Multichannel LSTM-CRF for Named Entity Recognition in Chinese Social Media

Chuanhai Dong[1,2(✉)], Huijia Wu[1,2], Jiajun Zhang[1,2], and Chengqing Zong[1,2,3]

[1] CASIA, National Laboratory of Pattern Recognition, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China
{chuanhai.dong,huijia.wu,jjzhang,cqzong}@nlpr.ia.ac.cn

Abstract. Named Entity Recognition (NER) is a tough task in Chinese social media due to a large portion of informal writings. Existing research uses only limited in-domain annotated data and achieves low performance. In this paper, we utilize both limited in-domain data and enough out-of-domain data using a domain adaptation method. We propose a multichannel LSTM-CRF model that employs different channels to capture general patterns, in-domain patterns and out-of-domain patterns in Chinese social media. The extensive experiments show that our model yields 9.8% improvement over previous state-of-the-art methods. We further find that a shared embedding layer is important and randomly initialized embeddings are better than the pretrained ones.

Keywords: Multichannel · Named entity recognition · Chinese social media

1 Introduction

Named Entity Recognition is a fundamental technique for many natural language processing applications such as information extraction [2] and entity linking [13]. With the development of Internet, more and more researches turn towards NER in social media [10,31]. Social media texts are informal and mixed with strong noise which makes it more challenging to recognize named entities. Research on NER in English has narrowed the gap between social media and formal domains [4], but NER in Chinese social media is still hard [26].

One important reason that limits NER in Chinese social media is that there is rare annotated data for supervised learning. For example, the training set of Weibo NER corpora [26], which come from Sina Weibo service (comparable in size and popularity to Twitter), is less than 1/30 of MSRA training set in the third SIGHAN Bakeoff Chinese NER shared task [21] in size. It's difficult to achieve comparable results using such rare data, let alone its informality and strong noise. However, since manual annotation is time consuming and costs expensive, we choose to use out-of-domain annotated data to improve in-domain NER results using domain adaptation method.

M. Sun et al. (Eds.): CCL 2017 and NLP-NABD 2017, LNAI 10565, pp. 197–208, 2017.
https://doi.org/10.1007/978-3-319-69005-6_17

We consider domain adaptation method for NER in Chinese social media. Generally, we would train a model on all available data for a given task and test it on the same domain. However, MSRA data and Weibo data are from different domains [25,32]. The former comes from news and the latter comes from social media. In this case, distribution changes in the different input domains make generalizing across them difficult [28].

There have been much progress in domain adaptation [1,8,33]. A notable example is the feature augmentation method [7], whose key insight is that if we partition the model parameters to those that handle general patterns and those that handle domain-specific patterns, the model is forced to learn from all domains yet preserve domain-specific knowledge [19].

In this paper, we extend the feature augmentation method to multichannel LSTM-CRF for NER in Chinese social media. We make the following contributions:

- We propose three LSTM channels where one captures general patterns and the other two capture source domain patterns and target domain patterns, respectively. After concatenating three LSTM channels to a shared hidden layer, we propose domain-specific CRF layers to decode for different domains.
- We find that a shared embedding layer is important for improving performance. Randomly initialized embeddings are better than pretrained embeddings for multichannel architecture.
- We improve model's generalization ability using multichannel LSTM-CRF and achieve significant performance improvement.

2 Related Work

The problem of NER requires both boundary identification and type classification [3]. As the DEFT ERE Annotation Guidelines[1] shows, there are five entity types: person (PER), titles (TTL), organizations (ORG), geopolitical entities (GPE) and locations (LOC). A mention is a single occurrence of a name (NAM), nominal phrase (NOM) or pronominal phrase (PRO) that refers to or describes a single entity [16]. In Weibo NER dataset [26], which comes from Chinese social media, PER, ORG, GPE and LOC are considered, all including NAM and NOM mentions. In the third SIGHAN Bakeoff Chinese NER shared task [21] MSRA dataset, which comes from news, there are three types: PER, ORG and LOC, only in NAM mentions. In this paper, we focus on named entities mentions: PER, ORG and LOC.

