Adversarial Learning for Multi-Lingual Entity Linking

Bingbing Wang[1], **Bin Liang**[2*], **Zhixin Bai**[3], **Yongzhuo Ma**[1]
[1] Harbin Institute of Technology, Shenzhen, China
[2] The Chinese University of Hong Kong, Hong Kong, China
[3] Harbin Institute of Technology, Harbin, China
{bingbing.wang,baizhixin,yanagichiaki}@stu.hit.edu.cn,
bin.liang.cuhk.edu.hk

Abstract

Entity linking aims to identify mentions from the text and link them to a knowledge base. Further, Multi-lingual Entity Linking (MEL) is a more challenging task, where the language-specific mentions need to be linked to a multi-lingual knowledge base. To tackle the MEL task, we propose a novel model that employs the merit of adversarial learning and few-shot learning to generalize the learning ability across languages. Specifically, we first randomly select a fraction of language-agnostic unlabeled data as the language signal to construct the language discriminator. Based on it, we devise a simple and effective adversarial learning framework with two characteristic branches, including an entity classifier and a language discriminator with adversarial training. Experimental results on two benchmark datasets indicate the excellent performance in few-shot learning and the effectiveness of the proposed adversarial learning framework.

1 Introduction

Entity linking (EL), a process of disambiguating entity mentions with a target knowledge base (KB), is one of the tasks in information retrieval (Joko et al., 2021) and real applications involving information extraction (Phan and Sun, 2018) and question answering (Li et al., 2020), etc. Many state-of-the-art studies generally pay attention to English KB and do not put enough energy into the low-resource and challenging languages, such as Persian. In addition, the vast majority of low-resource languages are only provided with a limited annotated text, even without labeled data. Therefore, the cross-lingual entity linking (XEL) task was proposed for several pairs of source text and KB languages (McNamee et al., 2011; Tsai and Roth, 2016; Sil et al., 2018; Upadhyay et al., 2018a), where mentions expressed in a language are linked to a KB delivered in another.

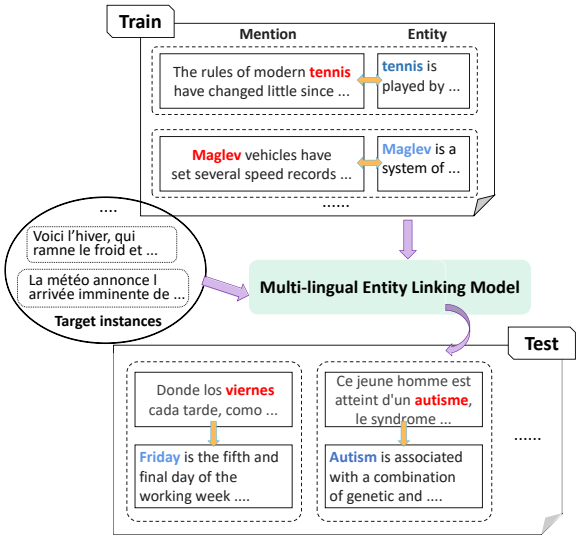

Figure 1: Multi-lingual entity linking task: training and test example of the source text in English and target text in French.

However, XEL restricted the scope of EL to some extent since this popular method generally relies on the hypothesis of one single pivotal KB language as well as one limited KB. Subsequently, Multi-lingual entity linking (MEL) has gained attention as the generalization of XEL and some datasets (Joko et al., 2021; Botha et al., 2020b; Ji et al., 2015) have been collected for it. Compared to TAC-KBP 2014, TAC-KBP 2015 (Ji et al., 2015) was broadened from monolingual to trilingual coverage in three languages. Recently, Mewsli-9 (Botha et al., 2020b) was introduced as a large dataset featuring all entities to numerous cross-lingual systems with almost 300,000 mentions through 9 kinds of languages.

