Recognizing Complex Entity Mentions: A Review and Future Directions

Xiang Dai
CSIRO Data61 and School of IT, University of Sydney
Sydney, Australia
dai.dai@csiro.au

Abstract

Standard named entity recognizers can effectively recognize entity mentions that consist of contiguous tokens and do not overlap with each other. However, in practice, there are many domains, such as the biomedical domain, in which there are nested, overlapping, and discontinuous entity mentions. These complex mentions cannot be directly recognized by conventional sequence tagging models because they may break the assumptions based on which sequence tagging techniques are built. We review the existing methods which are revised to tackle complex entity mentions and categorize them as token-level and sentence-level approaches. We then identify the research gap, and discuss some directions that we are exploring.

1 Introduction

Named entity recognition (NER), the task of identifying and classifying named entities (NE) within text, has received substantial attention. This is largely due to its crucial role in conducting several downstream tasks, such as entity linking (Limsopatham and Collier, 2016; Pan et al., 2017), relation extraction (Zeng et al., 2014), question answering (Mollá et al., 2007) and knowledge base construction (Zhang, 2015).

Traditionally, the NER problem can be defined as: given a sequence of tokens, output a list of tuples $< I_s, I_e, t >$, each of which is a NE mention in text. Here, I_s and I_e are the starting and ending index of the NE mention, respectively, and t is the type of the entity from a pre-defined category scheme. There are two assumptions associated with this perspective:

1. An NE mention consists of contiguous tokens, where all the tokens indexed between I_s and I_e are part of the mention; and,

2. These linear spans do not overlap with each other. In other words, no token in the text can belong to more than one NE mention.

Based on these two assumptions, the most common approach to NER is to use sequence tagging techniques with a BIO or BIOLU label set. Each token is assigned with a tag which is usually composed of a position indicator and an entity type. The position indicator is used to represent the token's role in a NE mention. In the BIOLU schema, B stands for the beginning of a mention, I for the intermediate of a mention, O for outside a mention, L for the last token of a mention, and U for a mention having only one token (Ratinov and Roth, 2009).

Sequential tagging models, such as linear-chain CRFs and BiLSTM-CRF, have achieved start-of-the-art effectiveness in many NER data sets (Lample et al., 2016; Ma and Hovy, 2016; Chiu and Nichols, 2016), since most training data sets are also annotated based on these two assumptions.

However, in practice, there are many domains, such as the biomedical domain, which involve nested, overlapping, discontinuous NE mentions that break the two assumptions mentioned above. We categorize these mentions as *complex entity mentions*, and note that standard tagging techniques cannot be applied directly to recognize these mentions (Muis and Lu, 2016; Dai et al., 2017). In the following paragraphs, we explain these complex entity mentions in details.

Nested NE mentions One NE mention is completely contained by the other. We call both of the mentions involved as nested entity mentions. Figure 1a is an example taken from the GENIA cor-

Proceedings of ACL 2018, Student Research Workshop, pages 37–44
Melbourne, Australia, July 15 - 20, 2018.

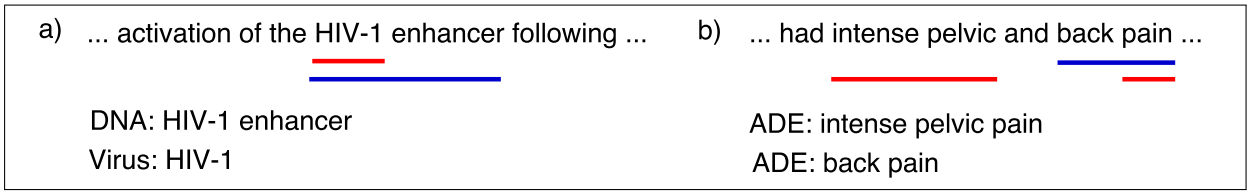

Figure 1: Examples involving overlapping, discontinuous and nested NE mentions. In (a), '*HIV-1 enhancer*' and '*HIV-1*' are nested NE mentions. In (b), '*intense pelvic pain*' and '*back pain*' overlap, meanwhile, '*intense pelvic pain*' is a discontinuous mention.

pus (Kim et al., 2003). Here, '*HIV-1 enhancer*' is a DNA mention, and it contains another mention '*HIV-1*', which is a virus.

