Learning Disentangled Semantic Representations for Zero-Shot Cross-Lingual Transfer in Multilingual Machine Reading Comprehension

Linjuan Wu[1*], Shaojuan Wu[1*], Xiaowang Zhang[1,2], Deyi Xiong[1†], Shizhan Chen[1†],
Zhiqiang Zhuang[1], Zhiyong Feng[1]

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China, 300350
[2]Tianjin University-Aishu Data Intelligence Joint Laboratory, Tianjin, China
`Tianjin_Research@aishu.cn`
`{wulinjuan1997,shaojuanwu,dyxiong,shizhan}@tju.edu.cn`

Abstract

Multilingual pre-trained models are able to zero-shot transfer knowledge from rich-resource to low-resource languages in machine reading comprehension (MRC). However, inherent linguistic discrepancies in different languages could make answer spans predicted by zero-shot transfer violate syntactic constraints of the target language. In this paper, we propose a novel multilingual MRC framework equipped with a Siamese Semantic Disentanglement Model (S^2DM) to disassociate semantics from syntax in representations learned by multilingual pre-trained models. To explicitly transfer only semantic knowledge to the target language, we propose two groups of losses tailored for semantic and syntactic encoding and disentanglement. Experimental results on three multilingual MRC datasets (i.e., XQuAD, MLQA, and TyDi QA) demonstrate the effectiveness of our proposed approach over models based on mBERT and XLM-100.

1 Introduction

Multilingual pre-trained language models (PLMs) (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020) have been widely explored in cross-lingual understanding tasks. However, zero-shot transfer method based on multilingual PLMs does not work well for low-resource language MRC. Such multilingual MRC models could roughly detect answer spans but may fail to predict the precise boundaries of answers (Yuan et al., 2020).

In order to address this issue, existing methods mainly resort to external resources. Based on the finding that 70% of answer spans are language-specific phrases (e.g., named entities, noun phrases) in MLQA (Lewis et al., 2020), Yuan et al. (2020) propose an additional language-specific knowledge
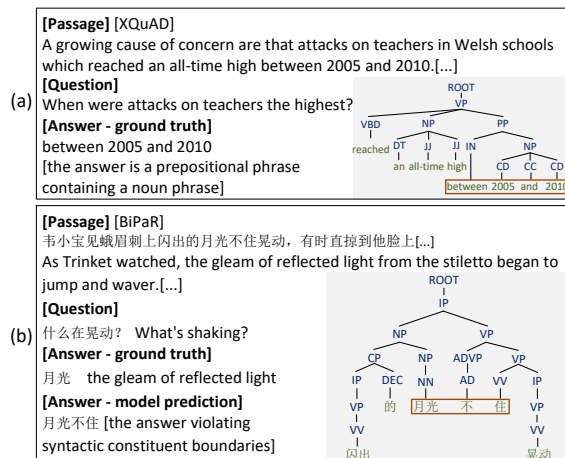

Figure 1: Relations between answer spans and syntactic constituents. (a) An example from XQuAD (Artetxe et al., 2020) where the ground-truth answer is a syntactic constituent. (b) A case from BiPaR (Jing et al., 2019) where the answer predicted by a model transferred from English to Chinese violates syntactic constituent boundaries in the target language.

phrase masking (LAKM) task to enhance boundary detection for low-resource languages. Liang et al. (2021) present a boundary calibration model stacked over a base sequence labeling module, introducing a phrase boundary recovery task to pretrain the calibration module on large-scale multilingual datasets synthesized from Wikipedia documents. These two methods rely on external resources, which are not always easily available.

As illustrated in Figure 1(b), the transfer model may violate syntactic constraints for answer spans in the target language (e.g., the predicted answer "月光不住" crossing the boundaries of two subtrees). An intuitive assumption is that the majority of answer spans respect syntactic constituency boundaries (i.e., syntactic constraint, illustrated by the case in Figure 1(a)). On four multilingual MRC datasets, we use Stanford CoreNLP[1] to collect syn-

---

[1]https://stanfordnlp.github.io/CoreNLP/

991

	XQuAD	MLQA	TyDi QA-GoldP	BiPaR
English	89.08%	90.11%	89.12%	90.99%
Chinese	88.05%	87.57%	-	95.73%

Table 1: The percentages of answer spans that respect syntactic constituent boundaries in four multilingual MRC datasets in both English and Chinese.

tax parse trees and calculate the percentages of ground-truth answers that respect syntactic constituent boundaries. As shown in Table 1, over 87% of answer spans respect the syntactic constraint.

On the bilingual parallel MRC corpus BiPaR (Jing et al., 2019), we have compared two MRC models: a monolingual MRC model trained on the Chinese data of BiPaR vs. an mBERT-based MRC model trained on the English data of BiPaR and adapted to Chinese via zero-shot transfer. For questions where the monolingual model correctly predicts the answer and respect syntactic constraint, 23.15% of them are incorrectly predicted by the transfer model, and the predicted answers violate the syntactic constraint, illustrated by the case in Figure 1(b). This suggests that the source language syntax may have a negative impact on the answer boundary detection in the target language during zero-shot transfer, due to the linguistic discrepancies between the two languages.

However, linguistic discrepancies are diverse and it is difficult to learn them. We hence propose to decouple semantics from syntax in pre-trained models for multilingual MRC, transforming the learning of linguistic discrepancies into universal semantic information. Specifically, we propose a Siamese Semantic Disentanglement Model (S^{2}DM) that utilises two latent variables to learn semantic and syntactic vectors in multilingual pre-trained representations. As shown in Figure 2(a), stacking a linear output layer for MRC over the disentangled semantic representation layer, we can fine-tune the multilingual PLMs on the rich-resource source language and transfer only disentangled semantic knowledge into the target language MRC. Our model aims to reduce the negative impact of the source language syntax on answer boundary detection in the target language.

