NSP-BERT: A Prompt-based Zero-Shot Learner Through an Original Pre-training Task —— Next Sentence Prediction

NSP-BERT：通过原始预训练任务的基于提示的零样本学习器——下一句预测

摘要Abstract

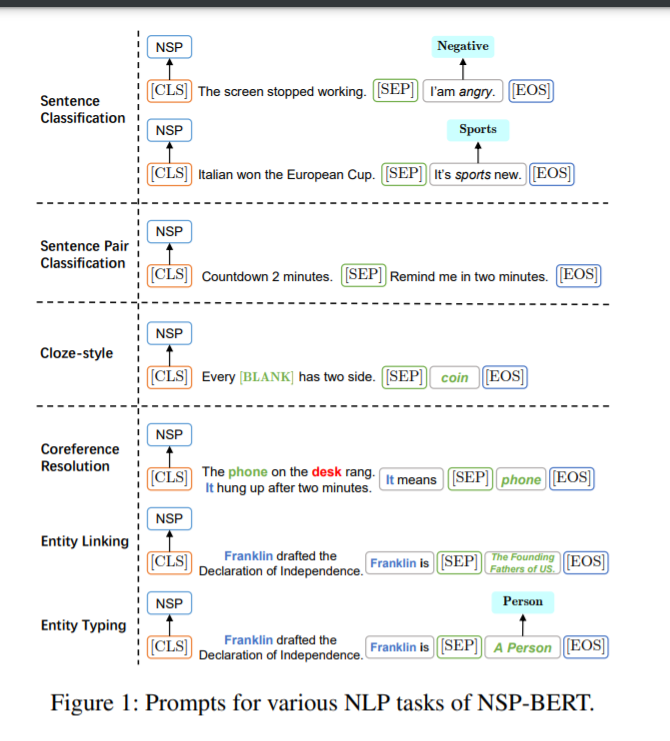
Using prompts to utilize language models to perform various downstream tasks, also known as prompt-based learning or prompt-learning, has lately gained significant success in comparison to the pre-train and fine-tune paradigm. Nonetheless, virtually all prompt-based methods are tokenlevel, meaning they all utilize GPT’s left-to-right language model or BERT’s masked language model to perform clozestyle tasks. In this paper, we attempt to accomplish several NLP tasks in the zero-shot scenario using a BERT original pre-training task abandoned by RoBERTa and other models—Next Sentence Prediction (NSP). Unlike token-level techniques, our sentence-level prompt-based method NSPBERT does not need to fix the length of the prompt or the position to be predicted, allowing it to handle tasks such as entity linking with ease. Based on the characteristics of NSPBERT, we offer several quick building templates for various downstream tasks. We suggest a two-stage prompt method for word sense disambiguation tasks in particular. Our strategies for mapping the labels significantly enhance the model’s performance on sentence pair tasks. On the FewCLUE benchmark, our NSP-BERT outperforms other zero-shot methods on most of these tasks and comes close to the few-shot methods.

使用提示利用语言模型执行各种下游任务，也称为基于提示的学习或提示学习，与预训练和微调范式相比，最近取得了重大成功。尽管如此，几乎所有基于提示的方法都是令牌级别的，这意味着它们都利用 GPT 的从左到右的语言模型或 BERT 的掩码语言模型来执行完形填空式任务。在本文中，我们尝试使用 RoBERTa 和其他模型放弃的 BERT 原始预训练任务——Next Sentence Prediction (NSP) 来完成零样本场景中的几个 NLP 任务。与标记级技术不同，我们的基于句子级提示的方法 NSPBERT 不需要固定提示的长度或要预测的位置，使其能够轻松处理实体链接等任务。基于 NSPBERT 的特点，我们为各种下游任务提供了几种快速构建模板。我们建议特别针对词义消歧任务采用两阶段提示方法。我们映射标签的策略显着提高了模型在句子对任务上的性能。在FewCLUE 基准测试中，我们的NSP-BERT 在这些任务中的大多数任务上都优于其他零样本方法，并且接近于少样本方法。

# 1 介绍introduction

GPT-2 (up to 1.5B (Radford et al. 2019)) and GPT-3 (up to 175B (Brown et al. 2020)) are ultra-large-scale language models with billions of parameters that have recently demonstrated outstanding performance in various NLP tasks. Compared with previous state-of-the-art finetuning methods, they can achieve competitive results without any or with just a limited quantity of training data. Although studies have shown that scaling up the model improves task-agnostic and few-shot performance, some studies have shown that by constructing appropriate prompts for the model, models like BERT (Devlin et al. 2018) or RoBERTa (Liu et al. 2019) can achieve similar performance despite having a parameter count that is several orders of magnitude smaller (Schick and Schutze ¨ 2021b,a; Wang et al. 2021).

GPT-2（高达 1.5B（Radford et al. 2019））和 GPT-3（高达 175B（Brown et al. 2020））是具有数十亿参数的超大规模语言模型，最近表现出出色的性能 在各种 NLP 任务中。 与之前最先进的微调方法相比，它们可以在没有任何训练数据或仅使用有限数量的训练数据的情况下获得有竞争力的结果。 尽管研究表明扩大模型可以提高与任务无关和少拍的性能，但一些研究表明，通过为模型构建适当的提示，BERT (Devlin et al. 2018) 或 RoBERTa (Liu et al. 2019) 等模型 ) 可以实现类似的性能，尽管参数数量要小几个数量级 (Schick and Schutze ¨ 2021b,a; Wang et al. 2021)。



Since then, the area of natural language processing has seen a fresh wave of developments, including the introduction of a new paradigm known as prompt-based learning or prompt-learning, which follows the ”pre-train, prompt, and predict” (Liu et al. 2021) process. In zero-shot and fewshot learning, prompt-learning has achieved a lot of success. Not only does it achieve outstanding performance, promptlearning better integrates pre-training and downstream tasks and brings NLP tasks closer to human logic and habits.

从那时起，自然语言处理领域出现了新的发展浪潮，包括引入了一种称为基于提示的学习或提示学习的新范式，它遵循“预训练、提示和预测”（刘 等人。2021）过程。 在零样本和少样本学习中，即时学习取得了很大的成功。 不仅表现出色，promptlearning 更好地整合了预训练和下游任务，让 NLP 任务更贴近人类的逻辑和习惯。

The input text for the classification task, for example, “The Italian team won the European Cup.”, should be assigned to one of the candidate labels, such as Gaming, Sports, or Finance. At this point, the template “This is [MASK] news.” will be added to the original text, and the model will be asked to predict the missing word or span. The model’s output will then be mapped to the candidate labels. We could utilize the pre-training tasks of several types of language models (LM) to predict the abovementioned templates, including but not limited to Left-to-right LM (GPT series (Radford et al. 2018, 2019; Brown et al. 2020)), Masked LM (BERT (Devlin et al. 2018), RoBERTa (Liu et al. 2019)), prefix LM (UniLM (Dong et al. 2019; Bao et al. 2020)) and Encoder-decoder LM (T5 (Raffel et al. 2019), BART (Lewis et al. 2020)).

分类任务的输入文本，例如“意大利队赢得欧洲杯”，应分配给候选标签之一，例如 Gaming、Sports 或 Finance。 此时，模板“This is [MASK] news”。 将被添加到原始文本中，并要求模型预测丢失的单词或跨度。 然后将模型的输出映射到候选标签。 我们可以利用几种语言模型（LM）的预训练任务来预测上述模板，包括但不限于从左到右的 LM（GPT 系列（Radford et al. 2018, 2019; Brown et al. 2020))、Masked LM (BERT (Devlin et al. 2018)、RoBERTa (Liu et al. 2019))、前缀 LM (UniLM (Dong et al. 2019; Bao et al. 2020)) 和 Encoder-decoder LM ( T5（Raffel 等人，2019 年），BART（刘易斯等人，2020 年））。0

Although most research on prompt-learning has been conducted, the majority of the pre-training tasks used in promptlearning are token-level, requiring the labels to be mapped to a fixed-length token span (Schick and Schutze ¨ 2021b,a; Cui et al. 2021). On the one hand, when the number of labels grows rapidly, this necessitates a lot of human labor. On the other hand, tasks with variable-length options make Left-toright LM (L2R LM) or masked LM (MLM) difficult to cope with. The length of each candidate entity’s description, for example, varies significantly in the entity linking task.

