Synthesized Corpora for Tagging Japanese Product Attribute Values in Low-resource Settings

Yuki Nakayama[*]
yuki.b.nakayama@rakuten.com
Rakuten Institute of Technology, Rakuten, Inc.
Tokyo, Japan

Koji Murakami
koji.murakami@rakuten.com
Rakuten Institute of Technology, Rakuten, Inc.
Tokyo, Japan

Erick Mendieta
erick.mendieta@rakuten.com
Rakuten Institute of Technology, Rakuten, Inc.
Tokyo, Japan

Chen Zhao
chen.a.zhao@rakuten.com
Rakuten Institute of Technology, Rakuten, Inc.
Tokyo, Japan

ABSTRACT

Tagging product information in search queries from user inputs remains a challenging issue despite recent advancements in general name entity recognition (NER) tasks. It is primarily due to the short-length nature of input search queries typically containing only a few keywords. For Japanese search queries in particular, this problem is further exacerbated by lack of public and reliably annotated datasets, as most product-related query information is kept confidential by major E-Commerce platforms. In this paper, we mitigate this issue by demonstrating a complete workflow of Japanese query text augmentation based on a variational auto-encoder (VAE). Once the training of the VAE is completed a massive amount of search queries which we call "*virtual queries*" can automatically be synthesized without expensive manual effort in annotation. Experiments show that training a sequence tagging model using purely synthesized datasets delivers consistently better NER performance in term of four product attributes, compared to manually labeled ground truth datasets in low-resource scenarios. Both the trained model of VAE and synthesized corpora will be released to help contribute to future NER-related research on short Japanese text.

KEYWORDS

NER, corpus creation, sequence labeling, auto-encoder, natural language generation

ACM Reference Format:
Yuki Nakayama, Koji Murakami, Erick Mendieta, and Chen Zhao. 2022. Synthesized Corpora for Tagging Japanese Product Attribute Values in Low-resource Settings. In *Proceedings of ACM SIGIR Workshop on eCommerce (SIGIR eCom'22).* ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

The classic name entity recognition (NER) [11] task has achieved quite a few milestones of groundbreaking success, thanks to recent contributions [2] [1] [5] to sequence labeling problems from the NLP research community. While NER has gained impressive performance on benchmark datasets like CoNLL2003 [14] numerous challenges remain in practical use cases. This paper tackles the resource problem in detecting product information from Japanese search queries collected from user inputs. Within the scope of this paper, four attributes are covered: product brand, item color, material and size depiction. In addition to dealing with short texts, there also exist several language-specific issues. For instance, the same product brand may possess several valid synonyms including Latin letters (Romaji) and Katakana symbols. Expression patterns of item sizes depend on item category because Japanese language follows separate rules of applying quantifiers depending on objects. Formulated as a general NER task, product attribute tagging for raw search queries remains a tough task since it requires massive amounts of well-labeled training examples. It further requires the training data to cover a wide range of synonyms due to aforementioned reasons. Unfortunately, search query data of both diversity and comprehensiveness is only practically available from E-Commerce platforms with a large user base. Very rarely would platforms like those willingly release non-trivial real-world search query logs as are commonly considered proprietary and confidential. The complexity of manually labeling Japanese search queries further leads to even rarer publicly accessible training data of high quality. In this paper, we propose an augmentation scheme tailored for Japanese product attributes to mitigate the issue of corpus scarcity. It is illustrated through experiments that purely augmented search query data is still of considerable quality on par with real data, given that only a limited amount of ground truth examples are useable.

Our data augmentation scheme is based on a simple variational auto-encoder (VAE) [10] that is trained on only 9,000 manually annotated samples followed by a much larger pool of raw search queries, among which product attributes are only approximately labeled using our pre-built Japanese product brand dictionaries. Given proper sampling noise, the decoder part of the VAE is capable of generating a magnitude more search queries than the labeled examples we feed into the VAE. As an interesting characteristic those synthesized search queries contain both real product brands and non-existent brands which can hardly be discerned by human observation without knowledge from external references. Nonetheless, a sequence tagger trained with a large volume of synthesized samples still exhibits quite decent performance on recognizing

[*] Authors contributed equally to this research.

