URL https://aclanthology.org/2023.findings-acl.549. xi, 7, 9, 10, 12, 15, 95, 96, 99, 102, 103, 104, 105

[97] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. 124

[98] Thy Thy Tran, Phong Le, and Sophia Ananiadou. Revisiting unsupervised relation extraction. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7498–7505, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.669. URL `https://aclanthology.org/2020.acl-main.669`. 1

[99] Bo-Hsiang Tseng, Jianpeng Cheng, Yimai Fang, and David Vandyke. A generative model for joint natural language understanding and generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1795–1807, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.163. URL `https://aclanthology.org/2020.acl-main.163`. 45

[100] Yuen-Hsien Tseng, Lung-Hao Lee, Shu-Yen Lin, Bo-Shun Liao, Mei-Jun Liu, Hsin-Hsi Chen, Oren Etzioni, and Anthony Fader. Chinese open relation extraction for knowledge acquisition. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 12–16, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-4003. URL `https://www.aclweb.org/anthology/E14-4003`. 22

[101] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3, 10, 42, 47

[102] Tu Vu, Minh-Thang Luong, Quoc Le, Grady Simon, and Mohit Iyyer. STraTA: Self-training with task augmentation for better few-shot learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5715–5731, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.462. URL `https://aclanthology.org/2021.emnlp-main.462`. 6, 7, 44, 60, 74, 96, 98

[103] Somin Wadhwa, Silvio Amir, and Byron Wallace. Revisiting relation extraction in the era of large language models. In Anna Rogers, Jordan

Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.868. URL `https://aclanthology.org/2023.acl-long.868`. 5, 122

[104] Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. *Advances in Neural Information Processing Systems*, 30, 2017. 60

[105] Xiang Wang, X. He, Yixin Cao, Meng Liu, and Tat-Seng Chua. Kgat: Knowledge graph attention network for recommendation. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019. 1, 19

[106] Xuan Wang, Yu Zhang, Qi Li, Yinyin Chen, and Jiawei Han. Open information extraction with meta-pattern discovery in biomedical literature. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 291–300, 2018. 3, 21

[107] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=1PL1NIMMrw`. 124

[108] Yaqing Wang, Subhabrata Mukherjee, Haoda Chu, Yuancheng Tu, Ming Wu, Jing Gao, and Ahmed Hassan Awadallah. Adaptive self-training for few-shot neural sequence labeling. *arXiv preprint arXiv:2010.03680*, 2020. 9

[109] Ye Wang, Xinxin Liu, Wenxin Hu, and Tao Zhang. A unified positive-unlabeled learning framework for document-level relation extraction with different levels of labeling. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4123–4135, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.276.

URL https://aclanthology.org/2022.emnlp-main.276. 3, 96, 104

[110] Yun Cheng Wang, Xiou Ge, Bin Wang, and C.-C. Jay Kuo. GreenKGC: A lightweight knowledge graph completion method. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10596–10613, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.591. URL https://aclanthology.org/2023.acl-long.591. 123

[111] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Loddon Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10852–10861, 2021. 7, 44

[112] Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=rC8sJ4i6kaH. 7, 9, 12, 41, 43, 71, 95, 96, 100, 104

[113] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022. 5, 122

[114] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903, 2022. URL https://arxiv.org/abs/2201.11903. 5, 124

[115] 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. 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, October 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`. 60, 79, 102

[116] Yingce Xia, Tao Qin, Wei Chen, Jiang Bian, Nenghai Yu, and Tie-Yan Liu. Dual supervised learning. In *International conference on machine learning*, pages 3789–3798. PMLR, 2017. 45

[117] Qizhe Xie, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. Self-training with noisy student improves imagenet classification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2020. 7, 12, 43, 44

[118] Chenyan Xiong, R. Power, and Jamie Callan. Explicit semantic ranking for academic search via knowledge graph embedding. *Proceedings of the 26th International Conference on World Wide Web*, 2017. 1, 19