尽管已经进行了大多数关于提示学习的研究，但提示学习中使用的大多数预训练任务都是令牌级别的，需要将标签映射到固定长度的令牌跨度（Schick and Schutze ¨ 2021b,a; Cui 等人，2021 年）。 一方面，当标签数量快速增长时，这就需要大量的人力。 另一方面，具有可变长度选项的任务使 Left-toright LM (L2R LM) 或 masked LM (MLM) 难以处理。 例如，每个候选实体的描述长度在实体链接任务中差异很大。

At the same time, we observed that there is an original sentence-level pre-training object in vanilla BERT——NSP (Next Sentence Prediction), which is a binary classification task that predicts whether two sentences appear consecutively within a document or not. Many models, like RoBERTa (Liu et al. 2019) and many others (Conneau and Lample 2019; Yang et al. 2019; Joshi et al. 2020), have questioned and abandoned this task during pre-training. Nevertheless, based on the task’s features and object, we believe it is appropriate to use in prompt-learning.

同时，我们观察到在vanilla BERT中有一个原始的句子级预训练对象——NSP（Next Sentence Prediction），这是一个二元分类任务，预测两个句子在一个文档中是否连续出现。 许多模型，如 RoBERTa (Liu et al. 2019) 和许多其他模型 (Conneau and Lample 2019; Yang et al. 2019; Joshi et al. 2020)，在预训练期间质疑并放弃了这项任务。 尽管如此，基于任务的特征和对象，我们认为它适合用于即时学习。

Unlike most prior work, we present NSP-BERT, a sentence-level prompt-learning method. The paper’s main contributions can be summarized as follows:

• We propose the use of NSP, a sentence-level pre-training task for prompt-learning. On the FewCLUE benchmark, NSP-BERT has achieved the SOTA performance among zero-shot models without using any task-specific training data. Its performance is comparable to that of several few-shot learning methods.

• Based on the features of the downstream tasks, we propose two alternative label/answer mapping methods that significantly improved prompt-learning performance in the sentence-pair task.

• We suggest to use soft-position and two-stage prompt construction methods to alleviate the problem that sentence-level prompt-based models are not sensitive to token positions, which further improves the performance of NSP-BERT on word sense disambiguation tasks.

• We demonstrate that a simple sentence-level contrastive learning pre-training task on interactive models can fit prompt-based learning well and solve various NLP tasks. It is very inspiring for zero-shot and few-shot learning.

与大多数先前的工作不同，我们提出了 NSP-BERT，一种句子级提示学习方法。该论文的主要贡献可以总结如下：

• 我们建议使用NSP，这是一种用于提示学习的句子级预训练任务。在FewCLUE 基准测试中，NSP-BERT 在不使用任何特定任务的训练数据的情况下，在零样本模型中实现了 SOTA 性能。其性能可与几种少样本学习方法相媲美。

• 基于下游任务的特征，我们提出了两种替代的标签/答案映射方法，它们显着提高了句子对任务中的提示学习性能。

• 我们建议使用软定位和两阶段提示构建方法来缓解基于句子级提示的模型对标记位置不敏感的问题，这进一步提高了 NSP-BERT 在词义消歧任务上的性能。

• 我们证明了一个简单的基于交互模型的句子级对比学习预训练任务可以很好地适应基于提示的学习并解决各种 NLP 任务。对于零样本和少样本学习非常有启发性。

# 2 相关工作

Many studies on prompt-learning for zero-shot or few-shot have been conducted. This section focuses on models at different levels and optimization methods.

已经进行了许多关于零样本或少样本的即时学习的研究。 本节重点介绍不同级别的模型和优化方法。

## 2.1 Token-Level and Sentence-Level令牌级和句子级

Token-Level Prompt-Learning Token-level pre-training tasks, such as MLM (Shown in the left part of Figure 2) (Jiang et al. 2020; Schick and Schutze ¨ 2021b,a) or L2R LM(Radford et al. 2019; Brown et al. 2020; Cui et al. 2021), are commonly used in token-level prompt-learning approaches. Although the expected answer may be in the form of tokens, spans, or sentences in token-level promptlearning, the predicted answer is always generated token by token. Tokens are usually mapped to the whole vocabulary or a set of candidate words (Petroni et al. 2019; Cui et al. 2021; Han et al. 2021; Adolphs, Dhuliawala, and Hofmann 2021; Hu et al. 2021). Take PET model (Schick and Schutze ¨ 2021b,a) as an example, the sentiment classification input/label pair is reformulated to “x: [CLS] The Italian team won the European Cup. This is [MASK] news. [EOS], y: Sports”.

Token-Level Prompt-Learning Token-level 预训练任务，例如 MLM（如图 2 左侧部分所示）（Jiang et al. 2020; Schick and Schutze ¨ 2021b,a）或 L2R LM（Radford et al. 2019; Brown et al. 2020; Cui et al. 2021)，通常用于令牌级提示学习方法。 尽管在令牌级提示学习中，预期答案可能是令牌、跨度或句子的形式，但预测答案始终是逐个令牌生成的。 标记通常映射到整个词汇表或一组候选词（Petroni et al. 2019; Cui et al. 2021; Han et al. 2021; Adolphs, Dhuliawala, and Hofmann 2021; Hu et al. 2021）。 以PET模型（Schick and Schutze ¨ 2021b,a）为例，将情感分类输入/标签对重新表述为“x：[CLS]意大利队赢得欧洲杯。 这是 [MASK] 新闻。 [EOS]，y：体育”。

Sentence-Level Prompt-Learning Sentence-level methods concentrate on the relationship between sentences, with the model’s output usually mapped to a relationship space. As far as we know, EFL (Wang et al. 2021) is the only sentence-level model. It reformulates NLP tasks into sentence entailment-style tasks. For example, the sentiment classification input/label pair is reformulated to “x: [CLS] The Italian team won the European Cup. [SEP] This is Sports news.[EOS], y: Entail”. The output of model is Entail or Not Entail. The EFL model can perform well on few-shot learning but not on Zero-shot tasks unless it is trained on labeled natural language inference (NLI) datasets like MNLI (Williams, Nangia, and Bowman 2018).

Sentence-Level Prompt-Learning Sentence-Level 方法专注于句子之间的关系，模型的输出通常映射到关系空间。 据我们所知，EFL (Wang et al. 2021) 是唯一的句子级模型。 它将 NLP 任务重新表述为句子蕴涵式任务。 例如，情感分类输入/标签对被重新表述为“x：[CLS]意大利队赢得了欧洲杯。 [SEP] 这是体育新闻。[EOS]，y：Entail”。 模型的输出是 Entail 或 Not Entail。 EFL 模型可以在少样本学习上表现良好，但在零样本任务上表现不佳，除非它在标记的自然语言推理 (NLI) 数据集上进行训练，例如 MNLI (Williams、Nangia 和 Bowman 2018)。

## 2.2 Optimization methods优化方法

Automated Prompt Manually designed prompts are highly unstable. Sometimes it is necessary to be familiar with the particular task and language model in order to construct a high-quality prompt. As a result, several studies attempt to automatically search for and generate prompts. LMBFF (Gao, Fisch, and Chen 2021) model use conditional likelihood to automatically select labels words, and use T5 (Raffel et al. 2019) to generate templates. AUTOPROMPT (Shin et al. 2020) uses a gradient-guided search to create prompts. Compared to the discrete prompt search methods mentioned above, P-tuning (Liu et al. 2021) employs trainable continuous prompt embeddings, with P-tuning, GPTs achieve comparable and sometimes better performance to similar-sized BERTs in supervised learning.