To disassociate semantic and syntactic information in PLMs well, we introduce objective functions of learning cross-lingual reconstruction and semantic discrimination together with losses of incorporating word order information and syntax structure information (Part-of-Speech tags and syntax parse

trees). We use a publicly available multilingual sentence-level parallel corpus with syntactic labels to train S^{2}DM.

To summarize, our main contributions are as follows.

- We propose a multilingual MRC framework that explicitly transfers semantic knowledge of the source language to the target language to reduce the negative impact of source syntax on answer span detection in the target language MRC.

- We propose a siamese semantic disentanglement model that can effectively separate semantic from syntactic information of multilingual PLMs with semantics/syntax-oriented losses.

- Experimental results on three multilingual MRC datasets (XQuAD, MLQA, and TyDi QA) demonstrate that our model can achieve significant improvements of 3.13 and 2.53 EM points over two strong baselines, respectively.

2 Related Work

Cross-lingual/Multilingual Machine Reading Comprehension Recent advances in multilingual MRC evaluation datasets (Artetxe et al., 2020; Lewis et al., 2020; Clark et al., 2020) trigger research interests in multilingual and cross-lingual MRC (Hsu et al., 2019; Cui et al., 2019; Yuan et al., 2020; Liu et al., 2020; Huang et al., 2021; Wu et al., 2021). Hsu et al. (2019) investigate cross-lingual transfer capability of multilingual BERT (mBERT) on MRC tasks and find that zero-shot learning based on PLM is feasible, even between distant languages, such as English and Chinese. Various approaches have been proposed on top of multilingual MRC based on PLMs. Cui et al. (2019) propose a method that combines multilingual BERT and back-translation for cross-lingual MRC. In order to effectively leverage translation data and reduce the impact of noise in translations, Liu et al. (2020) propose a cross-lingual training approach based on knowledge distillation for multilingual MRC. Yuan et al. (2020) present two auxiliary tasks: mixMRC and LAKM to introduce additional phrase boundary supervision into the fine-tuning stage. Liang et al. (2021) propose a pre-trained boundary calibration module based on the output
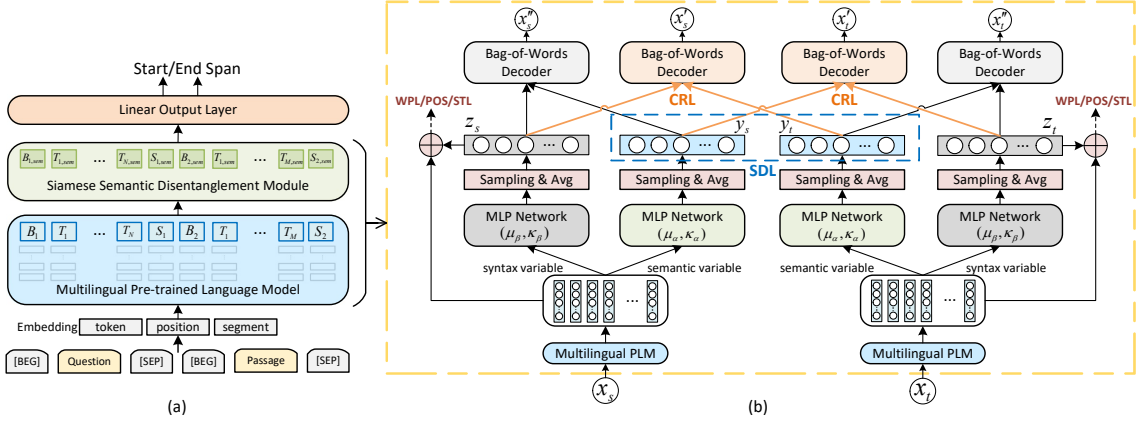
Figure 2: Diagram of the proposed zero-shot cross-lingual transfer framework for multilingual MRC. (a) The overview of our multilingual MRC framework. (b) The architecture of S^{2}DM that is composed of two siamese networks with shared parameters for the source and target language. Once trained, only the output of source language MLP network is fed into the linear output layer. The subscripts s and t in (b) represent the source and target language respectively. CRL: cross-lingual reconstruction loss. SDL: semantic discrimination loss. WPL: word position loss. POS: Part-of-Speech loss. STL: syntax tree loss.

of a base zero-shot transfer model, refining the boundaries of initial answers.

Different from the above studies, we mainly consider the impact of syntactic divergences between the source and target language in zero-shot cross-lingual transfer based on multilingual PLMs, and attempt to disassociate semantics from syntax and only transfer semantics to the target language.

Disentangled Representation Learning Recently, there has been a growing amount of work on learning disentangled latent representations in NLP tasks (Zhang et al., 2019; Hu et al., 2017; Yin et al., 2018). In this aspect, the most related work to our syntax-semantics decoupling method is the vMF-Gaussian Variational Autoencoder (VGVAE) model proposed by Chen et al. (2019). It is a generative model using two latent variables to represent semantics and syntax of the sentence, developed for monolingual setting and trained with paraphrases. It uses paraphrase reconstruction loss and a discriminative paraphrase loss to learn semantic representations and word order information for syntactic representations. We adapt this model to multilingual syntax-semantics disentanglement. We use bilingual sentence pairs to train our model with a cross-lingual reconstruction loss and semantic discrimination loss. To better disentangle semantics from complex and diverse syntax in multilingual PLMs, we introduce two additional syntax-related losses for incorporating POS tags and syntax trees.

3 Approach

Figure 2 shows the architecture of our multilingual MRC framework with the proposed siamese semantic disentanglement model.