Most related research regards NER as a sequence tagging task. Hidden Markov Model (HMM), Support Vector Machine (SVM) and Conditional Random Field (CRF) once achieved good results [12,14,22]. These classic methods need delicate hand-crafted features to obtain good results. In recent years, neural architectures show great learning power both for English NER [5,6,18,20,23] and Chinese NER [9]. These neural architectures don't need hand-crafted features

[1] Entities V1.7, Linguistic Data Consortium, 2014.

and some are end-to-end models which can be easily applied to other languages or similar tasks without data preprocessing. As Chinese words have no natural word boundaries, character-based tagging strategy simplifies NER without results of Chinese Word Segmentation (CWS) compared to word-based strategy. [9] achieve state-of-the-art performance on MSRA dataset using character-based LSTM-CRF architecture. We use character-based LSTM-CRF as our basic channel for a single domain.

Methods mentioned above have been restricted to formal text, e.g. news. It is difficult for NER in social media because of its informality and strong noise. There are many abbreviations, typos, novel words and ungrammatical constructions in social media. Chinese presents additional challenges, since it lacks of explicit word boundaries and other clues that indicate a named entity, e.g. capitalizations in English.

Related works on NER in Chinese social media focus on supervised learning using rare annotated corpora. [26] first release a Chinese social media corpora, namely Weibo NER data. They explore several types of embeddings using a CRF model and propose joint training objectives for embeddings and NER. [27] use Chinese word segmentation representation as features to improve NER. [15] propose a F-score driven training method through adding sentence level F-score to its label accuracy loss function. These methods only utilize rare in-domain corpora and get good precision but low recall rate, which is less than half of the recall rate trained on enough corpora [9]. [16] combine cross-domain supervised learning using out-of-domain annotated data with semi-supervised learning using in-domain unannotated data to achieve performance improvement. We extend feature augmentation method to multichannel LSTM-CRF and improve model's generalization ability with only out-of-domain data.

The feature augmentation method is first considered for sparse binary-valued features which underlie conventional NLP systems. They conjoin feature types with domain indicators as a kind of data preprocessing and use them alongside the original feature types. They extend the feature augmentation method to semi-supervised learning in [8]. [19] try a neural extension of the feature augmentation method on English slot tagging task and different domains in their task have different labels. Our work applies feature augmentation method to NER in Chinese social media using a novel multichannel LSTM-CRF model, which captures not only general patterns across formal and informal domains but also domain-specific patterns.

3 Model

We use character-based LSTM-CRF described in [9] as our single channel tagger. The architecture is shown in Fig. 1. We then extend it to multichannel LSTM-CRF with three LSTMs and two CRFs. These LSTMs respectively capture general patterns, source domain patterns and target domain patterns. Similarly, two CRF decoders are designed for source domain and target domain separately.

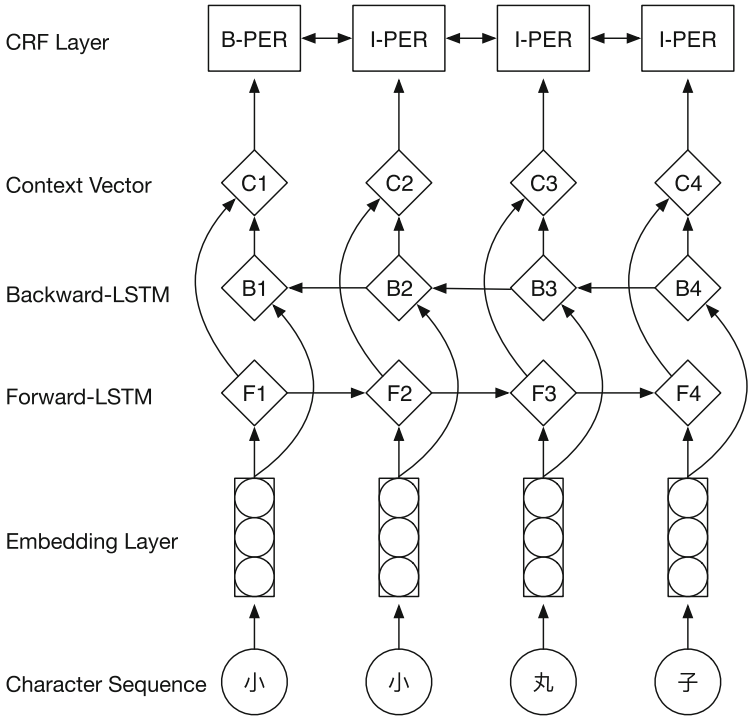

Fig. 1. The architecture of our character-based single channel BLSTM-CRF. The example means "little meat ball", a name in Chinese social media.