自动提示 手动设计的提示非常不稳定。 有时需要熟悉特定的任务和语言模型才能构建高质量的提示。 因此，多项研究试图自动搜索和生成提示。 LMBFF (Gao, Fisch, and Chen 2021) 模型使用条件似然自动选择标签词，并使用 T5 (Raffel et al. 2019) 生成模板。 AUTOPROMPT (Shin et al. 2020) 使用梯度引导搜索来创建提示。 与上述离散提示搜索方法相比，P-tuning (Liu et al. 2021) 采用可训练的连续提示嵌入，通过 P-tuning，GPT 在监督学习中实现了与类似大小的 BERT 相当甚至更好的性能。

Training Strategy There are many optimization methods in prompt-learning. ADAPET (Tam et al. 2021) uses more supervision by decoupling the losses for the label tokens and a label-conditioned MLM objective over the full original input. PTR (Han et al. 2021) incorporates logic rules to compose task-specific prompts with several simple sub-prompts. (Zhao et al. 2021) pointed out that there are 3 types of bias (majority label bias, recency bias and common token bias) in GPT. By using content-free inputs (e.g. “N/A”) to calibrate the model’s output probabilities, the performance of GPT-2 and GPT-3 has been substantially improved.

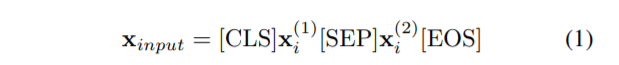
训练策略 即时学习中有许多优化方法。 ADAPET（Tam 等人，2021 年）通过在完整的原始输入上分离标签令牌的损失和标签条件的 MLM 目标来使用更多的监督。 PTR (Han et al. 2021) 结合了逻辑规则，用几个简单的子提示来组成特定于任务的提示。 (Zhao et al. 2021) 指出 GPT 中存在 3 种类型的偏差（多数标签偏差、新近偏差和普通标记偏差）。 通过使用无内容输入（例如“N/A”）来校准模型的输出概率，GPT-2 和 GPT-3 的性能得到了显着提高。

# 3 框架

## 3.1 Next Sentence Prediction下一句预测

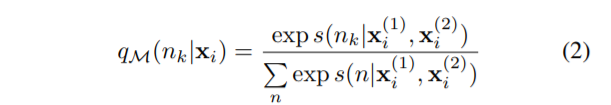
The next sentence prediction is one of the two basic pretraining tasks (the other is MLM) of the vanilla BERT model (Devlin et al. 2018) (Shown in the right part of Figure 2). This task inputs two sentences A and B into BERT at the same time to predict whether sentence B comes after sentence A in the same document. During specific training, for 50% of the time, B is the actual next sentence that follows A (IsNext), and for the other 50% of the time, we use a random sentence from the corpus (NotNext).

下一句预测是vanilla BERT模型（Devlin et al. 2018）的两个基本预训练任务之一（另一个是MLM）（如图2右侧部分所示）。 该任务同时将两个句子 A 和 B 输入到 BERT 中，以预测句子 B 是否在同一文档中的句子 A 之后。 在特定训练中，有 50% 的时间，B 是 A 之后的实际下一个句子（IsNext），另外 50% 的时间，我们使用语料库中的随机句子（NotNext）。



Let M denote the model trained on a large-scale corpus. This model is trained on both MLM task and NSP task at the same time. x (1) i and x (2) i denote sentence A and sentence B, respectively. The model’s input is xinput, and qM denotes the output probability of model’s NSP head (Eq. 2). s = Wnsph[CLS], where h[CLS] is the hidden vector of [CLS] and Wnsp is a matrix learned by NSP task, Wnsp ∈ R 2×H. The loss function of NSP task LNSP = − log qM(n|x), where n ∈ {IsNext, NotNext}.

让 M 表示在大规模语料库上训练的模型。 该模型同时在 MLM 任务和 NSP 任务上进行训练。 x (1) i 和 x (2) i 分别表示句子 A 和句子 B。 模型的输入是 xinput，qM 表示模型的 NSP 头部的输出概率（等式 2）。 s = Wnsph[CLS]，其中 h[CLS] 是 [CLS] 的隐藏向量，Wnsp 是 NSP 任务学习的矩阵，Wnsp ∈ R 2×H。 NSP 任务 LNSP = − log qM(n|x) 的损失函数，其中 n ∈ {IsNext, NotNext}。



NSP is a self-supervised task that is simple and weak. We believe the task is more likely to judge whether two sentences are from the same document since the negative sample is randomly picked from another unrelated document. In other words, rather of determining the order of two phrases, the NSP task may determine if they have the same topic and express the same semantics.

NSP 是一种简单而弱的自我监督任务。 我们认为该任务更有可能判断两个句子是否来自同一个文档，因为负样本是从另一个不相关的文档中随机选取的。 换句话说，NSP 任务不是确定两个短语的顺序，而是确定它们是否具有相同的主题并表达相同的语义。

The NSP task is quite similar to a contrastive learning task, as shown in Figure 3. So, does the NSP just compare sentence similarities or does it have the ability to reason logically? The following are the major reasons why we believe NSP has logical reasoning ability:

• The NSP task is interactive. Tokens in one sentence could interact with their own tokens while also interacting with tokens in the other sentence.

• The NSP task is trained alongside the MLM task. The MLM task provides a training basis for the self-attention mechanism of the entire model.

NSP 任务与对比学习任务非常相似，如图 3 所示。那么，NSP 只是比较句子相似性还是具有逻辑推理能力？ 以下是我们认为 NSP 具有逻辑推理能力的主要原因：

• NSP 任务是交互式的。 一个句子中的标记可以与自己的标记交互，同时也可以与另一句子中的标记交互。

• NSP 任务与 MLM 任务一起训练。 MLM任务为整个模型的self-attention机制提供了训练基础。

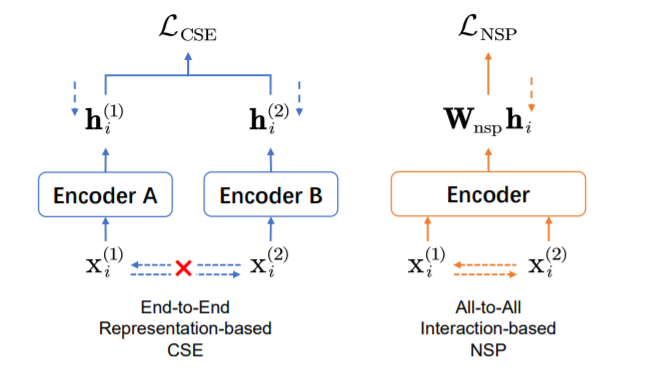


Figure 3: Conceptual comparison between End-to-End representation-based contrastive learning of sentence embeddings (CSE) and All-to-All interaction-based next sentence prediction (NSP). Except that the output of the model is not the representation of the sentence, the NSP task uses a weak self-supervision method to train the BERT。

图 3：基于端到端表示的句子嵌入对比学习 (CSE) 与基于 All-to-All 交互的下一句预测 (NSP) 之间的概念比较。 除了模型的输出不是句子的表示外，NSP 任务使用弱自监督的方法来训练 BERT

NSP-BERT is a true prompt-based learner, not a sentence similarity matcher, as determined by the above two points. This will be confirmed in our experiments. The model performs better the closer the template is to a fluent and logical natural language sentence.