3.1 Multilingual MRC Framework

Our multilingual MRC framework consists of three essential components: the multilingual PLM layer, the siamese semantic disentanglement module, and the linear output layer. The output representations from the multilingual PLM are fed into S^{2}DM to disassociate semantic and syntactic information. Only the disentangled semantic representations are input to the linear output layer for predicting answer spans in passages.

In order to facilitate the zero-shot cross-lingual transfer of only semantic knowledge from the rich-resource source language to the low-resource target language, we take a two-stage training strategy. First, we pre-train S^{2}DM with parallel data (see Section 3.2) while the parameters of the multilingual PLM are frozen. Once S^{2}DM is trained, only the output of source language MLP network is fed into the linear output layer for MRC. In the second step, we freeze the parameters of the S^{2}DM and fine-tune the entire multilingual MRC framework on MRC data of the source language.

3.2 Siamese Semantic Disentanglement Model

In S^{2}DM, we assume that a sentence x is generated by a semantic and syntactic variable, i.e., y and z, independently. We follow VGVAE Chen et al.

993

(2019) to use the von Mises-Fisher (vMF) distribution for the semantic variable and the Gaussian distribution for the syntactic variable. Formally, the joint probability of the sentence and its two latent variables can be factorized as:

$$p_\theta(x, y, z) = p_\theta(y)p_\theta(z)p_\theta(x|y, z) \qquad (1)$$

where $p_\theta(x|y, z)$ is a generative model consisting of bag-of-words decoder.

The variational inference process of VG-VAE uses a factorized approximated posterior $q_\phi(y|x)q_\phi(z|x) = q_\phi(y, z|x)$ with the objective function that maximizes a lower bound of the marginal log-likelihood:

$$L_{VGVAE} = L_{RL} + \text{KL}(q_\phi(z|x)||p_\theta(z)) \\ + \text{KL}(q_\phi(y|x)||p_\theta(y)), \qquad (2)$$

$$L_{RL} = \mathbb{E}_{\substack{y \sim q_\phi(y|x) \\ z \sim q_\phi(z|x)}} \big[-\log p_\theta(x|y, z) \big] \qquad (3)$$

where $q_\phi(y|x)$ is subject to $\text{vMF}(\mu_\alpha(x), \kappa_\alpha(x))$ while $q_\phi(z|x)$ follows $N(\mu_\beta(x), diag(\kappa_\beta(x)))$. The prior $p_\theta(y)$ and $p_\theta(z)$ follows the uniform distribution $\text{vMF}(\cdot, 0)$ and a standard Gaussian distribution respectively. Eq.(3) is the **reconstruction loss (RL)** of the generator. In our model, we adopt a multilayer perceptron (MLP) network to learn the mean (μ) and variance (κ) of two distributions. As pre-trained representations are contextually-encoded token vectors, latent variable vectors obtained by sampling from the distributions need to be averaged so as to output sentence-level semantic and syntactic vector.

Since S^{2}DM uses a Siamese network for both the source and target language, the disentanglement between semantics and syntax is conducted for the two languages simultaneously with two parameter-shared subnetworks, as shown in Figure 2(b).

We attempt to extract rich semantic information from multilingual representations which is universal for multiple languages and contains less syntactic information. Except for the conventional reconstruction loss, we propose two additional losses on parallel data to encourage the latent variable y to capture semantic information: a **Cross-lingual Reconstruction Loss (CRL)** and **Semantic Discrimination Loss (SDL)**. The former estimates the cross-entropy loss when we use the semantic representation y_t of the target language to reconstruct the source input and use the source semantic representation y_s for target reconstruction. The latter is used to force the learned source semantic representation y_s to be as close as possible to the target semantic representation y_t since the se-

mantic meanings of the parallel source and target sentence is equivalent to each other. The two losses are estimated as follows:

$$L_{CRL} = \mathbb{E}_{\substack{y_t \sim q_\phi(y|x_t) \\ z_s \sim q_\phi(z|x_s)}} \big[-\log p_\theta(x_s|y_t, z_s) \big] \\ + \mathbb{E}_{\substack{y_s \sim q_\phi(y|x_s) \\ z_t \sim q_\phi(z|x_t)}} \big[-\log p_\theta(x_t|y_s, z_t) \big], \qquad (4)$$

$$L_{SDL} = \max\big\{0, \delta - \text{sim}(y_s, y_t) + \text{sim}(y_s, n_t)\big\} \\ + \max\big\{0, \delta - \text{sim}(y_s, y_t) + \text{sim}(n_s, y_t)\big\} \qquad (5)$$

where $\text{sim}(\cdot, \cdot)$ is a cosine similarity score function. The margin δ is a hyperparameter to control the gap between parallel sentence pair (y_s, y_t) and two non-parallel sentence pairs (y_s, n_t) and (n_s, y_t). n_s is the semantic vector of a negative sample, which has the highest cosine similarity to y_s. Specially, as partial sentences in our corpus are parallel in more than two languages, we limit the data range of negative sampling to only 2-way parallel pairs. n_t are obtained in the similar way to n_s.

In order to guide S^{2}DM to disassociate syntactic information into the syntactic latent variable z, we also define three losses tailored for capturing different types of syntactic information. First, we employ **Word Position Loss (WPL)** , defined as follows:

$$L_{WPL} = \mathbb{E}_{z \sim q_\phi(z|x)} \big[-\sum_i \log \text{softmax}(f(h_i))_i \big], \qquad (6)$$

where $\text{softmax}(\cdot)_i$ indicates the probability of the ith word at position i, and $f(\cdot)$ is a three-layer feedforward neural network with input $h_i = [e_i; z]$ that is the concatenation of the syntactic variable z and the embedding vector e_i of the multilingual PLM for the ith token in the input sentence.