3.1 Multichannel LSTM-CRF

[7] propose the feature augmentation method for domain adaptation. All they do is to take each feature in the original problem and make three versions of it: a general version, a source-specific version and a target-specific version. The augmented source data contain general and source-specific versions. The augmented target data contain general and target-specific versions.

Formally, $\boldsymbol{x}$ is the input spaces; D^s is the source domain data set and D^t is the target domain data set. Define mappings Φ^s, Φ^t for mapping the source and target data to feature spaces respectively. Then after feature augmenting,

$$\Phi^s(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{x}, \mathbf{0} \rangle, \quad \Phi^t(\boldsymbol{x}) = \langle \boldsymbol{x}, \mathbf{0}, \boldsymbol{x} \rangle$$

where $\mathbf{0} = \langle 0, 0, ..., 0 \rangle$ is the zero vector. This approach can be easily applied to multi domains.

In our multichannel LSTM-CRF model, we use three LSTM channels: one general LSTM, one source domain LSTM and one target domain LSTM in Fig. 2. As there are general, source domain and target domain data, we try different pretrained embeddings (left part in Fig. 2). Each LSTM uses a domain-specific embedding layer which means different character vectors are put into different

LSTMs. All three LSTM outputs are concatenated through a mask vector. If the input data is from source domain, mask vector $m = [\vec{1}, \vec{1}, \vec{0}]$. Otherwise, if the input data is from target domain, mask vector $m = [\vec{1}, \vec{0}, \vec{1}]$. In this way, source domain training data help to learn general and source LSTM parameters and character embeddings, and target domain training data help to learn general and target LSTM parameters and character embeddings. We apply similar mask operation to CRF layer with two CRF decoders. Instinctively, source domain data and target domain data should have different transition probabilities between tags.

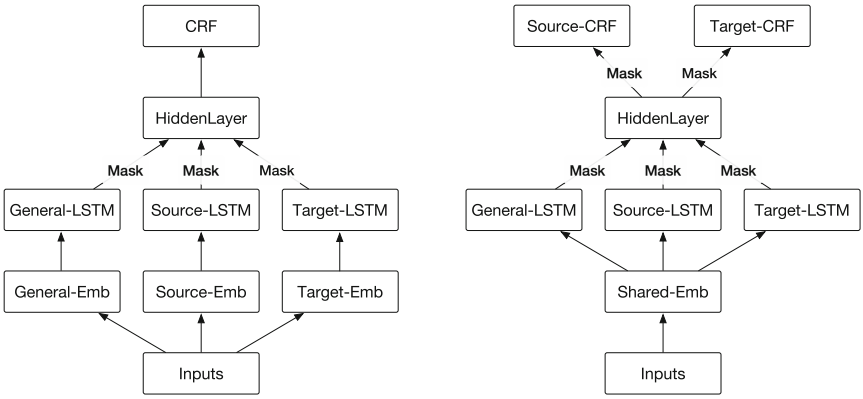

Fig. 2. The architecture of our model. First we use three embedding layers and LSTMs to learn general, source domain and target domain patterns (as shown in the left part). Then we adopt a shared embedding layer and 2 domain CRF decoders (as shown in the right part). We will explain the mask vector in Sect. 3.1.

3.2 Sharing Parameters

Our first model shares hidden layer and CRF layer parameters in the neural architecture. When the training sentence is from source domain, target domain embeddings and LSTM parameters are not updated. We notice that there is severe imbalance between source and target domain training data. Target domain training data is less than 1/30 of source domain training data. It means that most of the time during training, target channel LSTM is not trained. Related experiment is shown in Sect. 4.6.