Multi-type NE mentions An extreme case of nested NE mentions is one on which an NE mention has multiple entity types. For example, in the EPPI corpus (Alex et al., 2007), proteins can also be annotated as drug/compound, indicating that the protein is used as a drug to affect the function of a cell. Such a mention should be classified as both protein and drug/compound. In this case, we consider this mention as two mentions of different types, and these two mentions contain each other.

Overlapping NE mentions Two NE mentions overlap, but no one is completely contained by the other. Figure 1b is an example taken from the CADEC corpus (Karimi et al., 2015), which is annotated for adverse drug events (ADE) and relevant concepts. In this example, two ADEs: '*intense pelvic pain*' and '*back pain*', share a common token '*pain*', and neither is contained by the other.

Discontinuous NE mentions The mention consists of discontiguous tokens. In other words, the mention contains at least one gap. In Figure 1b, '*intense pelvic pain*' is a discontinuous NE mention since it is interrupted by '*and back*'.

These complex NE mentions can hold very useful information for downstream tasks. Sometimes, the nested and overlapping structure itself are already good indicators of the relationship between different entities involved. For example, an ORG mention '*University of Sydney*' contains a LOC mention '*Sydney*'. This structure has implied the location of the organization, and recognition of these mentions can potentially speed up the construction of a knowledge base. In addition, such entities often have fixed representations in different languages. Therefore, recognizing NE mentions, especially these discontinuous NE mentions, can improve the performance of a machine translation system (Klementiev and Roth, 2006). Furthermore, we notice that similar complex structures also exist in other NLP tasks, such as multi-word expressions recognition (Baldwin and Kim, 2010). The ideas proposed for a NER task can thus be applied to tackle similar difficulties in other tasks.

Below, we briefly review existing methods to recognize complex mentions and discuss their strengths and limitations. We also discuss the research directions we are exploring to address the research gaps.

2 Token-level Approach

Sequence tagging techniques take the representation of each token as input and output a label for each token. These local decisions are chained together to perform joint inference. Figure 2 is an illustration of a linear-chain CRF model where the tag of one token depends on both the features of that token in context and the tag of the previous token. The tag sequence predicted by the tagger is finally decoded into NE mentions using explicit rules. Here, the intermediate outputs for each token are usually BIO tags in standard NER tasks. However, since the BIO tags cannot effectively represent complex NE mentions, a natural direction is to expand the BIO tag set so that different kinds of complex entity mentions can be captured. We categorize the methods based on conventional sequence tagging as token-level approach.

Metke-Jimenez and Karimi (2015) introduced a BIO variant schema to represent discontinuous and overlapping NE mentions. Concretely, in addition to the BIO tags, four new position indicators, BD, ID, BH, and IH are proposed to denote **B**eginning of **D**iscontinuous body, **I**nside

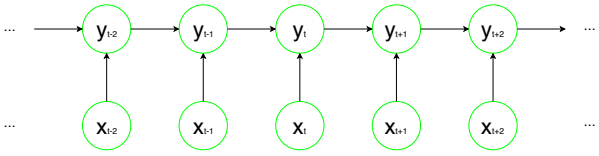

Figure 2: In a linear-chain CRF model, the output for each token depends on the features of that token in context and the output for the previous token.

had	intense	pelvic	and	back	pain	.
O	BD	ID	O	B	BH	O

Figure 3: An encoding example of two NE mentions: '*intense pelvic pain*' and '*back pain*'. Here, we keep only the position indicator and remove the entity type, since this schema can only represent overlapping mentions of the same entity type.

of **D**iscontinuous body, **B**eginning of **H**ead, and **I**nside of **H**ead. Here, the word sequences which are shared by multiple mentions are called head, and the remaining parts of the discontinuous mention are called body. Figure 3 is an encoding example using this schema. '*pain*' is the beginning of the head that is shared by two mentions, and therefore tagged as *BH*. '*intense pelvic*' is the body of a discontinuous mention, while '*back*' is the beginning of a continuous mention. We note that, even in this simple example, it is still impossible to represent several discontinuous mentions unambiguously. For example, this encoding can also be decoded as having three mentions: '*intense pelvic pain*', '*back pain*' and '*pain*'. Muis and Lu (2016) introduced the notion of *model ambiguity* and theoretically demonstrated that the models based on BIO variants usually have high ambiguity level, and therefore low precision in practice. Another limitation of this schema is that it supports only overlapping mentions of the same entity type.