SIGIR eCom'22, July 15, 2022, Madrid, Spain

brands, color/material and item size information. Consistently better metrics on all 4 attributes have been achieved, compared to either one or both baseline models used in our experiments.

The rest of this paper briefly introduces related works regarding low-resource sequence labeling tasks, then describes how the VAE is trained and used for generation. Next it mentions some important preprocessing and postprocessing steps that are critical to the quality of augmented dataset. Details on experiment and evaluation metrics are then discussed before we come to conclusion.

2 TASK DEFINITION

The core problem of this paper is to develop a sustainable way of synthesizing virtual queries that resemble the search queries collected by typical ECommerce search engines. Functionally, we require that virtual queries possess sufficient data integrity to be qualified as reasonable training data. With absolutely no real datasets, we desire to train a better sequence tagger than one we do with limited amounts of ground truth labels.

This work presents 2-fold contributions. First, the augmented corpus with labels will be released so that fellow researchers working on Japanese product attribute tagging would have easier access to a versatile yet virtual dataset for training a moderately functional sequence tagger. Second, we provide some insights about how VAE could help relieve resource limitation for short-text NER problems. Our well-tuned VAE will also be released so that anyone with access to the model is able to synthesize virtual queries of much higher quantities than what are used in our experiments.

3 RELATED WORKS

Recent works regarding NER tasks on E-Commerce search queries are mostly focused on English data, represented by: *TripleLearn* [4] that iteratively trains a model on multiple training sets of different quality; and the framework by Wen et al. [15] as an effective aspect recognition system to optimize search relevance. As for NER application under resource constraint, Hedderich et al. devised a noise layer on top of BiLSTM outputs [8] so that automatically labeled noisy datasets still enable a 35% F1 increase on *CoNLL2003* [14] test set. Other low-resource settings mainly involve cross-lingual knowledge sharing relying on either translation [16] or cross-lingual knowledge extraction [7] [13] through shared features in neural networks. New architecture has also been investigated [17] where adversarial training helps improve NER performance for Dutch and Spanish. A data augmentation technique is also introduced as *DAGA* [6] which is made of a simple RNN language model. However, no evaluation metrics for NER tasks on *DAGA*-generated Japanese corpora are ever reported. While general purpose NER corpora in Japanese are already available [9], they are not closely related to product attribute detection. Our VAE is trained directly on pseudo-labeled search queries and differs from a stand-alone RNN language model learning only to reconstruct training samples. Instead we enforce input sequences at the character level during the training phase and introduce sampling noise into decoding phase to avoid repeating identical sequences from the training data.

4 LANGUAGE MODEL, AUTO-ENCODER AND CORPUS AUGMENTATION

A simple language model is trained to learn the conditional probability distribution from example. Given an example sentence $\boldsymbol{w} = (w_1, ..., w_t, ...)$, the language model is supposed to learn $p = \Pi_t Pr(w_t | w_{<t}, w_{>t})$, where $(w_{<t}, w_{>t})$ refers to the context surrounding the tth word in the same sentence. For sentence reconstruction, the language model is expected to make prediction on word w_t based on inference probability $\hat{p} = \Pi_t \hat{Pr}(w_t | w_{<t}, w_{>t})$ so that $w_t = \text{argmax}_w \Pi_t \hat{Pr}(w_t | w_{<t}, w_{>t})$. A general RNN-based language model [12] can be trained by minimizing the cross-entropy loss $\mathcal{L}(p, \hat{p}) = -\sum_i p_i \log \hat{p}_i$ over the training set. For corpus augmentation purpose, we train a similar language model through VAE, which is basically a encoder-decoder structure based on double-layered BiLSTM cells. After an input sentence $\boldsymbol{w} = (w_1, ..., w_t, ...)$ is embedded as $\boldsymbol{x} = (x_1, ..., x_t, ...)$, it is fed into the encoder f_{enc} through embedding lookup so that the final hidden encoder state is $c = f^h_{enc}(\boldsymbol{x})$, which is mapped into an array of distributions $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. To ensure differentiability over the entire VAE, the following reparameterization tricks are applied.