[119] Benfeng Xu, Quan Wang, Yajuan Lyu, Dai Dai, Yongdong Zhang, and Zhendong Mao. S2ynRE: Two-stage self-training with synthetic data for low-resource relation extraction. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8186–8207, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.455. URL `https://aclanthology.org/2023.acl-long.455`. 96

[120] Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. How to unleash the power of large language models for few-shot relation extraction? In Nafise Sadat Moosavi, Iryna Gurevych, Yufang Hou, Gyuwan Kim, Young Jin Kim, Tal Schuster, and Ameeta Agrawal, editors, *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 190–200, Toronto, Canada (Hybrid), July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.sustainlp-1.13. URL `https://aclanthology.org/2023.sustainlp-1.13`. 5, 122

[121] Zihan Xu, Hai-Tao Zheng, Zuoyou Fu, and Wei Wang. Enhancing question understanding and representation for knowledge base relation detection. In *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018*, pages 1362–1367.

IEEE Computer Society, 2018. doi: 10.1109/ICDM.2018.00186. URL `https://doi.org/10.1109/ICDM.2018.00186`. 1, 19

[122] Nianwen Xue. Labeling chinese predicates with semantic roles. *Computational Linguistics*, 34:225–255, 2008. 26, 32

[123] Yan Yan and Yuhong Guo. Partial label learning with batch label correction. In *AAAI Conference on Artificial Intelligence*, 2020. 9

[124] Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.276. URL `https://aclanthology.org/2021.naacl-main.276`. 44

[125] Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. Generative data augmentation for commonsense reasoning. In *Findings of EMNLP*, pages 1008–1025, 2020. 10, 42

[126] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. DocRED: A large-scale document-level relation extraction dataset. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1074. URL `https://aclanthology.org/P19-1074`. 7, 96

[127] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196, 1995. 6, 41, 95

[128] Dani Yogatama, Cyprien de Masson d'Autume, Jerome T. Connor, Tomás Kociský, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. Learning and evaluating general linguistic intelligence. *ArXiv*, abs/1901.11373, 2019. 44

[129] Junjie Yu, Xing Wang, Jiangjiang Zhao, Chunjie Yang, and Wenliang Chen. STAD: Self-training with ambiguous data for low-resource relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2044–2054, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL `https://aclanthology.org/2022.coling-1.178`. 2, 4, 7, 8, 9, 95, 96

[130] Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. Double graph based reasoning for document-level relation extraction. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1630–1640, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.127. URL `https://aclanthology.org/2020.emnlp-main.127`. 3, 96

[131] Dongxu Zhang, Sunil Mohan, Michaela Torkar, and Andrew McCallum. A distant supervision corpus for extracting biomedical relationships between chemicals, diseases and genes. In *Proceedings of The 13th Language Resources and Evaluation Conference*, Marseille, France, June 2022. European Language Resources Association. 101

[132] Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. Document-level relation extraction as semantic segmentation. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3999–4006. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/551. URL `https://doi.org/10.24963/ijcai.2021/551`. Main Track. 3, 96

[133] Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen, and Jinjun Xiong. How unlabeled data improve generalization in self-training? a one-hidden-layer theoretical analysis. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=qiMXBIf4NfB`. 9, 41, 95

[134] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. Revisiting few-sample {bert} fine-tuning. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=cO1IH43yUF`. 41, 44, 79

[135] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015. 53

[136] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1004. URL `https://aclanthology.org/D17-1004`. 96

[137] Yao Zhao, Mikhail Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. Calibrating sequence likelihood improves conditional language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=0qSOodKmJaN`. 12, 95, 99

[138] Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. xii, 1, 3, 5, 9, 17, 94, 95, 96, 98, 101, 104, 122

[139] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & amp; Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3210080. URL `https://doi.org/10.1145/3209978.3210080`. 55