In addition, we define a Part-of-Speech and syntax tree loss to encourage S^{2}DM to isolate deeper syntactic information from pre-trained representations. POS tagging is a sequence labeling task, which can be regarded as a multi-class classification problem for each token in a sentence. Hence, we define **Part-of-Speech (POS) Loss** as a cross-entropy style loss as follows:

$$L_{POS} = \sum_i \big[\sum_{j=1}^m -\log \text{softmax}(g(h_i))_{j=class} \big] \qquad (7)$$

where $g(\cdot)$ is a linear layer, $\text{softmax}(\cdot)_{j=class}$ estimates the probability of gold POS tag $class$, m is the number of different POS tags.

For learning structural information, we design **Syntax Tree Loss (STL)**. Many studies have found

that PLMs can encode syntactic structures of sentences (Hewitt and Manning (2019); Chi et al. (2020)). Inspired by Hewitt and Manning (2019), we formulate syntactic parsing from pre-trained word representations as two independent tasks: depth prediction of a word and distance prediction of two words in the parse tree. Given a matrix $B \in \mathbb{R}^{k \times m}$ as a linear transformation, the losses of these two subtasks are defined as:

$$L_{depth} = \sum_i (\|w_i\| - \|Bh_i\|_2^2), \qquad (8)$$

$$L_{distance} = \sum_{i,j} \left| \mathrm{d}_T(w_i, w_j) - \mathrm{d}_B(h_i, h_j) \right| \quad (9)$$

where $\|w_i\|$ is the parse depth of a word defined as the number of edges from the root of the parse tree to w_i, and $\|Bh_i\|_2$ is the tree depth L2 norm of the vector space under the linear transformation. $\mathrm{d}_T(w_i, w_j)$ is the number of edges in the path between the ith and jth word in the parse tree T. As for $\mathrm{d}_B(h_i, h_j)$, it can be defined as the squared L_2 distance after transformation by B:

$$\mathrm{d}_B(h_i, h_j) = (B(h_i - h_j))^T (B(h_i - h_j)) \quad (10)$$

To induce parse trees, we minimize the summation of the above two losses L_{depth} and $L_{distance}$, and L_{STL} is defined as:

$$L_{STL} = L_{depth} + L_{distance} \qquad (11)$$

According to the different syntactic tasks, we train two S^{2}DM variants: S^{2}DM_POS and S^{2}DM_SP (SP for syntactic parsing), where their training objectives are defined as follows:

$$L_1 = L_{VGVAE} + L_{CRL} + L_{SDL} + L_{WPL} + L_{POS},$$

$$L_2 = L_{VGVAE} + L_{CRL} + L_{SDL} + L_{WPL} + L_{STL}$$

3.3 Generalization Analysis

In this section, we analyze the generalization of our decoupling-based multilingual MRC model.

By two reconstruction losses Eq.(3) and Eq.(4), we will prove that the syntactic and semantic vectors obtained by S^{2}DM are language-agnostic. Since the mathematic structures of Eq.(3) and Eq.(4) are the same, we take one part of Eq.(4) for analysis. Due to z_s and y_t are independent of each other, $p_\theta(x_s, z_s|y_t) = p_\theta(x_s, z_s)$. We obtain:

$$\mathbb{E}_{\substack{y_t \sim q_\phi(y|x_t) \\ z_s \sim q_\phi(z|x_s)}} \left[-\log p_\theta(x_s|y_t, z_s) \right]$$

$$= \mathbb{E}_{y_t \sim q_\phi(y|x_t)} \left(\sum_{z_s \sim q_\phi(z|x_s)} p_\theta(z_s) \log \frac{p_\theta(z_s)}{p_\theta(x_s, z_s|y_t)} \right)$$

$$= \mathrm{KL}(p_\theta(z_s)||p_\theta(x_s, z_s))$$

Similarly,

$$\mathbb{E}_{\substack{y_s \sim q_\phi(y|x_s) \\ z_t \sim q_\phi(z|x_t)}} [-\log p_\theta(x_t|y_s, z_t)] = \mathrm{KL}(p_\theta(z_t)||p_\theta(x_t, z_t))$$

$$L_{RL} = \mathrm{KL}(p_\theta(y_s)||p_\theta(x_s, y_s)) + \mathrm{KL}(p_\theta(y_t)||p_\theta(x_t, y_t))$$

Minimizing $\mathrm{KL}(q_\phi(z|x)||p_\theta(z))$ and $\mathrm{KL}(q_\phi(y|x)||p_\theta(y))$ will eventually fit both $p_\theta(x_s, z_s)$ and $p_\theta(x_t, z_t)$ into the same distribution. In the same way, both $p_\theta(x_s, y_s)$ and $p_\theta(x_t, y_t)$ also fit to the same distribution, no matter what the target language is. This is consistent with our motivation to use the siamese network.

Furthermore, the semantic discrimination loss in Eq.(5) guarantees that the semantic vectors of the source language and the target language are similar to each other. Minimizing Eq.(5) can be equivalent to:

$$\begin{cases} \mathrm{sim}(y_s, y_t) > \mathrm{sim}(y_s, n_t) + \delta \\ \mathrm{sim}(y_s, y_t) > \mathrm{sim}(n_s, y_t) + \delta \end{cases}$$

which is to maximize $\mathrm{sim}(y_s, y_t)$ to encourages the target semantic vector to approach parallel source semantic vector.

In summary, S^{2}DM can obtain language-agnostic semantic and syntactic vectors. Therefore, our multilingual MRC model is suitable even for low-resource languages without training data for the decoupling model.

4 Experiment

4.1 Datasets

To verify the effectiveness of our multilingual MRC model, we conducted experiments on three multilingual question answering benchmarks:

XQuAD (Artetxe et al., 2020) consists of 11 datasets of different languages translated from the SQuAD v1.1 (Rajpurkar et al., 2016) development set, including Spanish (es), German (de), Greek (el), Russian (ru), Turkish (tr), Arabic (ar), Vietnamese (vi), Thai (th), Chinese (zh), Hindi (hi), and Romanian (ro).