$$\boldsymbol{\mu} = cW^{\mu}, \ \boldsymbol{\sigma} = cW^{\sigma}$$

$$\boldsymbol{z} = \boldsymbol{\mu} + \epsilon \boldsymbol{\sigma}, \epsilon \sim \mathcal{N}(0, 1)$$

$$\boldsymbol{h} = zW^z$$

The decoder f_{dec} takes the hidden and cell states to compute the final decoder outputs as $\hat{x} = f^c_{dec}(\boldsymbol{h}, z)$.[1] To train the VAE, we require that the auto-encoder to learn input sentence patterns and the variational parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ to move towards $\mathcal{N}(0, 1)$. For each training sample w, this leads to a loss function consisting of cross-entropy $\mathcal{L}^{ce}(w_t, \hat{x}_t) = \sum_t -w_t \log \hat{x}_t$ and $KL[\mathcal{N}(0, 1), \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})]$, yielding the total loss $\mathcal{L}(w_t, \hat{x}_t)$, w_t being one-hot encoded. The training loss is summarized as

$$\mathcal{L}(w, \hat{x}) = \sum_t \mathcal{L}^{ce}(w_t, \hat{x}_t) + KL[\mathcal{N}(0, 1), \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})]$$

where

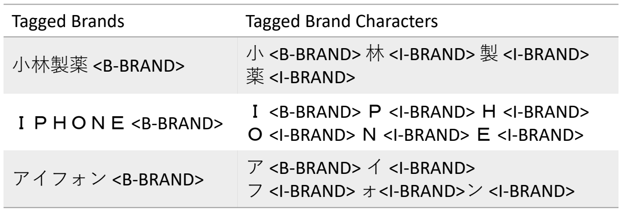
$$KL[\mathcal{N}(0, 1), \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})] = -\log \boldsymbol{\sigma} + \frac{\boldsymbol{\sigma}^2 + \boldsymbol{\mu}^2}{2} - \frac{1}{2}$$

For sequence augmentation, only the decoder part is used. Note that during the generation phase $\boldsymbol{z}$ and $\boldsymbol{h}$ become stochastic. The output token is selected on top of another fully connected layer followed by softmax. We also perform double sampling on $\boldsymbol{z}$ for every output sentence such that $\hat{w} = \frac{1}{2}[f^c_{dec}(\boldsymbol{h}, \boldsymbol{z}) + f^c_{dec}(\boldsymbol{h}, \boldsymbol{z})]W^o$. The final word is determined by the vocabulary index $\tilde{w} = \text{argmax}_i \hat{w}$ where $\hat{w}_i = \frac{e^{\hat{w}_i}}{\sum_j e^{\hat{w}_j}}, j \in |V|$.

5 PREPROCESSING AND POST PROCESSING

Very little preprocessing is needed for training a VAE. We break up brands in every tagged sentence into characters following *BIO*-based tagging convention, as is shown in Figure 1. Note that the same process is done on every brand regardless of representation including alphanumeric brands and brands in Kanji/Katakana. This

[1] The input sequence is omitted from the notation for simplicity. Teacher forcing is adopted so the input to each decoder cell is the previous token in the same sentence.



Tagged Brands	Tagged Brand Characters
小林製薬 <B-BRAND>	小 <B-BRAND> 林 <I-BRAND> 製 <I-BRAND> 薬 <I-BRAND>
ＩＰＨＯＮＥ <B-BRAND>	Ｉ <B-BRAND> Ｐ <I-BRAND> Ｈ <I-BRAND> Ｏ <I-BRAND> Ｎ <I-BRAND> Ｅ <I-BRAND>
アイフォン <B-BRAND>	ア <B-BRAND> イ <I-BRAND> フ <I-BRAND> ォ<I-BRAND>ン <I-BRAND>

Figure 1: Brand as Character Sequences

enables the VAE to synthesize non-existing brands with Japanese brand style.