NSP-BERT 是真正的基于提示的学习器，而不是句子相似度匹配器，由以上两点决定。 这将在我们的实验中得到证实。 模板越接近流畅且符合逻辑的自然语言句子，模型的性能就越好。

## 3.2 Prompts in NSP-BERT NSP-BERT 中的提示

NSP-BERT, like other prompt-based learning methods, requires the construction of appropriate templates for various tasks. Since NSP-BERT does not rely on the training data of any downstream tasks, the template’s building form must closely match the original NSP task. In this section, we’ll show how to construct templates for different tasks.

NSP-BERT 与其他基于提示的学习方法一样，需要为各种任务构建合适的模板。 由于 NSP-BERT 不依赖任何下游任务的训练数据，因此模板的构建形式必须与原始 NSP 任务紧密匹配。 在本节中，我们将展示如何为不同的任务构建模板。

Single Sentence Task Samples must be classified into different topics in the single sentence task. Sentiment analysis, for example, is the classification of texts into various sentiment trends. Suppose that the training dataset of a single sentence classification task D = {(xi , yi)} N i=1, xi is the ith sentence in the total N samples, and the label of xi is yi , which can be mapped to y (j) ∈ Y, where |Y| = M, M is the number of topics in this dataset. For each y (j) , it will be mapped to a template p (j) ∈ P. And the input of the model will be,

单句任务样本必须在单句任务中分类为不同的主题。 例如，情感分析是将文本分类为各种情感趋势。 假设单句分类任务的训练数据集 D = {(xi , yi)} N i=1, xi 是总共 N 个样本中的第 i 个句子，xi 的标签为 yi ，可以映射到 y (j) ∈ Y，其中 |Y| = M，M 是这个数据集中的主题数。 对于每个 y (j) ，它将被映射到一个模板 p (j) ∈ P。模型的输入将是，



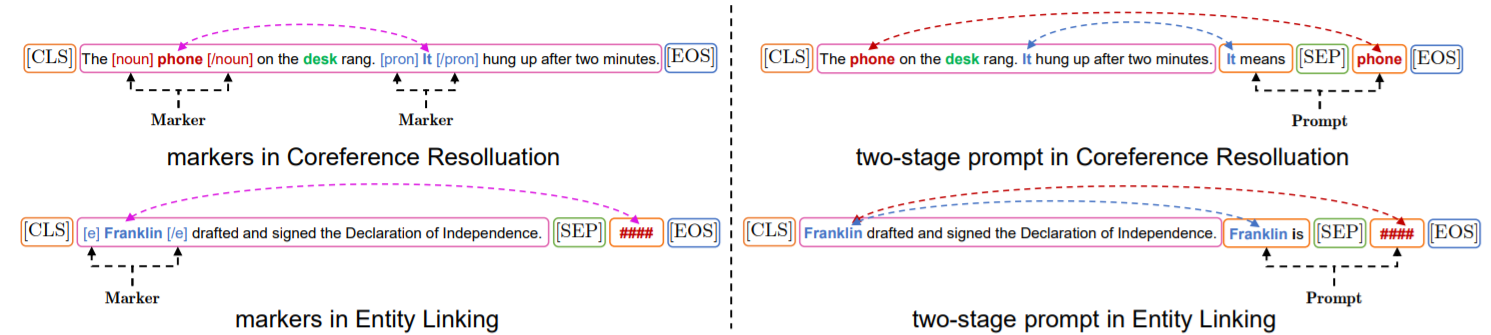
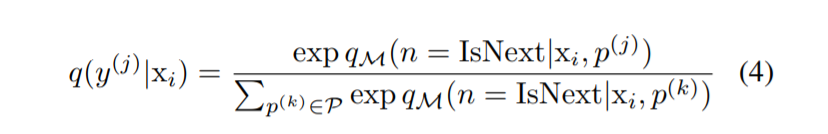


Figure 4: The comparison of markers (Left) and two-stage prompt (Right), examples in coreference resolution and entity linking/typing tasks.

图 4：标记（左）和两阶段提示（右）的比较、共指解析和实体链接/键入任务中的示例。

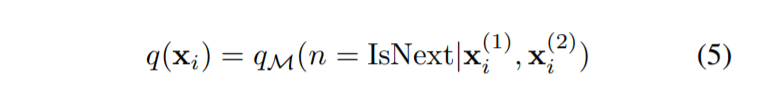
the probability when the label of sample xi is y (j) is:

样本xi的标签为y(j)时的概率为：



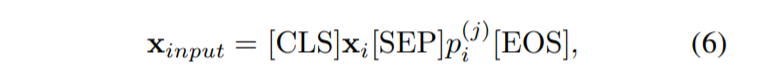
Sentence Pair Task The sentence pair tasks aim to identify the relationship between two sentences. Such as OCNLI (Hu et al. 2020), BUSTM (of OPPO XiaoBu 2021) and CSL (Xu et al. 2020) used in this paper. This type of dataset D = {(x (1) i , x (2) i , yi)} N i=1 contains N samples, each with 2 sentences x (1) i and x (2) i . The relationship between them is yi , which can be mapped to y (j) ∈ Y, where |Y| = M, is the number of relationship types. Its input is the same as Eq. 1. The output of the NSP model qM(xi) is shown in Eq. 5. (We do not directly associate the output of the NPS model directly with the labels here.)

句子对任务句子对任务旨在识别两个句子之间的关系。 如本文使用的OCNLI (Hu et al. 2020)、BUSTM (of OPPO XiaoBu 2021) 和CSL (Xu et al. 2020)。 这种类型的数据集 D = {(x (1) i , x (2) i , yi)} N i=1 包含 N 个样本，每个样本有 2 个句子 x (1) i 和 x (2) i 。 它们之间的关系是 yi ，可以映射到 y (j) ∈ Y，其中 |Y| = M，是关系类型的数量。 它的输入与方程相同。 1. NSP 模型 qM(xi) 的输出如等式所示。 5.（我们不直接将 NPS 模型的输出与这里的标签直接关联。）



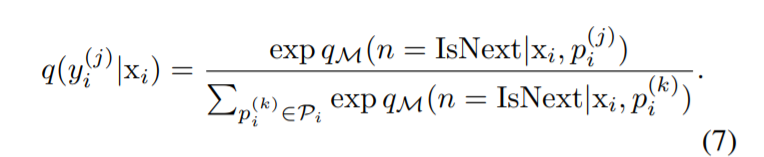
Cloze-Style Task The cloze-style task is to give a sentence with blanks, and the model must find the most appropriate tokens or spans to fill in the blanks. The dataset D = {(xi , c (1) i , ..., c (j) i , ..., yi)} N i=1. For each sample, there is a sentence xi with a [BLANK], and there are Ki candidates {c (j) i } Ki j=1 to be chosen. For each option c (j) i , there is a template p (j) i ∈ Pi corresponding to it. Given the input

完形填空式任务 完形填空式任务是给出一个带有空格的句子，模型必须找到最合适的标记或跨度来填充空格。 数据集 D = {(xi , c (1) i , ..., c (j) i , ..., yi)} N i=1。 对于每个样本，有一个 [BLANK] 的句子 xi，并且有 Ki 候选 {c (j) i } Ki j=1 可供选择。 对于每个选项 c (j) i ，都有一个模板 p (j) i ∈ Pi 与之对应。 鉴于输入



the output of model is:

输出模型是：



As shown in Figure 5, in cloze-style task (such as ChID), if we use the regular position embeddings (hard-position embeddings), the candidate word coin can’t perceive the [BLANK] and the context. Inspired by (Liu et al. 2020) and (Sun et al. 2020), we adopt the soft-position index which allows NSP to work similarly like MLM. We align the position index of the candidate word coin with [BLANK].