MLQA (Lewis et al., 2020) consists of over 5K extractive MRC instances in 7 languages: English (en), Arabic (ar), German (de), Spanish (es), Hindi (hi), Vietnamese (vi) and Chinese (zh). MLQA is also highly parallel, with MRC instances parallel across 4 different languages on average.

TyDi QA-GoldP is the gold passage task in TyDi QA (Clark et al., 2020) covering 9 typologically diverse languages: Arabic (ar), Bengali (bg), English (en), Finnish (fi), Indonesian (id), Korean (ko), Russian (ru), Swahili (sw), Telugu (te). It

is a more challenging MRC benchmark as questions have been written without seeing the answers, leading to 3 and 2 times less lexical overlap than XQuAD and MLQA, respectively (Hu et al., 2020).

4.2 Baseline Models

We used the following two multilingual PLMs to build our MRC model to conduct experiments:

mBERT is the multilingual version of BERT Devlin et al. (2019), with 177M parameters, is pretrained on the Wikipedia of 104 languages to optimize the masked language modeling objective.

XLM-100 uses a pre-training objective similar to that of mBERT but with a larger number of parameters (578M) and a larger shared vocabulary than mBERT, and is trained on the same Wikipedia data covering 100 languages as mBERT.

Furthermore, we compared with a strong baseline that uses external knowledge to enhance cross-lingual MRC:

LAKM is a pre-trained task proposed in (Yuan et al., 2020) by introducing external sources for phrase-level masked language modeling task. The external corpus contain 363.5k passages and 534k knowledge phrases in four languages: English (en), French (fr), German (de), and Spanish (es).

4.3 Setup

For S^2DM, we collected approximately 26k labelled parallel sentence pairs from the Universal Dependencies (UD 2.7) Corpus (Zeman et al., 2020) as the training set. The training set covers 20 languages and overlap with 13 languages of three MRC datasets. We used Universal POS tags and HEAD tags in UD 2.7 for the POS tagging and syntactic parsing task. We chose data from the Chinese semantic textual similarity (STS) task (Tang et al., 2016) as the development set. For hyper-parameters in S^2DM, the learning rate was set to 5e-5, the margin δ was 0.4, and the latent variable dimensions was 200.

For our multilingual MRC models and two baseline models, we fine-tuned them on the SQuAD v1.1 (Rajpurkar et al., 2016) and evaluated them on the test data of the three multilingual MRC datasets. For models based on mBERT, we fine-tuned them for 3 epochs with a training batch size of 32 and a learning rate of 2e-5. We fine-tuned models based on XLM-100 for 2 epochs with a training batch size of 16 and a learning rate of 3e-5.

		XQuAD		MLQA		TyDi QA	
		EM	F1	EM	F1	EM	F1
XLM-100	XLM-100	45.3	70.9	38.5	66.4	33.4	61.7
	XLM+S^2DM_POS	46.6	72.7	40.1	67.3	35.7	63.6
	XLM+S^2DM_SP	**47.7**	**73.5**	**41.4**	**68.9**	**37.5**	**65.5**
mBERT	mBERT	48.5	63.3	41.2	58.5	43.6	57.6
	mBERT+S^2DM_POS	49.4	63.7	42.8	59.9	46.3	58.7
	mBERT+S^2DM_SP	**49.8**	**64.1**	**43.3**	**60.3**	**47.8**	**60.1**

Table 2: The average experimental results on XQuAD, MLQA and TyDi QA dataset.

4.4 Experiment Results

The overall experimental results are shown in Table 2. All our tests were conducted under the conditions of zero-shot transfer. Our models (S^2DM_POS, S^2DM_SP combined with XLM-100 or mBERT) significantly outperform both XLM-100 and mBERT baselines on three datasets. S^2DM_SP achieves the best performance, indicating that the learning of deeper syntax information is compelling. Especially, compared with baselines on the TyDi QA-Gold dataset, S^2DM_SP based on XLM-100 and mBERT gains 4.1%, 4.2% EM improvements on average across 9 languages, respectively.

The results of 12 languages in XQuAD and MLQA are shown in Table 3. For cross-lingual transfer performance, our models are better than the two baselines in terms of either EM or F1 on all 11 low-resource target languages. On the MLQA dataset, LAKM uses a larger extra corpus to train a better backbone language model, while our method with less external data can still achieve similar performance in German (de) and Spanish (es).

The TyDi QA-GoldP dataset is more challenging than XQuAD and MLQA. The results of TyDi QA-GoldP are shown in Table 4, and our models are superior to the baselines in terms of either EM or F1 for all 8 low-resource target languages. Significantly, XLM+S^2DM_SP outperforms the XLM-100 baselines by 8.4%, 9.5% in EM for Finnish (fi), Russian(ru), respectively. The language families of these two languages are different from that of English. The evaluation results on these three datasets verify the effectiveness of our proposed method.