Schneider et al. (2014) also proposed several BIO schema variants to encode multiword expressions with gaps and nested structure. They include two strict restrictions which are motivated linguistically in their work:

1. An expression can be completely contained within another expression, but no overlapping is allowed; and,
2. A contained expression cannot contain other expressions. In other words, the nested structure has maximum two levels.

We note that these strict restrictions cannot be applied directly on our NER tasks.

Alex et al. (2007) proposed three approaches based on a maximum entropy model (Curran and Clark, 2003) to deal with nested NE mentions:

Layering The tagger first identifies the innermost (or outermost) mentions, then the following taggers are used to identify increasingly next level mentions. Finally, the output of the taggers on different layers is combined by taking the union.

Joined labeling Each word is assigned a tag by concatenating the tags of all levels of nesting. Then a tagger is trained on the data containing the joined labels. During inference, the joined labels are decoded into their original BIO format for each entity type.

Cascade Separate models are trained for each entity type or by grouping several entity types without nested structures. Similar to the layering approach, the latter models can utilize the outputs from previous models as input features. Despite the difficulty of ordering and grouping entity models and the fact that this approach cannot deal with nested mentions of the same entity type, the cascade approach still achieves the best results among these three approaches.

Byrne (2007) and Xu et al. (2017) used a similar approach to deal with nested NE mentions. They concatenated adjacent tokens (up to a certain length) into potential mention spans. Then these spans, together with their left and right contexts, are fed into a classifier (a maximum entropy tagger in (Byrne, 2007) and a feedforward neural network in (Xu et al., 2017)). The classifier is trained to first predict whether the span is a valid NE mention, and then its entity type if it is a NE mention.

3 Sentence-level Approach

Instead of predicting whether a specific token or several tokens belong to a NE mention and its role in the mention, some methods predict directly a combination of NE mentions within a sentence. We categorize these methods as sentence-level approach.

Bill	and	Hilary	Clinton	traveled	to	Canada	.
P	O	O	P	O	O	O	O
O	O	P	P	O	O	O	O
O	O	O	O	O	O	L	O

Figure 4: An example of sentence with three NE mentions. P(ER) and L(OC) refer to the entity types.

McDonald et al. (2005) proposed a new perspective of NER as structured multi-label classification. Instead of starting index and ending index, they represent each NE mention using the set of token positions that belong to the mention. Figure 4 is an example of this representation, with each token tagged using an I/O schema. This representation is very flexible as it allows NE mentions consisting of discontiguous tokens and does not require mentions to exclude each other. Using this representation, the NER problem is converted into the multi-label classification problem of finding up to k correct labels among all possible labels, where k is a hyper-parameter of the model. Labels can be decoded to all possible NE mentions in the sentence. They do not come from a pre-defined category but depend on the sentence being processed. McDonald et al. (2005) used large-margin online learning algorithms to train the model, so that the scores of correct labels (NE mentions) are higher than those of all other possible incorrect mentions. Another advantage of this method is that the outputs of the model are unambiguous for all kinds of complex entity mentions and easy to be decoded, although the method suffers from a $O(n^3T)$ inference algorithm, where n is the length of the sentence and T is the number of entity types.

Finkel and Manning (2009) used a discriminative constituency parser to recognize nested NE mentions. They represent each sentence as a constituency tree, where each mention corresponds to a phrase in the tree. In addition, each node needs to be annotated with its parent and grandparent labels, so that the CRF-CFG parser can learn how NE mentions nest. Ringland (2016) also explored a joint model using the Berkeley parser (Petrov et al., 2006), and showed that it performed well even without specialized NER features. However, one disadvantage of their models, as in (McDonald et al., 2005), is that their time complexity is cubic in the number of tokens in the sentence. Furthermore, the high quality parse training data, which is not always available, plays a crucial role in the success of the joint model (Li et al., 2017).