To make our model share more parameters and make target channel learn more, we adjust our original architecture. We propose shared embedding layer but still keep multichannel LSTMs. No matter which domain the training sentence is from, the shared embedding layer can get updates. Meanwhile, multichannel LSTMs learn different domain patterns all the same. The proposed model can remember general, source and target patterns through multichannel

LSTMs and can be trained more effectively. At the same time, we use 2 different CRF layers respectively for source and target domain. The right part in Fig. 2 shows our improved architecture.

4 Experiments

To demonstrate the correctness and effectiveness of our framework, we do some experiments on NER datasets. We will describe the details of datasets, tagging scheme, pretrained embeddings, baselines, settings and results in our experiments.

4.1 Datasets

We use the same annotated corpora[2] as [16] which is called Weibo NER dataset for NER in Chinese social media. Weibo NER contains PER, ORG, GPE and LOC for both named and nominal mention. We use the training set of the third SIGHAN Bakeoff MSRA NER dataset as out-of-domain data. MSRA NER dataset contains only PER, ORG and LOC for named mention. We merge GPE and LOC as LOC for consistency. As we focus on named mention, we ignore nominal mention. The details of Weibo NER are shown in Table 1. The details of MSRA NER are shown in Table 2.

4.2 Tagging Scheme

As we use a character-based tagging strategy, we need to assign a named entity label to every character in a sentence. Many named entities span multiple characters in a sentence. Sentences are usually represented in the IOB format (Inside, Outside, Beginning). In this paper, we use IOBES tagging scheme. Using this scheme, more information about the following tag is considered [9,20]. Related works show that using a more expressive tagging scheme like IOBES improves performance [11,29].

4.3 Pretrained Embeddings

Previous works show that both pretrained word embeddings for English and pretrained character embeddings for Chinese improve performance significantly than randomly initialized embeddings [9,20]. We first use different embedding layers for different LSTM channels, because these LSTM channels tend to capture different domain patterns. For source domain, which is from news, we use unlabeled texts from the People's Daily (1994–2003) to pretrain Chinese character embeddings. Here we use gensim[3] [30], which contains a python version

[2] We just fix four obvious annotating errors with starting PER character tagged as 'I-PER' in the training set.

[3] https://radimrehurek.com/gensim/index.html.

Table 1. Details of Weibo NER dataset.

	Named Entity	Sentence
Train Set	957	1350
Dev Set	153	270
Test Set	211	270

Table 2. Details of MSRA NER dataset.

Entity Type	Train Set	Test Set
PER	17615	1973
LOC	36517	2877
ORG	20571	1331
Sentence	46364	4365

implementation of word2vec [24]. After simple preprocessing, such as unifying different styles of punctuations, we use CBOW model to train the embeddings because it is faster than skip-gram model. For target domain, which is from social media, we use the character embeddings in [26], which are pretrained using 2,259,434 unlabeled Weibo messages. Although [26] report 3 kinds of Chinese embeddings, we use the character embeddings without position information in order to be consistent with the other two domain pretrained embeddings. For general domain, we use the character embeddings in [9], which are pretrained using unlabeled Chinese Wikipedia backup dump of 20151201. All the embeddings have a dimension of 100. Source and general embeddings are trained using the same parameters while target embeddings is directly adopted from [26].

4.4 Baselines

We regard character-based single channel LSTM-CRF (described in Fig. 1) as baseline, which achieves state-of-the-art performance in news dataset [9]. We use the same LSTM block in all experiments and other settings are showed in Sect. 4.5.

4.5 Settings

We use dropout training [17] before the input to LSTM layer with a probability of 0.5 in order to avoid overfitting. We train our network using the back-propagation algorithm updating our parameters on every training example, one at a time. We use stochastic gradient decent (SGD) algorithm with a learning rate of 0.05 for 100 epochs on all training sets. Dimension of LSTM is 100. Dimension of Chinese character embeddings is also 100. Fine tuning is applied in all experiments to adjust the character embeddings. We adopt these hyper parameters according to [9,20].