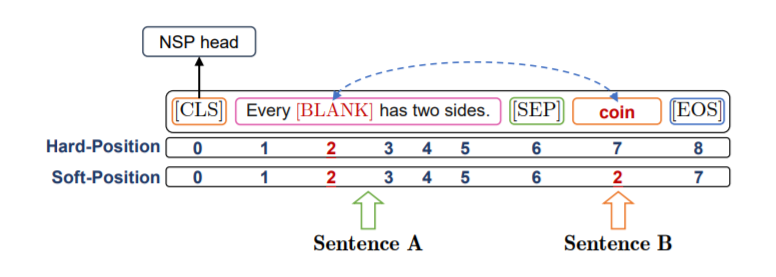
如图 5 所示，在 cloze-style 任务（例如 ChID）中，如果我们使用常规位置嵌入（hard-position embeddings），候选词 coin 无法感知 [BLANK] 和上下文。 受 (Liu et al. 2020) 和 (Sun et al. 2020) 的启发，我们采用了软位置指数，它允许 NSP 像 MLM 一样工作。 我们将候选词coin的位置索引与[BLANK]对齐。

图5：NSP-BERT 中的软定位和硬定位。

Word Sense Disambiguation In a fully supervised training scenario, we may add markers before and after the word to identify the word to be disambiguated (Huang et al. 2019; Soares et al. 2019; Wu and He 2019), as illustrated in Figure 4. Because there is no downstream tasks training data for sentence-level prompt-learning, it is impossible to identify the target word’s position by markers. We propose a TwoStage Prompt construction method to indicate the target word using natural language descriptions in our NSP-BERT.

• Stage 1: Prompt the target word at the end of sentence A. This stage’s purpose is to provide enough context for the target word.

• Stage 2: Prompt the description of the candidate word sense in sentence B.

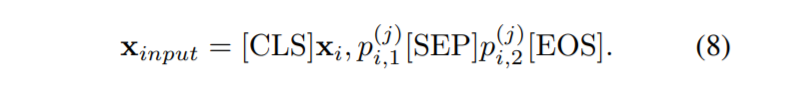
Feed the two-stage prompt into the language model, and it will determine if the sentence is fluent and reasonable. Let p (j) i,1 and p (j) i,2 denote the first and the second part of the prompt. The model’s input is:

词义消歧 在完全监督的训练场景中，我们可以在单词前后添加标记来识别要消歧的单词（Huang et al. 2019; Soares et al. 2019; Wu and He 2019），如图 4 所示 . 因为没有用于句子级提示学习的下游任务训练数据，所以无法通过标记来识别目标词的位置。 我们提出了一种 TwoStage Prompt 构造方法，在我们的 NSP-BERT 中使用自然语言描述来指示目标词。

• 第一阶段：在句子 A 的末尾提示目标词。这个阶段的目的是为目标词提供足够的上下文。

• 阶段 2：提示对句子 B 中候选词义的描述。

将两阶段提示输入语言模型，判断句子是否流畅合理。 让 p (j) i,1 和 p (j) i,2 表示提示的第一部分和第二部分。 模型的输入是：



## 3.3 Answer Mapping

It’s easy to observe that not all probability outputs in the above tasks are directly linked with labels. This is because not all datasets can provide contrastive candidate objections (sentiments/topics/idioms/entities). Pre-trained language models, on the other hand, are not susceptible to negative inference (Kassner and Schutze ¨ 2020), the NSP model is no exception. As a result, we propose two answer mapping methods, candidates-contrast answer mapping and samples-contrast answer mapping, for different situations.

很容易观察到，并非上述任务中的所有概率输出都与标签直接相关。 这是因为并非所有数据集都可以提供对比性的候选反对意见（情绪/主题/习语/实体）。 另一方面，预训练的语言模型不易受到负面推理的影响（Kassner 和 Schutze ¨ 2020），NSP 模型也不例外。 因此，我们针对不同的情况提出了两种答案映射方法，候选-对比答案映射和样本-对比答案映射。

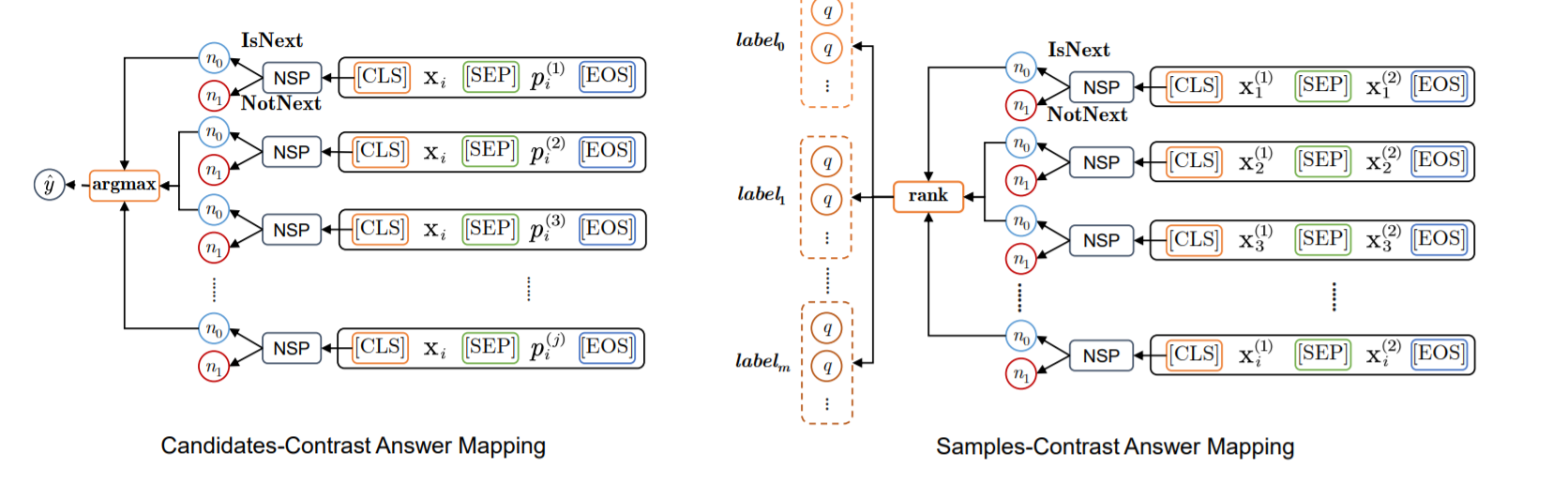
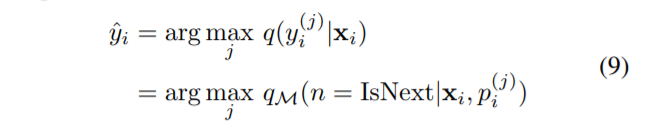


图6：两种答案映射方法候选-对比法（左）和样本-对比法（右）。

Candidates-Contrast For datasets with multiple candidates, such as candidate sentiments (EPRSTMT), candidate topics (TNEWS, IFLYTEK and CSLDCP), candidate idioms (ChID) and candidate entities (DuEL). For the above datasets, there is a template p (j) i (or pi) corresponding to the label y (j) i (or yi). As show in Figure 6. We take the highest probability output by M among the candidates as the final output answer where the condition is IsNext:

Candidates-Contrast 用于具有多个候选的数据集，例如候选情绪（EPRSTMT）、候选主题、候选习语和候选实体。 对于上述数据集，有一个模板p(j)i(或pi)对应标签y(j)i(或yi)。 如图6所示，在条件为IsNext的情况下，我们将候选者中M输出的概率最高的作为最终输出答案：



Samples-Contrast For datasets with no candidate for contrast (such as OCNLI, BUSTM, CSL and CLUEWSC), we propose the samples-contrast answer mapping method (Figure 6), this procedure is summarized in Algorithm 1. Considering the fairness of the comparative experiment, we consider two preconditions. One is that a complete development set and a test set can be obtained at the same time; the other is that only the development set can be obtained, and the test samples must be predicted one by one or batch by batch during testing. In our experiment, we use the development set to determine the thresholds of probability, and use these thresholds to predict the test set.