In Section 3.3, we theoretically analyze the generalization of our model. The results on the three datasets show the effectiveness on five languages not included in the training target languages for S^2DM. The five languages are Romanian (ro), Vietnamese (vi) in XQuAD and Bengali (bg), Swahili (sw), Telugu (te) in TyDi QA-GoldP, which are resource-scarce and have differ-

		XQuAD (EM/F1)											
		en	ar	de	el	es	hi	ro	ru	th	tr	vi	zh	avg
XLM-100	XLM-100	66.5/86.5	35.6/72.4	53.8/80.9	37.9/66.3	54.6/81.0	39.9/64.9	56.6/79.6	54.0/**79.5**	10.3/27.0	42.0/72.4	49.5/75.4	42.7/65.4	45.3/70.9
	XLM+S²DM_POS	67.5/87.4	**40.2**/**74.9**	54.2/80.8	41.9/71.3	55.4/82.1	40.0/66.2	56.4/79.6	54.0/79.3	**13.8**/**38.9**	41.9/70.8	50.6/75.8	42.9/65.1	46.6/72.7
	XLM+S²DM_SP	68.3/88.0	39.8/**74.9**	55.8/81.7	44.1/72.4	56.8/82.5	40.5/66.5	59.0/81.7	54.2/79.5	13.3/38.5	44.5/72.9	51.3/76.1	44.5/67.6	47.7/73.5
mBERT	mBERT	72.6/83.6	44.3/**60.6**	54.0/69.6	46.0/61.1	57.3/74.9	38.3/53.3	58.3/72.5	54.0/69.6	30.9/39.9	33.8/50.9	46.1/65.9	46.3/57.4	48.5/63.3
	mBERT +S²DM_POS	73.4/83.2	**44.9**/59.9	55.6/**71.9**	44.8/59.7	**57.4**/**75.0**	41.3/55.7	58.1/72.4	55.3/**71.2**	**32.7**/**40.7**	34.0/50.8	48.2/67.4	47.1/56.9	49.4/63.7
	mBERT +S²DM_SP	73.2/**84.0**	43.3/60.0	55.2/70.7	**46.6**/**61.8**	57.1/74.1	42.7/56.5	59.5/73.4	54.6/70.3	30.4/38.9	36.3/51.4	**49.8**/**69.7**	48.9/58.5	**49.8**/**64.1**
		MLQA (EM/F1)											
XLM-100	XLM-100	59.1/81.8	27.0/62.8	43.5/71.3	-	42.7/73.8	29.3/56.4	-	-	-	-	37.4/65.0	30.1/53.7	38.5/66.4
	XLM+S²DM_POS	61.1/82.8	30.5/65.7	43.9/71.2	-	43.1/73.5	31.5/58.0	-	-	-	-	39.7/66.7	30.5/53.1	40.1/67.3
	XLM+S²DM_SP	61.1/83.0	31.2/67.1	45.9/72.9	-	43.6/74.1	34.1/61.2	-	-	-	-	41.4/68.5	32.3/55.6	41.4/68.9
mBERT	mBERT	67.0/79.3	31.5/49.5	43.8/58.3	-	45.8/64.1	29.4/45.2	-	-	-	-	37.5/57.3	34.5/56.1	41.2/58.5
	LAKM	66.8/**80.0**	-	**45.5**/**60.5**	-	**48.0**/**65.9**	-	-	-	-	-	-	-	-
	mBERT+S²DM_POS	66.3/79.5	**32.4**/50.2	45.1/59.7	-	46.8/65.1	30.8/46.0	-	-	-	-	39.5/59.4	**38.4**/59.1	42.8/59.9
	mBERT+S²DM_SP	67.5/79.8	32.1/**50.5**	45.3/59.9	-	47.2/65.0	**32.0**/46.9	-	-	-	-	**41.1**/**60.6**	38.0/**59.3**	**43.3**/**60.3**

Table 3: EM and F1 score of 12 languages on the XQuAD and MLQA dataset.

		TyDi QA-GoldP (EM/F1)									
		en	ar	bg	fi	id	ko	ru	sw	te	avg
XLM-100	XLM-100	52.9/78.1	31.1/69.8	29.2/57.7	39.3/65.3	42.8/69.0	1.4/24.9	36.8/70.2	32.9/59.2	34.4/61.1	33.4/61.7
	XLM_S²DM_POS	52.3/76.1	30.4/69.5	**37.2**/66.1	37.5/64.6	44.1/68.4	**1.8**/25.3	39.4/72.4	**41.9**/62.7	37.1/**67.4**	35.7/63.6
	XLM_S²DM_SP	**53.6**/78.5	34.4/72.3	33.6/**66.8**	47.7/72.7	45.5/69.4	1.5/28.8	46.3/75.6	37.5/**63.1**	**37.2**/62.6	37.5/65.5
mBERT	mBERT	65.5/75.3	43.8/59.5	39.8/54.9	44.0/**56.9**	45.3/59.8	41.7/49.8	41.4/64.4	32.3/50.0	39.0/48.2	43.6/57.6
	mBERT+S²DM_POS	**66.1**/74.8	44.2/**60.9**	41.6/53.3	41.9/55.6	46.5/60.2	45.3/51.7	42.9/63.6	43.3/55.8	**44.8**/**52.7**	46.3/58.7
	mBERT+S²DM_SP	65.9/**76.6**	**44.7**/60.7	**44.2**/55.2	45.1/56.5	47.3/60.9	48.2/55.0	44.3/65.5	45.9/58.1	44.4/52.0	**47.8**/**60.1**

Table 4: EM and F1 score of 9 languages on the TyDi QA-GoldP dataset.

ent language families from English. Significantly, mBERT+S²DM_SP outperforms the mBERT baseline by 13.6% in EM for Swahili (sw).

5 Analysis

5.1 Ablation Study

We further conducted an ablation study based on the mBERT and VGVAE model with different combinations of losses (introduced in the Section 3.2). The results are shown in Figure 3. Our mBERT+S²DM_SP MRC model achieves the strongest performance among all variants, surpassing the model w/ all losses. According to the results shown in Figure 3, we can summarize that each loss is essential and suitable to our model.