4.6 Results

Multichannel and Embeddings: To demonstrate the effectiveness of our proposed multichannel LSTMs with shared embedding layer and 2 CRF layers, we compare results of variants in Table 3. We adopt single channel LSTM-CRF

Table 3. Results of variants. 'S' stands for source domain training data. 'T' stands for target domain training data. 'S+T' means using merged data for training. 'random' means using randomly initialized embeddings. '3random' means using 3 different randomly initialized embeddings. '1news' '1weibo' '1wiki' '1random' respectively means a shared pretrained embeddings using news, weibo, wikipedia texts and a shared randomly initialized embeddings. '3embs' means using 3 pretrained embeddings together. '2CRF' stands for 2 domain-specific CRF decoder and '1CRF' stands for a shared CRF decoder. '2 channels' removes general LSTM.

ID	Models	Precision	Recall	F1
1	Single channel (T, random)	55.90	50.00	52.78
2	Single channel (S+T, random)	55.25	45.87	50.13
3	Multichannel (S+T, 3random, 1CRF)	64.50	50.00	56.33
4	Multichannel (S+T, 3embs, 1CRF)	58.25	54.30	56.21
5	Multichannel (S+T, 1news, 1CRF)	58.82	54.30	56.47
6	Multichannel (S+T, 1weibo, 1CRF)	57.14	48.87	52.68
7	Multichannel (S+T, 1wiki, 1CRF)	58.25	52.94	55.19
8	Multichannel (S+T, 1random, 1CRF)	66.48	54.75	60.05
9	Multichannel (S+T, 1random, 2CRF)	65.45	**56.56**	**60.68**
10	2 channels (S+T, 1random, 1CRF)	**71.14**	47.96	57.30

described in Fig. 1 as baselines (#1)(#2). (#1) only uses source domain training data and (#2) uses mixed source and target data. We can find that using mixed training data without distinction leads to the model bias to large source domain and the model labels less target domain entities in test data with a decline of recall.

(#3) and (#4) use different initialized character embeddings with the same multichannel architecture. Pretrained embeddings have important effect on single domain NER [9,20], but (#4) shows no improvement in overall F1. One possible reason is that three pretrained embeddings are not in the same vector space. Another important reason we have verified is that target channel parameters are not trained enough due to the imbalance source/target ratio. We can measure the changes of a character vector before (randomly initialized) and after training (fine tuned) using cosine similarity. We respectively compute the sum of cosine similarity of top 10 characters appeared with a tag of 'B-PER', which means Chinese family name, in source domain and target domain in (#3). The bigger cosine similarity means the less changes compared to randomly initialized vector. We can see target domain cosine similarity sum is much greater than source domain in Table 4. We find target parameters are not updated enough times because there is such few target domain training data.

Nearly all the multichannel architectures ((#3)∼(#9)) perform better than single channel LSTM-CRF model. With general LSTM capturing general patterns on both source and target data, these multichannel architectures get better recall rate which is the main bottleneck that limits NER performance in

Table 4. Top 10 family name character embeddings changes before and after training in form of cosine similarity sum.

	Cosine Similarity Sum
Top 10 in Source	0.714075
Top 10 in Target	5.229697

Chinese social media. Through comparing (#8)(#9) to (#5)∼(#7), we find that randomly initialization gets better results than using pretrained embeddings as initialization. It is different from single domain supervised NER because it's hard to define a domain corpora to pretrain embeddings, especially when the pretrained embeddings are shared to three different domain LSTMs. But we can learn the shared embeddings with random initialization instead of pretraining one. (#6) has poor performance because we directly use the Weibo embeddings provided in [26]. It uses different training parameters. After using shared embedding layer which is randomly initialized in (#8)(#9), we achieve good results on F1 with improvement on both precision and recall. (#9) uses domain-specific CRF decoders which gain more improvement on recall.

We remove general LSTM in (#8) and thus get a 2 channels architecture (#10). The dramatic fall of recall rate in (#10) shows the importance of general LSTM on improving model's generalization ability.

Our multichannel LSTM-CRF (#9) exceeds overall F1 by +10.55 than baseline (#2). Multichannel variants gain widespread better recall rate.