We apply postprocessing with the following 3 criteria.

(1) All the augmented brands require a minimum length of 3 characters.
(2) Augmented brands of purely numeric characters are not allowed.
(3) Augmented brands of purely alphanumeric characters must have 5 characters at minimum.

Augmented search queries which do not comply with any of the three criteria are discarded, in order to eliminate outrightly ridiculous brands generated by the VAE.

6 EXPERIMENTS

6.1 Dataset Overview

Every non-synthesized dataset involved in this paper comes from raw user inputs that are logged by the search engine of our main ECommerce service. Table 2 lists out a few raw samples together with tagged ones.

Experiments in this paper are designed to validate the effectiveness of synthesized datasets in low-resource setting. Three independent search query datasets are prepared and used as candidate training sets for training sequence tagging models. They include a small manually labelled dataset, a pseudo-labelled dataset and a purely virtual dataset augmented by the VAE. The pseudo-labelled dataset comes from raw search queries without annotation, whose keywords are then automatically matched and tagged in a non-comprehensive approach based on our internal product attribute dictionary. All the samples in our datasets are collected from recent search query logs except the augmented ones. Human annotation only covers 9,000 real search queries. Whereas several different types of characters constitutes all our datasets, as is mentioned in Section 1, some basic patterns of query words are analyzed in Table 4.

Distinct Count	Brand	Size	Color	Material
Manual	1,673	159	64	261
Dict. Matched Keywords	11,331	306	162	546
Augmented	202,728	3,296	239	597
Evaluation	965	86	42	113

Table 1: Attribute Value Counts per Dataset

Raw Search Queries	Annotated Search Queries
子供 ドレス 150cm	子供　ドレス　１５０ＣＭ <B-SIZE>
パブロン 風邪薬 キッズ	パブロン <B-BRAND> 風邪薬 キッズ
蛍光灯 直管	蛍光灯　直管
スマホ グーグルPixel6保護フィルム	スマホ　グーグル <B-BRAND> ＰＩＸＥＬ<B-BRAND> ６保護フィルム
こだわり 安眠館 布団カバーガーゼ	こだわり <B-BRAND> 安眠館 <I-BRAND>　布団カバーガーゼ <B-MATERIAL>
白い翡翠	白い <B-COLOR> 翡翠

Table 2: Examples of the Raw Query Log and Labeled Samples

Table 1 displays distinctive value counts of all 4 attributes in every candidate dataset. The augmented set stands out with a magnitude higher count of distinctive brands and sizes. In case of brands, it is expected behavior since every input brand was disassembled before training the VAE model. This results in a high number of virtual brands that possess brand-ish patterns meanwhile never existing in the real-world. As for size attribute, the main cause is its frequent involvement with numbers, which tend to fuse diversity into token generation process, the simplest example being "10ML"→"100ML","1000ML".

To further investigate attribute value patterns among augmented search queries, Table 3 tracks for each attribute how many distinctive values are purely synthesized, i.e., augmented attribute values that never exist in real datasets. As is shown by examples, it is not surprising to see that a majority of brand/size attribute values in the augmented training set are imaginary values. Yet they share common patterns with input phrases from real users, despite a few non-idiomatic tokens that do not make perfect sense.

Attribute	Total	Examples
Brand	199,135	ＺＡＮＨＯ, 銀鳥フーズ, 田乃谷純銅, 犬のＨＯＮＥＹ, 小ふるさと製薬, ...
Size	2,955	１セット××１０ＭＭ, 四つ折り, ７７２ＭＬ, １フィート, ...
Color	80	黒金,緑黄, ホワイトオーク, ＷＨＩＴＥＲ, ...
Material	114	過炭酸, 人工皮革, 炭酸脂肪, ...