Given a text and entity mentions, there are two primary steps for multi-lingual entity linking: (1) Candidate Generation, possible entities are engendered for the mention, and (2) Entity Ranking, a score between the representation of mention and a candidate entity is computed. In this work, we con-

*Corresponding Authors

Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10), pages 28–35

sider the multi-lingual entity linking task illustrated in Figure 1 that mainly takes Entity Ranking into consideration, and adopts a language-adversarial training approach to improve the performance.

First, a multi-lingual pre-trained transformer model XLM-Roberta (XLM-R) (Conneau et al., 2020) which builds robust representations of text in a wide range of languages, is utilized to build a single representation of mention including surrounding context and name of mention, and entity with description. The abundant source languages are leveraged to compute the similarity between mention and entity.

Second, we design a dedicated and simple adversarial learning approach to construct a language discriminator, which cleverly selects a small part from the test data (excluded during testing) and effectively generalizes to unseen languages for better robustness. In addition, previous studies (Arjovsky et al., 2017) contended that an adversarial training network could be regarded as a way that minimizes the Jensen-Shannon (JS) divergence between two distributions, in our case the feature distributions of the source language and target language. For the discontinuities of JS divergence, Wasserstein distance was proposed to enhance the stability of hyperparameter selection. Furthermore, a gradient penalty is introduced in our adversarial training approach to optimize the discriminator loss that hopes to enlarge the difference between source and target language as much as possible.

The main contributions of our work are summarized as follows:

- A novel adversarial learning framework for the multi-lingual entity linking task in few-shot learning is proposed with the purpose of English bias reduction and generalization improvement.
- We introduce a simple but effective adversarial training approach that randomly selects a certain proportion of test data, and optimizes the feature distributions between source and target languages by minimizing the Wasserstein distance with an additional gradient penalty.
- State-of-the-art results of the experiment on few-shot learning reveal the robustness of our model in the multi-lingual entity linking task.

2 Related Work

2.1 Entity Linking

A series of previous works paid attention to entity linking which develops a model to link textual mentions to entities in KB. (De Cao et al., 2020) proposed a system that retrieves entities by generating their unique names in an autoregressive manner, processing each token sequentially from left to right while conditioning on the given context. (Liu et al., 2022) introduced a scalable and effective BERT-based entity linking model that balances accuracy and speed. Their two-stage zero-shot linking algorithm defines each entity with only a short textual description, and they provide an extensive evaluation of the model's performance. (Botha et al., 2020a) developed a dual encoder model that significantly enhances feature representation, incorporates negative mining, and includes an auxiliary entity-pairing task. This approach resulted in a single-entity retrieval model capable of handling over 100 languages and 20 million entities.

2.2 Multi-lingual Entity Linking

Building on this foundation, researchers gradually shifted their focus to Cross-Language Entity Linking (XEL). (Upadhyay et al., 2018b) devised the first XEL approach that integrates supervision from multiple languages. This method enhances the limited supervision in the target language with additional supervision from a high-resource language, allowing for the training of a single entity linking model across multiple languages. (Zhou et al., 2019a) examined the impact of resource availability on the quality of existing XEL systems and quantified this effect. They proposed three improvements to entity candidate generation and disambiguation, which optimize the use of limited data in resource-scarce scenarios. (De Cao et al., 2022) designed a sequence-to-sequence approach for multilingual entity linking that enhances the interaction between mention strings and entity names. This method cross-encodes mentions and entity names, capturing more complex interactions than the traditional dot product between mention and entity vectors.

3 Methodology

3.1 Task Definition and Overview

Multi-lingual entity linking is a task that links an entity mention in some context languages to the corresponding entity in a language-agnostic KB.