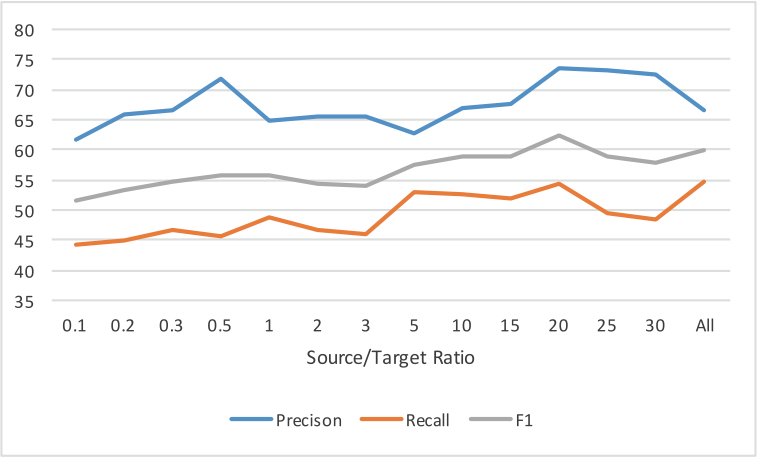

Fig. 3. Results of different source/target ratios.

Effects of Ratios: Source domain data are more than 30 times of target domain data in size. It's important to know the effects of ratios. We keep the size of target

data unchanged and gradually increase the size of source data. We compare results with different amount source/target rations in Fig. 3. We use multichannel LSTM model with shared embeddings and CRF layer in this experiment. We can see that F1 goes up in overall along with more source training data adding in, then F1 reaches 62.50% as maximum when source/target ratio equals 20. We notice that the maximum ratio is nearly 35 and results may be different if more source data are used.

Table 5. Results compared to other works. * indicates results using out-of-domain annotated data.

Models	Precision	Recall	F1
[26] Peng2015	**74.78**	39.81	51.96
[27] Peng2016	66.67	47.22	55.28
[15] He2016	66.93	40.67	50.60
[16] He2017*	61.68	48.82	54.50
Our Model*	65.45	**56.56**	**60.68**

Compared with Other Works: Table 5 shows results compared with other works[4]. [15,26,27] just use in-domain annotated data for training and large unlabeled Weibo messages for pretraining. Except out-of-domain annotated data, [16] also use unlabeled Weibo messages for semi-supervised learning. We can see that we get the best result which owns a good recall rate due to the use of general LSTM and domain-specific CRF decoders. Our model exceeds previous best recall by +7.74 and previous best overall F1 by +5.4. We achieve state-of-the-art performance on NER in Chinese social media with significant improvement.

5 Conclusion

In this paper, we have proposed a multichannel LSTM-CRF neural model using out-of-domain annotated data for NER in Chinese social media. Three LSTM channels sharing the same Chinese character embedding are designed respectively to capture general, in-domain and out-of-domain patterns. The experiments demonstrate that our model achieves significantly better performance compared to the previous state-of-the-art methods on NER in Chinese social media. Through deep analysis, we find that different channels sharing the same character embedding is important for performance improvement. And randomly initialized embeddings perform better than the pretrained ones for multichannel architecture. Domain-specific CRF decoders also help to improve recall rate.

Acknowledgments. The research work has been supported by the Natural Science Foundation of China under Grant No. 61403379 and No. 61402478.

[4] [26,27] update their results here http://www.cs.jhu.edu/~npeng/papers/golden_horse_supplement.pdf.