The results without POS and STL loss (e.g., w/ CRL+SDL+WPL) on the MLQA dataset validate the effectiveness of our losses (POS or STL loss) tailored for capturing syntactic information. The performance of models that only contain two losses in CRL, SDL, and WPL drops significantly compared with the w/ CRL+SDL+WPL model. The results of models that only contain one of the losses in CRL, SDL drop slightly, but the EM of the model with only WPL is better than w/ CRL+WPL and w/ SDL+WPL, which further demonstrates the importance of the syntax-oriented loss. All ablation models do not exceed our best model, illustrating the importance of all proposed losses.
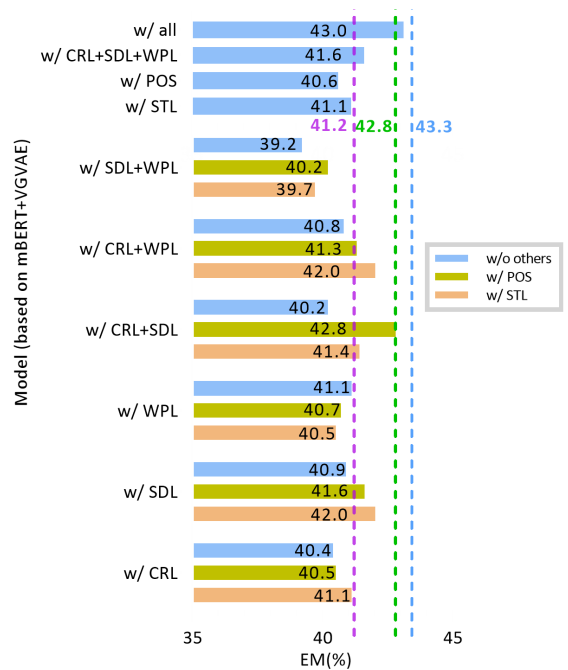

Figure 3: The ablation study results on the MLQA dataset. The three dotted lines indicate the results of baseline mBERT, S²DM_POS, and S²DM_SP from left to right, respectively.
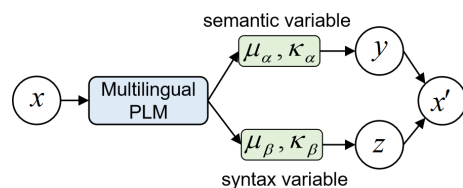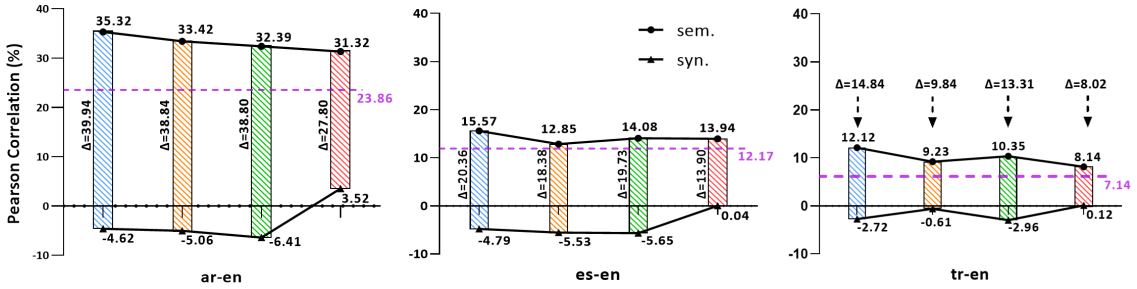

Figure 4: A single network of S²DM.

Figure 5: Pearson correlation (%) on the cross-lingual STS tasks. The length of the bar represents the gap of two vectors. The four bars with different colors represent the results of S^2DM_SP, $S^2DM\_single\_SP$, S^2DM_POS, and $S^2DM\_single\_POS$ from left to right. Purple dotted line: the result of mBERT.
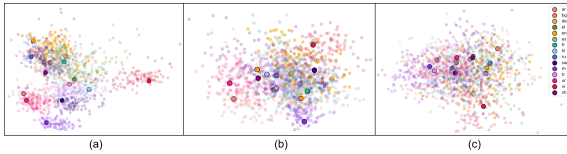

Figure 6: PCA visualization of hidden representations from the last layer of mBERT (a) and semantic representations of S^2DM_POS (b) and S^2DM_SP (c). Darker dots: the same 15-way parallel sentence in different languages.

	XQuAD		MLQA		TyDi QA	
	EM	F1	EM	F1	EM	F1
S^2DM_POS	49.4	63.7	42.8	59.9	46.3	58.7
$S^2DM\_single\_POS$	48.6	62.6	42.4	59.2	43.9	56.4
S^2DM_SP	**49.8**	**64.1**	**43.3**	**60.3**	**47.8**	**60.1**
$S^2DM\_single\_SP$	49.2	63.6	42.5	59.6	45.3	58.4

Table 5: Performance of our multilingual MRC model with different S^2DM variants based on mBERT.

5.2 Why Use a Siamese Network in S^2DM?

In order to separate semantic information from PLMs, an alternative way is to train a single network based on the VGVAE model as shown in Figure 4. Compared with S^2DM, the single-network model does not use the CRL and SDL loss and only requires labeled monolingual data. Corresponding to S^2DM, there are also two single-network variants: $S^2DM\_single\_POS$ and $S^2DM\_single\_SP$. Since there is no explicit semantics learning across the source and target language, we conjecture that the single-network S^2DM will affect the quality of learned semantic vectors and the degree of semantics-syntax decoupling. As shown in Table 5, the performance of the single-network S^2DM is worse than the siamese-network model.

5.3 Why the S^2DM Works?

Our method mainly aims to reduce the potential negative impact of syntactic differences of lan-

guages in the zero-shot transfer process by explicitly isolating semantics from syntax in representations from multilingual pre-trained models. Therefore, we hope to obtain multilingual semantic representations with rich semantic information to guide the machine to read and understand texts. In order to examine (1) whether semantic vectors y in S^2DM encode rich semantic information, and (2) whether semantics is sufficiently separated from syntax, and (3) whether semantic disentanglement can improve predicted answer spans in matching syntactic structures of the target language, we conducted additional experiments and analyses.