Table 3: Purely Synthesized Attribute Values

Dataset Name (# Query Examples)	Average Word Length	% Kanji Characters	% Hiragana/Katakana Characters	% Alphanumeric Characters
Manual (9,000)	3.23	0.12	0.71	0.17
Dict. Matched Keywords (28,000)	2.94	0.12	0.72	0.16
Augmented (673,967)	2.74	0.15	0.61	0.23
Evaluation (2,250)	3.21	0.14	0.71	0.15

Table 4: Word/Char Patterns among Datasets

	Training Data	Attributes (P/R/F1)			
		Brand	Size	Color	Material
Baseline Model	9K Manually Labeled	0.58	0.49	0.81	0.58
		0.59	0.39	0.55	0.39
		0.59	0.44	0.65	0.47
Baseline Model (Better)	9K Manual + 28K Keyword Match	0.67	0.63	**0.93**	0.87
		0.56	0.85	**0.86**	0.72
		0.59	0.72	**0.89**	0.78
Target Model	673K Augmented Queries	**0.59**	**0.75**	0.88	**0.93**
		0.68	**0.86**	0.84	**0.89**
		0.63	**0.80**	0.86	**0.92**

Table 5: Evaluation Metrics per Dataset

6.2 Training Details

As the building blocks of the VAE encoder/decoder, we stack two layers of BiLSTM cells with hidden size 1024 and a dropout rate of 0.5 in between. The embedding dimension is constantly 300. The input data for training the VAE include both manually labeled and pseudo-labeled datasets. In addition to preprocessing steps mentioned in Section 4, we truncate the maximum vocabulary size to 50,000. A mini-batch size of 24 is selected. During the generation phase, the maximum length of output sequences is restricted at 24.

For sequence tagging models, we adopt the same architecture as was used for NER tasks with contextual string embedding [1]. We stack three layers of embeddings above the general BiLSTM-CRF structure: trainable character embedding of hidden size 25, pretrained fastText [3] word embedding and pretrained Flair embedding. Both the fastText word vectors and Flair embedding are learnt from massive amounts of Japanese product description text. The Flair embedding we pretrained contains both forward and backward representations. The upper limit of training epochs is 100 per sequence tagging model. A mini-batch size of 64 is chosen regardless of which training set is used. The sequence tagging model is trained with a simple SGD optimizer of decay rate 0.5 while the learning rate starts from 0.1. No development set is held out in advance as the devopment score is computed on random 10% samples from training data. Early stop is also enforced when the learning rate drops below 10^{-4}.

6.3 Evaluations and Discussions

The comparative experiments are respectively conducted on separate datasets described in Section 5.1. Sequence tagging performance is evaluated on our own benchmark dataset that consists of 2250 real search queries with reliable NER tags from human annotations. This evaluation set features intentional emphasis on product brands, such that a significant proportion of brands are rarely known in real life and nor do they ever exist in any of the three training sets. Table 1 also lists out distinctive counts of all four attributes in the evaluation set. The second row counts attribute values that are only present in the test set without being visible in any training samples. To emphasize, more than one third brands are exclusive to evaluation, which makes "brand" the most difficult attribute to recognize among all the four attributes.

Details of evaluation metrics (precision/recall/F1) are listed in Table 5. The first model stands for a typical low-resource scenario where ground truth samples are limited to 9,000. The second model is trained with a similar dataset but all its 28,000 samples are pseudo-labeled. While the training set is still quite limited, significantly better scores show up for three attributes excluding brands. Recall of brands remain a problem, which is expected behavior caused by rare brands. The target model trained with and **only** with a large chunk of 673K purely augmented search queries achieves the best scores other than "color" that loses a small margin to the second baseline. Its drastic increase in brand recall verifies the efficacy of augmented "virtual" brands at the inevitable cost of some precision.[2] On the other, size/material attributes turn out to be remarkable success. It is likely due to the clear textual patterns indicating those two attributes, which is definitely not the case for brands. Another possible reason of the success of "virtual" dataset is that augmented samples tend to be more focused on kanji/alphanumeric words (Table 4) complementing what are fewer in real datasets. The overall evaluations validate the usefulness of purely synthesized search query datasets for product-related NER tasks.