References

1. Blitzer, J., McDonald, R., Pereira, F.: Domain adaptation with structural correspondence learning. In: Proceedings of the 2006 conference on empirical methods in natural language processing, pp. 120–128. Association for Computational Linguistics (2006)
2. Chang, C.Y., Teng, Z., Zhang, Y.: Expectation-regulated neural model for event mention extraction. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 400–410. Association for Computational Linguistics, San Diego, California, June 2016
3. Chen, Y., Zong, C., Su, K.Y.: On jointly recognizing and aligning bilingual named entities. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 631–639. Association for Computational Linguistics (2010)
4. Cherry, C., Guo, H.: The unreasonable effectiveness of word representations for twitter named entity recognition. In: HLT-NAACL, pp. 735–745 (2015)
5. Chiu, J.P., Nichols, E.: Named entity recognition with bidirectional lstm-cnns. arXiv preprint (2015). arXiv:1511.08308
6. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. **12**(Aug), 2493–2537 (2011)
7. Daumé III., H.: Frustratingly easy domain adaptation. arXiv preprint (2009). arXiv:0907.1815
8. Daumé III., H., Kumar, A., Saha, A.: Frustratingly easy semi-supervised domain adaptation. In: Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing, pp. 53–59. Association for Computational Linguistics (2010)
9. Dong, C., Zhang, J., Zong, C., Hattori, M., Di, H.: Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition. In: Lin, C.-Y., Xue, N., Zhao, D., Huang, X., Feng, Y. (eds.) ICCPOL/NLPCC -2016. LNCS, vol. 10102, pp. 239–250. Springer, Cham (2016). doi:10.1007/978-3-319-50496-4_20
10. Dredze, M., McNamee, P., Rao, D., Gerber, A., Finin, T.: Entity disambiguation for knowledge base population. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 277–285. Association for Computational Linguistics (2010)
11. Dyer, C., Ballesteros, M., Ling, W., Matthews, A., Smith, N.A.: Transition-based dependency parsing with stack long short-term memory. arXiv preprint (2015). arXiv:1505.08075
12. Fu, G., Luke, K.K.: Chinese named entity recognition using lexicalized hmms. ACM SIGKDD Explor. Newslett. **7**(1), 19–25 (2005)
13. Gottipati, S., Jiang, J.: Linking entities to a knowledge base with query expansion. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 804–813. Association for Computational Linguistics (2011)
14. Han, A.L.-F., Wong, D.F., Chao, L.S.: Chinese Named Entity Recognition with Conditional Random Fields in the Light of Chinese Characteristics. In: Kłopotek, M.A., Koronacki, J., Marciniak, M., Mykowiecka, A., Wierzchoń, S.T. (eds.) IIS 2013. LNCS, vol. 7912, pp. 57–68. Springer, Heidelberg (2013). doi:10.1007/978-3-642-38634-3_8
15. He, H., Sun, X.: F-score driven max margin neural network for named entity recognition in chinese social media. arXiv preprint (2016). arXiv:1611.04234

16. He, H., Sun, X.: A unified model for cross-domain and semi-supervised named entity recognition in chinese social media. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
17. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint (2012). arXiv:1207.0580
18. Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. arXiv preprint (2015). arXiv:1508.01991
19. Kim, Y.B., Stratos, K., Sarikaya, R.: Frustratingly easy neural domain adaptation. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. The COLING 2016 Organizing Committee, Osaka, Japan, pp. 387–396, December 2016
20. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint (2016). arXiv:1603.01360
21. Levow, G.A.: The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 108–117 (2006)
22. Li, L., Mao, T., Huang, D., Yang, Y.: Hybrid models for chinese named entity recognition. In: COLING• ACL 2006, p. 72 (2006)
23. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional lstm-cnns-crf. arXiv preprint (2016). arXiv:1603.01354
24. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp. 3111–3119 (2013)
25. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)
26. Peng, N., Dredze, M.: Named entity recognition for chinese social media with jointly trained embeddings. In: EMNLP, pp. 548–554 (2015)
27. Peng, N., Dredze, M.: Improving named entity recognition for chinese social media with word segmentation representation learning. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 2, pp. 149–155 (2016)
28. Peng, N., Dredze, M.: Multi-task multi-domain representation learning for sequence tagging. arXiv preprint (2016). arXiv:1608.02689
29. Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning, pp. 147–155. Association for Computational Linguistics (2009)
30. Rehurek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Citeseer (2010)
31. Ritter, A., Clark, S., Etzioni, O., et al.: Named entity recognition in tweets: an experimental study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1524–1534. Association for Computational Linguistics (2011)
32. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. J. Big Data **3**(1), 1–40 (2016)
33. Yang, Z., Salakhutdinov, R., Cohen, W.W.: Transfer learning for sequence tagging with hierarchical recurrent networks. arXiv preprint (2017). arXiv:1703.06345