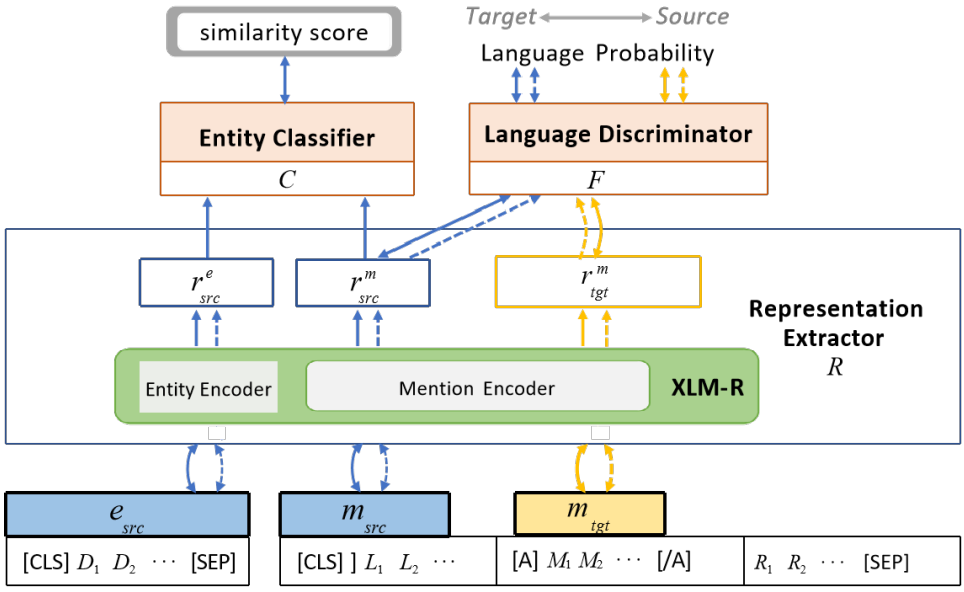

Figure 2: Proposed adversarial learning framework. Blue lines show the flow of source texts and the yellow ones are of target texts. The parameters of R and C are updated and shown as solid lines. The parameters of F are updated and demonstrated as dotted lines.

On this foundation, we employ a few-shot multilingual entity linking task aiming at reducing English bias in EL and improving the generalization for unseen entity set in KBs.

As illustrated in Figure 2, there are three primary components: Representation Extractor R that attains feature representations, Entity Classifier C that aims to compute similarity scores of entity-mention pairs, and Language Discriminator F that identifies whether the input text is from source or target language. Going forward, we assume that if the well-trained language discriminator F can't distinguish the language of the given representation extracted by pre-trained transformer model, these representations can be regarded as language-invariant. That's the motivation we introduce adversarial F to achieve better performance of representation extraction and effectiveness of language invariance.

A representation extractor is designed for the labeled source text T_{src} and unlabeled target text T_{tgt} given as input data. We then conduct a two-step training procedure in each training iteration. First, a small amount of unlabeled source (blue lines) and unlabeled target data (yellow lines) treated by representation extractor R, pass through a language discriminator F for adversarial training. And the labeled source data are put into C to calculate the similarity of mention and entity.

3.2 Representations Extractor

To extract the representation of mention and entity respectively, XLM-Roberta (XLM-R) (Conneau et al., 2020), a transformer representation model that is well-performed in the multi-lingual task is applied as the encoder to represent text into hidden representations. Mention-entity pair in source language is defined as $(m_{src}, e_{src}) \in T_{src}$, while $m_{tgt} \in T_{tgt}$ in target language. m_{src} and m_{tgt} are the combination of local context (the mention span M_i separated by [A] and [/A] markers , left of the mention L_i, right of the mention R_i). The source entity e_{src} is simply the entity description.

Mention in source text m_{src} is fed into XLM-R, and we use max pooling to create a single representation r^m_{src}. A similar method is used for entity in source text and mention in target text to obtain representation r^e_{src} and r^m_{tgt} respectively. Furthermore, r^m_{src} and r^e_{src} are then fed into entity classifier C to produce a score using cosine similarity shown in Eq. 1, while the language discriminator F is exposed to both r^m_{src} and r^m_{tgt}.

$$s(m_{src}, e_i) = \cos(r^m_{src}, r^e_i) = \frac{r^m_{src} \cdot r^e_i}{\|r^m_{src}\| \cdot \|r^e_i\|} \quad (1)$$

where the mention representation r^m_{src} is compared with candidate entity representation $r^e_i (i = 1, 2, ..., k)$ in source text.