7 CONCLUSION

In this paper, we bring forward the persistent issue of language resource constraint as a hurdle against Japanese short-text NER tasks. We take advantage of existing product attribute information we have and craft pseudo-labeled datasets. With the help of the pseudo-labeled datasets along with very few manually labeled examples, a variational auto-encoder can be trained into a search query synthesizer that is capable of generating quasi-unlimited search queries as training data.[3] Comparative experiments indicate that a sequence tagging model is able to attain impressive performance on four product-related attributes without direct involvement of real datasets. Henceforth we confirm that the synthesized corpora are valid resource for tagging Japanese search queries when large volumes of user queries are not accessible or ground truth labels are prohibitively expensive to obtain. We also consider expanding our scope into many more product attributes such as item condition, person names, target users, etc. in future.

REFERENCES

[1] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference*

[2] The vagueness of "what is a brand" makes it challenging even for humans to correctly label brand-like tokens. To clarify whether a text chunk is officially a brand, we rely on its domestic registration status from https://www.j-platpat.inpit.go.jp/

[3] The final checkpoint of data augmenter and samples of virtual datasets will be documented at https://bitbucket.org/vivi489/rit_sqa_ner/src/master/.

on Computational Linguistics. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1638–1649. https://aclanthology.org/C18-1139

[2] Bernd Bohnet, Ryan McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler, and Joshua Maynez. 2018. Morphosyntactic Tagging with a Meta-BiLSTM Model over Context Sensitive Token Encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 2642–2652. https://doi.org/10.18653/v1/P18-1246

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (07 2016). https://doi.org/10.1162/tacl_a_00051

[4] Xiang Cheng, Mitchell Bowden, Bhushan Ramesh Bhange, Priyanka Goyal, Thomas Packer, and Faizan Javed. 2021. An End-to-End Solution for Named Entity Recognition in eCommerce Search. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 17 (May 2021), 15098–15106.

[5] 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, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[6] Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data Augmentation with a Generation Approach for Low-resource Tagging Tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6045–6057. https://doi.org/10.18653/v1/2020.emnlp-main.488

[7] Xiaocheng Feng, Xiachong Feng, Bing Qin, Zhangyin Feng, and Ting Liu. 2018. Improving Low Resource Named Entity Recognition using Cross-lingual Knowledge Transfer. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 4071–4077. https://doi.org/10.24963/ijcai.2018/566

[8] Michael Hedderich and Dietrich Klakow. 2018. Training a Neural Network in a Low-Resource Setting on Automatically Annotated Noisy Data. 12–18. https://doi.org/10.18653/v1/W18-3402

[9] Tomoya Iwakura, Kanako Komiya, and Ryuichi Tachibana. 2016. Constructing a Japanese Basic Named Entity Corpus of Various Genres. In *Proceedings of the Sixth Named Entity Workshop*. Association for Computational Linguistics, Berlin, Germany, 41–46. https://doi.org/10.18653/v1/W16-2706

[10] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. arXiv:http://arxiv.org/abs/1312.6114v10 [stat.ML]

[11] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 282–289.

[12] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010* 2, 1045–1048.

[13] Rudra Murthy, Mitesh M Khapra, and Pushpak Bhattacharyya. 2018. Improving NER tagging performance in low-resource languages via multilingual learning. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 18, 2 (2018), 1–20.

[14] Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. https://aclanthology.org/W02-2024

[15] Musen Wen, Deepak Kumar Vasthimal, Alan Lu, Tian Wang, and Aimin Guo. 2019. Building Large-Scale Deep Learning System for Entity Recognition in E-Commerce Search. In *Proceedings of the 6th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies* (Auckland, New Zealand) *(BDCAT '19)*. Association for Computing Machinery, New York, NY, USA, 149–154. https://doi.org/10.1145/3365109.3368765

[16] Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural Cross-Lingual Named Entity Recognition with Minimal Resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 369–379. https://doi.org/10.18653/v1/D18-1034

[17] Joey Tianyi Zhou, Hao Zhang, Di Jin, Hongyuan Zhu, Meng Fang, Rick Siow Mong Goh, and Kenneth Kwok. 2019. Dual Adversarial Neural Transfer for Low-Resource Named Entity Recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3461–3471. https://doi.org/10.18653/v1/P19-1336