样本对比 对于没有对比候选的数据集（如OCNLI、BUSTM、CSL和CLUEWSC），我们提出了样本-对比答案映射方法（图6），这个过程总结在算法1中。考虑到对比实验的公平性，我们 考虑两个前提。 一是可以同时获得完整的开发集和测试集； 另一种是只能得到开发集，测试时必须一个一个或一个批次地预测测试样本。 在我们的实验中，我们使用开发集来确定概率的阈值，并使用这些阈值来预测测试集。

# 4 实验

## 4.1 任务和数据集

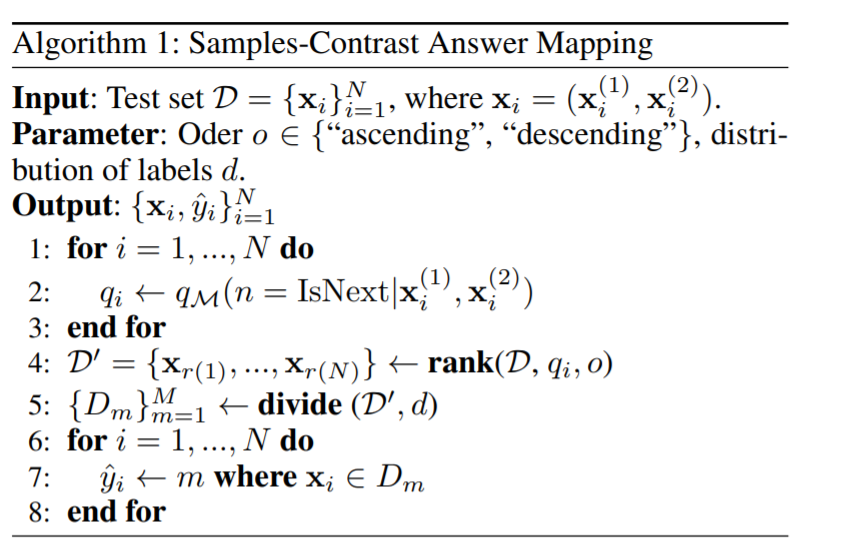
FewCLUE We evaluate our model mainly on FewCLUE (Xu et al. 2021), a Chinese Few-shot Learning Evaluation Benchmark, which contains 9 NLU tasks in Chinese, with 4 single sentence tasks, 3 sentence pair tasks and 2 reading comprehension tasks. The details of the datatsets is shown in Appendix. The number of samples in every training set is few, each label corresponds to 8 or 16 samples. We report the accuracy on all 9 tasks following FewCLUE.

FiveCLUE 我们主要在FewCLUE (Xu et al. 2021) 上评估我们的模型，这是一个中文Few-shot 学习评估基准，其中包含9 个中文NLU 任务，4 个单句任务、3 个句子对任务和2 个阅读理解任务。 数据集的详细信息显示在附录中。 每个训练集中的样本数量很少，每个标签对应8或16个样本。 我们报告了FewCLUE之后所有9个任务的准确性。

DuEL2.0 In order to further verify the ability of NSPBERT for word sense disambiguation, the entity linking dataset DuEL2.0 was added. In particular, we divide

DuEL2.0 into two parts. In the first part, the entity linking part, there are 26586 samples. All the samples’ mention can be mapped to single or multiple entities in the knowledge base, and each mention can be linked to 5.37 entities on average. In the second part, the entity typing part, there are 6465 samples. Those samples’ mention cannot be found in the knowledge base, but they will be divided into their corresponding upper entity types. There are a total of 24 upper entity types.

DuEL2.0 为了进一步验证NSPBERT在词义消歧方面的能力，增加了实体链接数据集DuEL2.0。 特别地，我们将 DuEL2.0 分为两部分。 第一部分，实体链接部分，有26586个样本。 所有样本的mention 可以映射到知识库中的单个或多个实体，每个mention 平均可以链接到5.37 个实体。 第二部分，实体类型部分，有6465个样本。 这些样本的提及在知识库中是找不到的，但是它们会被划分到它们对应的上层实体类型中。 共有 24 种上层实体类型。



## 4.2 Baselines 基线

Refer to the FewCLUE (Xu et al. 2021) 2 , we mainly choose 3 training scenarios, fine-tuning, few-shot and zero-shot.

参考FewCLUE (Xu et al. 2021) 2 ，我们主要选择了3个训练场景，fine-tuning、few-shot和zero-shot。

Fine-Tuning Standard fine-tuning of the pre-trained language model on the FewCLUE training set. The models are fine-tuned with cross entropy loss and using the BERT-style model’s hidden vector of [CLS] h[CLS] with a classification layer softmax(Wh[CLS]), where W ∈ RM×H, M is the number of labels.

Few-Shot In few-shot scenario, we choose token-level model PET (Schick and Schutze ¨ 2021b,a) and its opitmized models ADAPET (Tam et al. 2021), P-tuning (Liu et al. 2021) and LM-BFF(Gao, Fisch, and Chen 2021). We also choose sentence-level model EFL (Wang et al. 2021). All few-shot models are trained on FewCLUE’s training set.

Zero-Shot In zero-shot scenario, there are two ways to realize, one is GPT-ZERO using L2R LM (Radford et al. 2018, 2019; Brown et al. 2020), the other is PET-ZERO using MLM (Schick and Schutze ¨ 2021b,a).

Fine-Tuning 在FewCLUE 训练集上对预训练语言模型进行标准微调。 模型使用交叉熵损失进行微调，并使用 BERT 样式模型的隐藏向量 [CLS] h[CLS] 和分类层 softmax(Wh[CLS])，其中 W ∈ RM×H，M 是数字 的标签。

少镜头在少镜头场景中，我们选择令牌级模型 PET (Schick and Schutze ¨ 2021b,a) 及其优化模型 ADAPET (Tam et al. 2021)、P-tuning (Liu et al. 2021) 和 LM -BFF（高、菲施和陈 2021）。 我们还选择句子级模型 EFL (Wang et al. 2021)。 所有的小样本模型都是在FewCLUE 的训练集上训练的。

Zero-Shot 在零样本场景中，有两种实现方式，一种是使用 L2R LM 的 GPT-ZERO (Radford et al. 2018, 2019; Brown et al. 2020)，另一种是使用 MLM 的 PET-ZERO (Schick 和 Schutze ¨ 2021b,a)。