3.3 Adversarial Training

In order to aid the training model to learn representations preferably fitted for transferring to unseen languages, we further investigate a simple but effective adversarial training approach, which randomly selects test data (excluded during testing) as target instances according to a proportion of 1%, 5%, 10%. And the distribution of the representation extractor for both source and target instances are defined as below:

$$Y_R^{src} \triangleq Y(r_{src}^m = R(x)|x \in m_{src})$$

$$Y_R^{tgt} \triangleq Y(r_{tgt}^m = R(x)|x \in m_{tgt})$$

Our goal is to make these two distributions as close as possible to get better multi-lingual generalization. Traditional adversarial approaches suffer from convergence and unstable min-max game originating from the discontinuous JS divergence. To settle down this problem, Wasserstein Generative Adversarial Networks (WGAN) (Arjovsky et al., 2017) using Wasserstein distance is proposed. Enlightened by this, we minimize the Wasserstein distance W between Y_R^{src} and Y_R^{tgt} based on Kantorovich-Rubinstein duality (Villani, 2009).

$$W(Y_F^{src}, Y_F^{tgt}) = \sup_{||t||_L \leq 1} \mathop{\mathbb{E}}_{r_{src}^m \sim Y_R^{src}}[t(r_{src}^m)] - \mathop{\mathbb{E}}_{r_{tgt}^m \sim Y_R^{tgt}}[t(r_{tgt}^m)] \quad (2)$$

where the supremum is over all the set of 1-Lipschitz functions t. For convenience, we instead the function as the language discriminator F. The adversarial loss is given as:

$$L_{adv} = \mathop{\mathbb{E}}_{r_{tgt}^m \sim Y_R^{tgt}}[F(r_{tgt}^m)] - \mathop{\mathbb{E}}_{r_{src}^m \sim Y_R^{src}}[F(r_{src}^m)] + \lambda_p L_p \quad (3)$$

where λ_p is the gradient penalty coefficient. The intuition is that F should output the scores of the source language much higher than the target one. Moreover, WGAN also proposes weight clipping to meet the requirement that the discriminator must lie within the space of 1-Lipschitz functions. Unfortunately, it's exactly what leads to optimization difficulties of gradient vanishing and explosion. A gradient penalty is hence introduced to the optimization function and constrains the output relative to the gradient norm of the input:

$$L_p = \mathop{\mathbb{E}}_{\tilde{r} \sim Y_R'}[(||\nabla_{\tilde{r}} F(\tilde{r})||_2 - 1)^2] \quad (4)$$

$$\tilde{r} = \mu r_{src}^m + (1-\mu) r_{tgt}^m, \quad \mu \sim U[0,1] \quad (5)$$

Where $\tilde{r}$ is obtained by sampling from the sample space of the Y_R' distribution, which is implicitly defined sampling randomly along straight lines between a pair of points sampled in the source and target distribution of the mention representation.

3.4 The whole training process

We adopt the original cross-entropy loss expressed as $L_{CE}(\tilde{z}, z)$, where $\tilde{z}$ and z represent the predicted label distribution and the corresponding true label. Finally, combined with entity classifier and the adversarial training, the entire training loss that should be minimized, is given as:

$$L = L_{CE} + \lambda(\mathop{\mathbb{E}}_{r_{tgt}^m \sim Y_R^{tgt}}[F(r_{tgt}^m)] - \mathop{\mathbb{E}}_{r_{src}^m \sim Y_R^{src}}[F(r_{src}^m)]) \quad (6)$$

where λ is the balance factor. The training process of our proposed adversarial learning framework is illustrated in the Algorithm 1.