Here we used three datasets of cross-lingual semantic textual similarity (STS) in SemEval-2017[2] to evaluate the quality of semantic vectors learned by S^2DM. The three datasets are for Arabic to English (ar-en), Spanish to English (es-en), and Turkish to English (tr-en) cross-lingual STS. We report the results of our models in Figure 5 based on mBERT. We also evaluated learned syntactic vectors in cross-lingual STS, hoping that the performance gap between semantic vectors (i.e., y in S^2DM) and syntactic vectors (i.e., z in S^2DM) is as large as possible. As shown in Figure 5, disentangled semantic representations significantly improve Pearson correlation over the baseline in ar-en, es-en, and tr-en by 11.46%, 3.40%, 4.98%, respectively. Additionally, disentangled syntactic representations are negatively correlated to STS in most cases. These results suggest that disentangled semantic vectors indeed learn rich universal semantic information.

We visualize hidden representations of the last layer of mBERT and semantic representations of mBERT+S^2DM_POS and mBERT+S^2DM_SP in Figure 6, in which the parallel sentences are from

---

[2]https://alt.qcri.org/semeval2017/task1/

a 15-way parallel corpus (Conneau et al., 2018). It is clear to see that disentangled semantic representations learned by S^2DM make parallel sentences in 15 languages (semantically equivalent to each other) closer to one another in space, blending language boundaries clearly seen from mBERT representations (Figure 6(a)). Combined with the negative/positive results of syntactic/semantic vectors in the cross-lingual STS task in SemEval-2017, the visualization demonstrates that S^2DM can efficiently disassociate semantics from syntax.

Finally, we evaluated the degree of consistency to syntactic constituents of predicted answer spans. As described in Section 1, 23.15% of the non-transfer predicted correct answers violate syntactic constraint of the target language during the raw zero-shot cross-lingual transfer on BiPaR. By contrast, S^2DM_POS and S^2DM_SP drop this percentage to 12.98% and 6.60%, respectively. Moreover, on the entire test set of BiPaR (Jing et al., 2019) in Chinese, 93.27% answers predicted by S^2DM_SP exactly span syntactic constituents, which is 8.14% higher than the mBERT model.

6 Conclusions

In this paper, we have presented a novel multilingual MRC model for zero-shot cross-lingual transfer, which can disentangle semantic from syntactic representations and explicitly transfer semantic information from rich-resource language to low-resource languages, reducing the influence of syntactic differences between languages on the answer span prediction of the target language. To disassociate semantics from syntax in multilingual pre-trained representations, we propose the siamese semantic disentanglement model that semantics/syntax-oriented losses to guide latent variables to learn corresponding information. For low-resource languages without training data for the decoupling model, our theoretical analysis and experiments verify the generalization of our multilingual MRC model. Further in-depth analyses suggest that the proposed S^2DM can efficiently disentangle semantics from syntax and significantly improve syntactic consistency of answer predictions on the target language after zero-shot cross-lingual transfer.

Acknowledgments

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 4623–4637.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. A multi-task approach for disentangling syntax and semantics in sentence representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 2453–2464.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 5564–5577.

Jonathan H. Clark, Jennimaria Palomaki, Vitaly Nikolaev, Eunsol Choi, Dan Garrette, Michael Collins, and Tom Kwiatkowski. 2020. Tydi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Trans. Assoc. Comput. Linguistics*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of the 2019 Annual Conference on Neural Information Processing Systems, NeurIPS 2019*, pages 7057–7067.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 2475–2485.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2019. Cross-lingual machine reading comprehension. In *Proceedings of*

the *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 1586–1595.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 4171–4186.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 4129–4138.

Tsung-Yuan Hsu, Chi-Liang Liu, and Hung-yi Lee. 2019. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 5932–5939.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, pages 4411–4421.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, pages 1587–1596.

Wei-Cheng Huang, Chien-yu Huang, and Hung-yi Lee. 2021. Improving cross-lingual reading comprehension with self-training. *arXiv preprint arXiv:2105.03627*.

Yimin Jing, Deyi Xiong, and Yan Zhen. 2019. Bi-PaR: A bilingual parallel dataset for multilingual and cross-lingual reading comprehension on novels. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 2452–2462.

Patrick S. H. Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 7315–7330.

Shining Liang, Linjun Shou, Jian Pei, Ming Gong, Wanli Zuo, and Daxin Jiang. 2021. Calibrenet: Calibration networks for multilingual sequence labeling. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*, pages 842–850.

Junhao Liu, Linjun Shou, Jian Pei, Ming Gong, Min Yang, and Daxin Jiang. 2020. Cross-lingual machine reading comprehension with language branch knowledge distillation. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020*, pages 2710–2721.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, pages 2383–2392.

Shancheng Tang, Yunyue Bai, and Fuyu Ma. 2016. Chinese semantic text similarity trainning dataset. Xi'an University of Science and Technology.

Gaochen Wu, Bin Xu, Yuxin Qin, Fei Kong, Bangchang Liu, Hongwen Zhao, and Dejie Chang. 2021. Improving low-resource reading comprehension via cross-lingual transposition rethinking. *arXiv preprint arXiv:2107.05002*.

Pengcheng Yin, Chunting Zhou, Junxian He, and Graham Neubig. 2018. StructVAE: Tree-structured latent variable models for semi-supervised semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, pages 754–765.

Fei Yuan, Linjun Shou, Xuanyu Bai, Ming Gong, Yaobo Liang, Nan Duan, Yan Fu, and Daxin Jiang. 2020. Enhancing answer boundary detection for multilingual machine reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 925–934.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, et al. 2020. Universal dependencies 2.7. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Xinyuan Zhang, Yi Yang, Siyang Yuan, Dinghan Shen, and Lawrence Carin. 2019. Syntax-infused variational autoencoder for text generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 2069–2078.