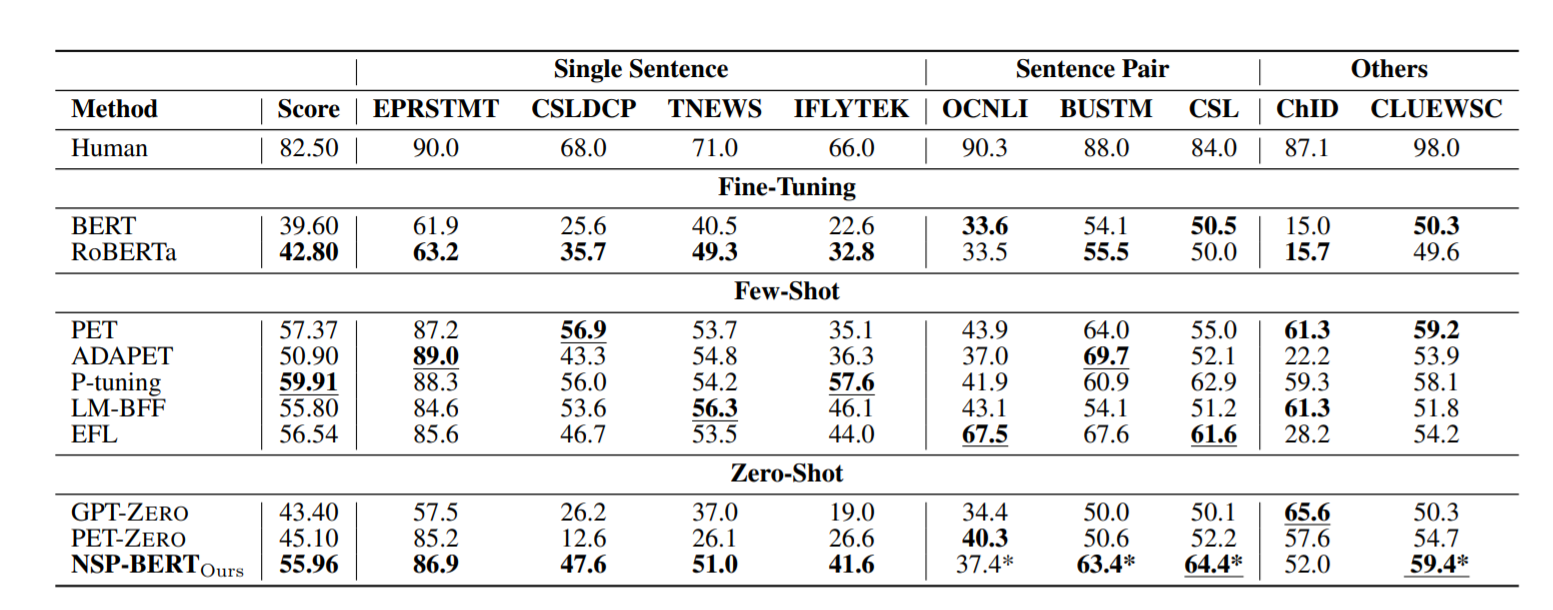


表1：FiveCLUE 基准测试的主要结果。 我们报告了所有 9 个任务的准确度，并计算平均准确度作为很少 CLUE 基准的分数。 星号\*表示使用样本对比答案映射方法。

## 4.3 实验设置

For the baselines, we follow the settings in FewCLUE. The BERT model is using RoBERTa-wwm-ext (Cui et al. 2019, 2020) 3 , a Chinese RoBERTa-BASE with whole-word-mask, which is expected to have better performance on cloze-style tasks. The GPT model is NEZHA-Gen (Wei et al. 2019) 4 . Because of the need to utilize the model pre-trained by the NSP task, none of the RoBERTa models are suitable for our NSP-BERT. So we adopt the vanilla BERT trained by UER using MLM and NSP (Zhao et al. 2019) 5 . The pretraining corpus is a large mixed corpus in Chinese. Along with the basic model (L=12, H=768, A=12, Total Parameters=110M), we conduct experiments using UER-BERTs of various scales (tiny, small, and big) to validate the effect of varied scale models on NSP-BERT. Meanwhile, we use models trained by other organizations (Google6 and HFL3 ), to evaluate the robustness of our optimization methods.

对于基线，我们遵循FewCLUE 中的设置。 BERT 模型使用 RoBERTa-wwm-ext (Cui et al. 2019, 2020) 3 ，这是一个带有全字掩码的中文 RoBERTa-BASE，预计在完形填空式任务上有更好的性能。 GPT 模型是 NEZHA-Gen (Wei et al. 2019) 4 。 由于需要利用 NSP 任务预训练的模型，没有一个 RoBERTa 模型适合我们的 NSP-BERT。 所以我们采用了 UER 使用 MLM 和 NSP 训练的 vanilla BERT (Zhao et al. 2019) 5 。 预训练语料库是一个大型的中文混合语料库。 连同基本模型（L=12，H=768，A=12，总参数=110M），我们使用各种尺度（小、小和大）的UER-BERTs进行实验，以验证各种尺度模型的效果 在 NSP-BERT 上。 同时，我们使用由其他组织（Google6 和 HFL3）训练的模型来评估我们优化方法的稳健性。

## 4.4 主要结果

The table 1 reports the main results on FewCLUE. Our NSP-BERT model outperformed all other zero-shot learning methods on 7 out of 9 datasets. Its performance is comparable to the best few-shot methods currently available.

When using the same size model, it outperforms GPTZERO (based on L2R LM) and PET-ZERO (based on MLM) significantly on the single sentence classification tasks (CSLDCP, TNEWS and IFLTEK). It demonstrates NSP’s remarkable ability to distinguish across sentence topics. Nonetheless, as discussed in the previous section, the sentence-level prompt-learning methods have a number of drawbacks when used with cloze-style tasks, and NSPBERT is no exception. This demonstrates that we have a gap in ChID when compared to token-level methods. The following (§ 4.5) section attempts to narrow the gap using soft-position methods.

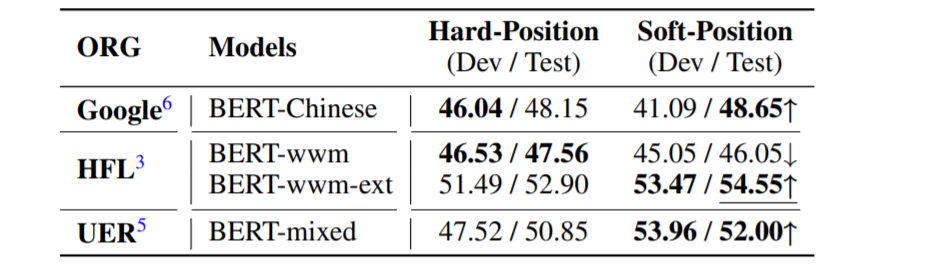
表 1 报告了FewCLUE 的主要结果。 我们的 NSP-BERT 模型在 9 个数据集中的 7 个数据集中优于所有其他零样本学习方法。 它的性能可与目前可用的最好的少拍方法相媲美。

当使用相同大小的模型时，它在单句分类任务（CSLDCP、TNEWS 和 IFLTEK）上明显优于 GPTZERO（基于 L2R LM）和 PET-ZERO（基于 MLM）。 它展示了 NSP 区分句子主题的卓越能力。 尽管如此，如上一节所述，句子级提示学习方法在与完形填空式任务一起使用时有许多缺点，NSPBERT 也不例外。 这表明与令牌级方法相比，我们在 ChID 方面存在差距。 以下（第 4.5 节）部分尝试使用软定位方法缩小差距。

## 4.5 优化和分析

Soft-Position for Cloze-Style Task As seen in Table 2, while the soft-position approach clearly has an effect on ChID, it exhibits some instability. This could be because the model was not fine-tuned sufficiently on the training set.

完形填空式任务的软定位如表 2 所示，虽然软定位方法显然对 ChID 有影响，但它表现出一些不稳定性。 这可能是因为模型没有在训练集上进行足够的微调。

  
表2：与完形填空式任务 ChID 上的硬位置相比，软位置 NSP-BERT 的准确性。 我们采用的模型来自不同的组织。

Two-Stage Prompt In §3.2, we introduced a two-stage prompt method for word sense disambiguation tasks. We compare its effect with a one-stage prompt on dataset DuEL2.0. We divide the dataset into two parts, entity linking and entity typing. In the entity typing part, we do not remove the Other type. Our model has satisfactory performance on DuEL2.0 without relying on any training data, especially for entity linking, NSP-BERT can handle entity descriptions of different lengths well, which is something that models such as PET can hardly achieve.