Algorithm 1 The training process of our proposed adversarial learning framework

Require: Labeled source text T_{src} (mention m_{src}, entity description e_{src}), unlabeled target text T_{tgt} (mention m_{tgt}), gradient penalty coefficient λ_p, hyper-parameter $\lambda > 0$ number of critic iterations per generator n_{critic}, maximum number of iterations n_{epoch}, and number of batches n_{batch}.

for t = 0 to n_{epoch} **do**
  for i = 0 to n_{batch} **do**
    for j = 0 to n_{critic} **do**
      Sample unlabeled source data m_{src} from T_{src}
      Sample unlabeled target data m_{tgt} from T_{tgt}
      A random number $\mu \sim U[0,1]$
      $r_{src}^m = R(m_{src})$
      $r_{tgt}^m = R(m_{tgt})$
      $\tilde{r} = \mu r_{src}^m + (1-\mu) r_{tgt}^m$
      ▷ Calculate loss
      $L_p = \mathbb{E}[(||\nabla_{\tilde{r}} F(\tilde{r})||_2 - 1)^2]$
      $L_{adv} = -\mathbb{E}[F(r_{src}^m)] + \mathbb{E}[F(r_{tgt}^m)] + \lambda_p L_p$
    end for
    Update F parameters with Adam to minimize L_{adv}
  end for
  ▷ Main iterations
  Sample labeled source data m_{src} and e_{src} from T_{src}
  Sample unlabeled target data m_{tgt} from T_{tgt}
  $r_{src}^m = R(m_{src})$
  $r_{tgt}^m = R(m_{tgt})$
  ▷ Calculate loss
  $L = L_{CE}(C(r_{src}^m); e_{src}) + \lambda(\mathbb{E}[F(r_{src}^m)] - \mathbb{E}[F(r_{tgt}^m)])$
  Updata R parameters with Adam to minimize loss.
end for

Table 1: Accuracy (acc), precision (p), recall (r), and F1 of four languages in three few-shot of 1%, 5%, and 10%.

	es				zh				de			
Split	acc	p	r	F1	acc	p	r	F1	acc	p	r	F1
1%	85.6	92.4	67.9	78.3	90.0	87.8	76.5	81.8	66.1	80.2	86.1	83.0
5%	86.3	93.2	66.8	77.8	91.5	90.3	84.9	87.5	68.4	82.3	86.7	84.4
10%	88.9	93.8	68.2	79.0	92.4	93.5	88.6	91.0	70.1	82.5	87.2	84.8

4 Experiment

4.1 Datasets and Settings

We conduct our evaluation on two well-known entity linking datasets.

- **TAC-KBP 2015**(Ji et al., 2015): following (Sil et al., 2018), we use Spanish and Chinese on TAC-KBP 2015 Tri-Lingual Entity Linking Track, which contains 166 Chinese documents (82 discussion forum articles and 84 news) and 167 Spanish documents (83 discussion forum articles and 84 news).

- **TR 2016**^{hard}(Tsai and Roth, 2016): is a cross-lingual dataset based on Wikipedia. It's constructed to contain difficult mention-entity pairs and removed the mention overlapping between training and test data.

In our experiment, the balance factor in Eq. 3 and Eq. 6 are set to 1. For all the experiments on each language, R and C are optimized by Adam (Kingma and Ba, 2015) with a learning rate of 0.0005, while F is trained through different Adam optimizers with the same learning rate. In order to present the effectiveness of the language discriminator F intuitively, our model using the adversarial approach is referred to as Model X^+, and the model without the adversarial approach is described as Model X. Except for training data, the target instances were selected randomly from test data to implement adversarial training at a small amount proportion of 1%, 5%, and 10% respectively. As for entity candidates, we use FAISS (Johnson et al., 2021) IndexFlatIP index type to obtain the top 100 entity candidates.

4.2 Main results

We first explore the performance of the English training model in an unseen language. This result presents the challenge of solving the entity linking task with a few examples per language. We carry out three settings used in few-shot learning (Gao et al., 2021): taking 1%, 5% and 10% test data as target instances. For each language in two datasets - Spanish (es) and Chinese (zh) in TAC-KBP 2015, German (de) in TR 2016^{hard}, we train our proposed model and demonstrate four indicators including accuracy (acc), precision (p), recall (r), and F1 in difference few-shot settings shown in Table 1. As we can see, with the increase in the few-shot examples, indicators show an upward trend more or less.

Table 2: Accuracy (%) results of ablation study in four languages under the circumstance of 10% few-shot setting. AT represents adversarial learning.

Model	es	zh	de
BERT	78.4	80.3	59.2
BERT + AT	81.2	86.8	65.4
XLM-R	83.5	89.2	64.6
XLM-R + AT	**88.9**	**92.4**	**70.1**

4.3 Ablation study

We launched an ablation study to explore the impact of different components in the proposed adversarial learning framework, and the results are reported in Table 2. From two components between the pre-trained model and whether there is an adversarial training approach or not, we additionally introduce a BERT pre-trained model (Vaswani et al., 2017; Devlin et al., 2019) initialized by Botha et al. (Botha et al., 2020b) using the first 4 layers. Note that XLM-R pre-trained model which can extract robust representations in a wide range of languages, performs better than BERT. Moreover, the removal of the adversarial training approach leads to performance degradation. This implies that the proposed adversarial learning framework with XLM-R pre-trained model and adversarial training advances the performance.

Table 3: Accuracy (%) on TAC-KBP 2015 and TR 2016^{hard}

Model	TAC-KBP2015		TR2016^{hard}			
	es	zh	de	es	fr	it
Sil et al.(Sil et al., 2018)	82.3	84.4	-	-	-	-
Upadhyay et al.(Upadhyay et al., 2018a)	84.4	86.0	55.2	56.8	51.0	52.3
Zhou et al.(Zhou et al., 2019b)	85.5	83.3	-	-	-	-
Botha et al.(Botha et al., 2020b)	-	-	62.0	58.0	54.0	56.0
Model X	84.6	87.2	61.2	58.3	55.2	55.5
Model X^+	**85.5**	**89.3**	**65.3**	**63.4**	**63.9**	**64.2**

4.4 Influence of Adversarial Training Approach

Many recent researchers fix their attention on the zero-shot setting that no mention is available during inference. Therefore, we conduct the following experiment based on a zero-shot setting. On this foundation, we investigate the influence of the adversarial training approach. From Table 2, it's concluded that the adversarial training approach helps better performance. More concretely, this section compares our model with the recent study in zero-shot setting, and the results are reported in Table 3 for TAC-KBP 2015 and TR 2016^{hard} using Model X and Model X^+. We can observe that our Model X^+ consistently outperforms all compared models at the same time. For the model proposed by Upadhyay et al (Upadhyay et al., 2018a), the best-improved results of TAC-KBP 2015 and TR 2016^{hard}respectively are 3.3% and 12.9%.

4.5 Visualization

To qualitatively demonstrate how our proposed adversarial learning framework affects the distribution between English and Chinese instances, we present a t-SNE (Van der Maaten and Hinton, 2008) visualization analysis of feature representations with 10 random mention texts from English and Chinese validation set respectively in Figure 3. Figure 3a shows representation distributions without adversarial training. Note that the two languages mention texts are not translations of each other. To shed light on the effect of our architecture, a significant reduction after adversarial training is presented in Figure 3b where we can see a more mixed distribution of representation between English and Chinese instances. This further indicates that our proposed adversarial learning framework effectively narrows the distance of representation distribution in different languages using the adversarial training approach.

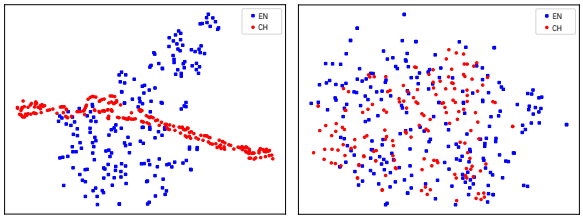

Figure 3: Results of t-SNE visualization. (a) The distribution of representation between English and Chinese instances without adversarial training presents a language gap. (b) A more mixed distribution of representation between English and Chinese instance with adversarial training at the end of the representation extractor shows a smaller language gap.