两阶段提示 在 §3.2 中，我们介绍了一种用于词义消歧任务的两阶段提示方法。 我们将其效果与数据集 DuEL2.0 上的单阶段提示进行比较。 我们将数据集分为两部分，实体链接和实体类型。 在实体类型部分，我们没有删除其他类型。 我们的模型在不依赖任何训练数据的情况下在DuEL2.0上有令人满意的表现，尤其是在实体链接方面，NSP-BERT可以很好地处理不同长度的实体描述，这是PET等模型难以实现的。

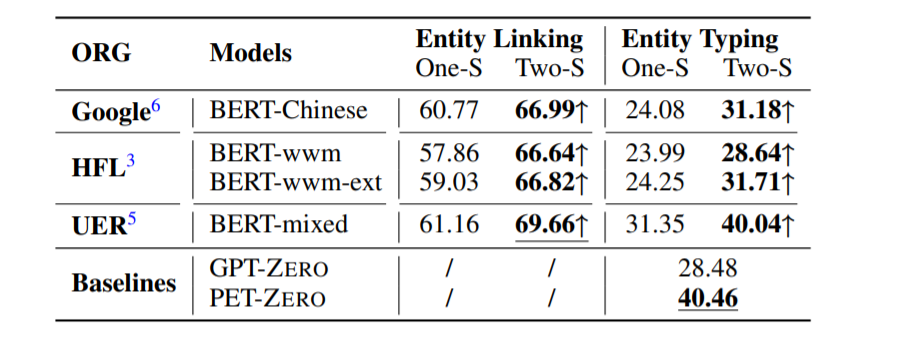
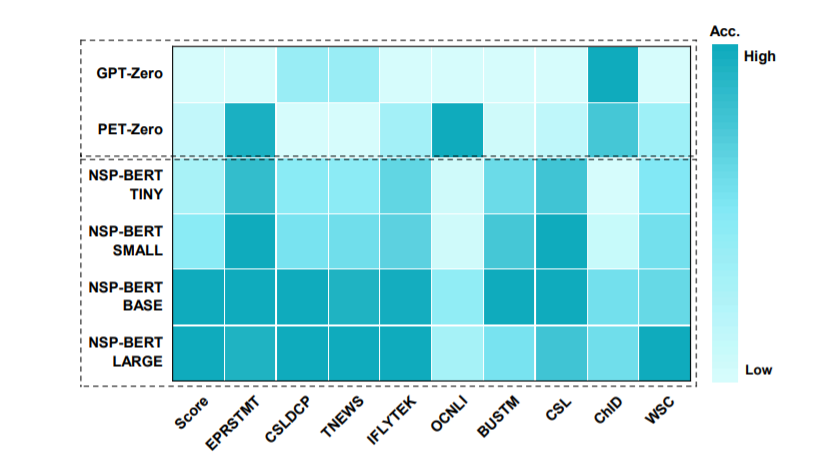


表3：NSP-BERT 在 DuEL2.0 上的结果 (Acc.)，带有一个阶段提示 (One-S) 和两个阶段提示 (Two-S)。 我们不会删除实体类型中的其他类型。

图8：不同比例模型的精度草图。 X 轴代表FewCLUE 中的任务，y 轴代表基线（GPT-ZERO 和PET-ZERO）和不同模型尺度（tiny、small、base 和 large）下的NSPBERT。

Size of Models Shown in Figure 8, we compared the impact of the models’ scale on FewCLUE. The average accuracy of tiny, small, base and large BERT models are 47.35, 49.69, 56.95 and 57.0 respectively, when the baselines GPTZERO and PER-ZERO are 43.40 and 45.10.

Influence of Prompt’s Logic and Fluency The biggest difference between NSP-BERT and contrast learning is that the prompts in NSP-BERT need to be close to natural language habits. As shown in Figure 9, based on the 3 prompt templates, according to the logic, T3 > T2 > T1, the accuracy increased significantly, on 4 datasets (EPRSTMT, TNEWS, CSLDCP and IFLYTEK).

模型的大小 如图 8 所示，我们比较了模型的规模对 FiveCLUE 的影响。 当基线 GPTZERO 和 PER-ZERO 为 43.40 和 45.10 时，tiny、small、base 和 large BERT 模型的平均准确率分别为 47.35、49.69、56.95 和 57.0。

Prompt 逻辑和流畅度的影响 NSP-BERT 和对比学习最大的区别在于 NSP-BERT 中的提示需要接近自然语言习惯。 如图 9 所示，基于 3 个提示模板，按照 T3 > T2 > T1 的逻辑，在 4 个数据集（EPRSTMT、TNEWS、CSLDCP 和 IFLYTEK）上，准确率显着提高。

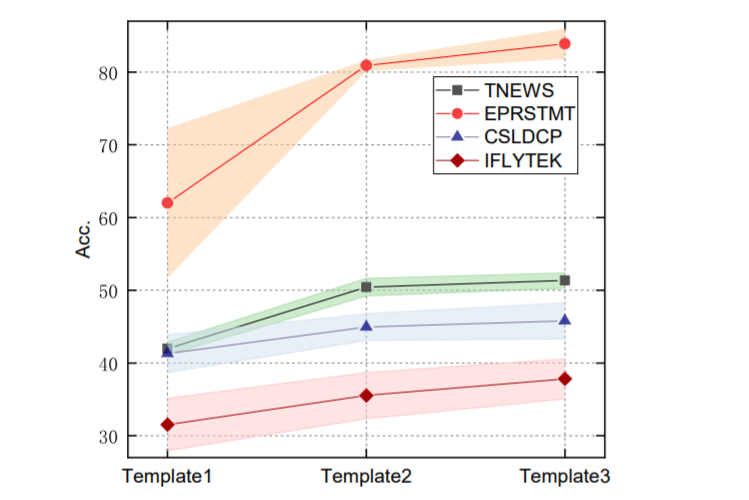


图9：当提示变得更加流畅和合乎逻辑时，NSP-BERT 的准确性就会提高。

# 5 结论

In this paper, we introduce NSP-BERT, which uses an unexpected pre-training task called Next Sentence Prediction (NSP) of BERT to perform various NLP tasks using prompts. As a sentence-level prompt-learning method, NSP-BERT not only can achieve SOTA results on multiple tasks, but it also has an impressive improvement over prior zero-shot methods (GPT and PET). NSP-BERT can accomplish non-fixed length tasks that are difficult to be solved by token-level methods, such as entity linking tasks with variable-length entity descriptions. Although it is based on sentence-level, our method is different from the traditional sentence embedding similarity contrast learning since it must be prompted by natural language. The model’s performance improves as the prompt becomes more logical and fluent. Our NSP-BERT is inspiring for prompt-based learning owing to our experiments show that a simple pre-training task can efficiently solve various downstream tasks without any task-specific training data. In future work, it is essential to extend NSP-BERT to the few-shot scenario.

在本文中，我们介绍了 NSP-BERT，它使用了一种名为 Next Sentence Prediction (NSP) 的 BERT 的意外预训练任务，使用提示来执行各种 NLP 任务。作为一种句子级的提示学习方法，NSP-BERT 不仅可以在多个任务上取得 SOTA 结果，而且比之前的零样本方法（GPT 和 PET）也有令人印象深刻的改进。 NSP-BERT 可以完成令牌级方法难以解决的非固定长度任务，例如具有可变长度实体描述的实体链接任务。虽然它是基于句子级别的，但我们的方法与传统的句子嵌入相似度对比学习不同，因为它必须由自然语言提示。随着提示变得更加合乎逻辑和流畅，模型的性能会提高。我们的 NSP-BERT 对基于提示的学习很有启发，因为我们的实验表明，一个简单的预训练任务可以有效地解决各种下游任务，而无需任何特定于任务的训练数据。在未来的工作中，有必要将 NSP-BERT 扩展到少拍场景。