5 Conclusion

In this paper, we propose a novel model that applies adversarial learning and few-shot learning method to better generalize the learning ability across languages for the multi-lingual entity linking task. To be more exact, a fraction of language-agnostic unlabeled data are selected randomly as the language signal to build the language discriminator. Moreover, we design a simple and effective adversarial learning framework with two branches of an entity classifier and a language discriminator. Experimental results on two benchmark datasets empirically illustrate that the proposed adversarial learning framework is significantly effective.

Limitations

The current exploration, while demonstrating promising advancements, has areas for potential enhancement. Firstly, the study's focus on a limited number of languages may not fully capture the

breadth of linguistic diversity, potentially affecting the model's adaptability in multilingual scenarios. Secondly, variations in data quality could impact the robustness of the model's generalization capabilities.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (62176076), the Natural Science Foundation of Guangdong (2023A1515012922), the Shenzhen Foundational Research Funding (JCYJ20220818102415032), and the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (2022B1212010005).

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, Sydney Australia. PMLR.

Jan A Botha, Zifei Shan, and Dan Gillick. 2020a. Entity linking in 100 languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845.

Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020b. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online. Association for Computational Linguistics.

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*, pages 8440–8451, Online. Association for Computational Linguistics.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. 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 3816–3830, Online. Association for Computational Linguistics.

Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. 2015. Overview of tac-kbp 2015 tri-lingual entity discovery and linking. In *Proceedings of Eighth Text Analysis Conference (TAC 2015)*, Maryland, USA. National Institute of Standards and Technology.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Hideaki Joko, Faegheh Hasibi, Krisztian Balog, and Arjen P. de Vries. 2021. *Conversational Entity Linking: Problem Definition and Datasets*, page 2390–2397. Association for Computing Machinery, New York, NY, USA.

Diederik P Kingma and JL Ba. 2015. Adam: A method for stochastic optimization. In *Conference Track Proceedings*, San Diego, CA, USA. 3rd International Conference on Learning Representations, {ICLR} 2015.

Belinda Z. Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. 2020. Efficient one-pass end-to-end entity linking for questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6433–6441, Online. Association for Computational Linguistics.

Shengzhe Liu, Xin Zhang, and Jufeng Yang. 2022. Ser30k: A large-scale dataset for sticker emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 33–41.

Paul McNamee, James Mayfield, Dawn Lawrie, Douglas Oard, and David Doermann. 2011. Cross-language entity linking. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 255–263, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Minh C. Phan and Aixin Sun. 2018. Conerel: Collective information extraction in news articles. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1273–1276, New York, NY, USA. Association for Computing Machinery.

Avirup Sil, Gourab Kundu, Radu Florian, and Wael Hamza. 2018. Neural cross-lingual entity linking. In

Thirty-Second AAAI Conference on Artificial Intelligence, Louisiana, USA. Association for the Advancement of Artificial Intelligence.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598, San Diego, California. Association for Computational Linguistics.

Shyam Upadhyay, Nitish Gupta, and Dan Roth. 2018a. Joint multilingual supervision for cross-lingual entity linking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2495, Brussels, Belgium. Association for Computational Linguistics.

Shyam Upadhyay, Nitish Gupta, and Dan Roth. 2018b. Joint multilingual supervision for cross-lingual entity linking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2495.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998,6009, Long Beach, CA, USA. Curran Associates, Inc.

Cédric Villani. 2009. *Optimal transport: old and new*, volume 338. Springer.

Shuyan Zhou, Shruti Rijhwani, and Graham Neubig. 2019a. Towards zero-resource cross-lingual entity linking. *EMNLP-IJCNLP 2019*, page 243.

Shuyan Zhou, Shruti Rijhwani, and Graham Neubig. 2019b. Towards zero-resource cross-lingual entity linking. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 243–252, Hong Kong, China. Association for Computational Linguistics.