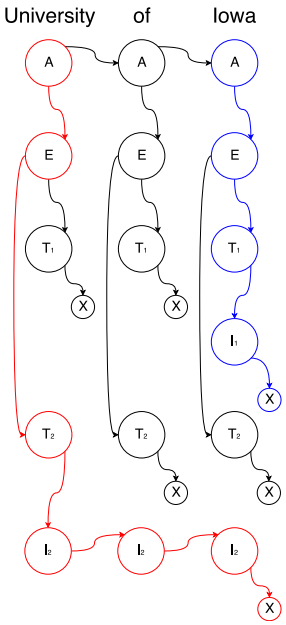

Figure 5: An example sub-hypergraph with two nested NE mentions: '*University of Iowa*' (ORG) and '*Iowa*' (LOC). Here, one mention corresponds to a path consisting of (AETI+X) nodes. Note that this hypergraph cannot be used to represent discontinuous mentions, but, in (Muis and Lu, 2016), they expand the hypergraph representation to capture discontinuous mentions through two new node types: B for within the mention, and O for part of the gap.

Lu and Roth (2015), extended by Muis and Lu (2016), proposed a novel hypergraph to compactly represent exponentially many possible nested mentions in one sentence, and one sub-hypergraph of the complete hypergraph can therefore be used to represent a combination of mentions in the sentence. Figure 5 is an example of such a sub-hypergraph, which represents two nested NE mentions.

The training objectives of these models are to maximize the log-likelihood of training instances consisting of the sentence and mention-encoded hypergraph. During inference, the model will first predict a sub-hypergraph among all possible sub-hypergraph of the complete hypergraph, and predicted mentions can be decoded from the output sub-hypergraph.

This hypergraph representation still suffers from some degree of ambiguity during decoding stage. For example, when one mention is

contained by another mention with the same entity type and their boundaries are all different, the hypergraph can be decoded in different ways. This ambiguity comes from the fact that, if one node has multiple parent nodes and multiple child nodes, there is no mechanism to decide which of the parent node is paired with which child node.

4 Research Plan

We note that the drawbacks of existing methods can be broadly categorized into: (1) lack of expressivity; and (2) computational complexity. Most of token-level approaches were proposed for some specific scenarios or data sets, therefore usually with strict restrictions. For example, the BIO variant schema in (Schneider et al., 2014) was designed for nested structure with maximum depth of two. Therefore, it is difficult to be applied on GENIA corpus (Kim et al., 2003), which contains nested entities up to four layers of embedding. In addition, these token-level methods are usually devised to deal with only either nested or discontinuous mentions, and seldom can be used to tackle all kinds of complex entity mentions simultaneously.

In contrast, sentence-level approaches are overall more flexible and less ambiguous, however with higher computational cost. For example, both Finkel and Manning (2009) and McDonald et al. (2005) methods suffer from a high time complexity which is cubic in the number of tokens in the sentence. Our aim is to propose a model that recognizes all kinds of complex entity mentions, with low ambiguity level and low computational complexity. Some specific directions include:

- The representation in (McDonald et al., 2005), introduced in Section 3, is most flexible and straightforward among all schemes designed for representing complex entity mentions. It can be used to represent all nested, overlapping and discontinuous entity mentions with unbounded length and depth. We are exploring recent advances in *multi-label classification* methods (Xu et al., 2016; Shi et al., 2017) to reduce the computational complexity of this approach.

- Sequence-to-sequence models (Sutskever et al., 2014; Cho et al., 2014) had achieved great success in machine translation and text generation tasks, especially after enhanced by attention mechanisms (Luong et al., 2015). We are exploring extending the encoder-decoder architecture to recognize complex entity mentions. During inference stage, instead of one tag sequence capturing all mentions in the sentence, the decoder can produce multiple sequences, each of which corresponds to one possible mention combination, analogous to several possible target sentences in machine translation tasks.

- Supervised learning NER methods are affected by the quantity and quality of the available annotated corpora. However, since annotating mentions with complex structure requires more human efforts than annotating only the outermost or longest continuous spans, training data for complex entity mention recognition is rare. Furthermore, the medical domain is where complex mentions widely exist, such as disorder mentions and adverse drug events. The cost of producing gold standard corpus in such a domain is very high, due to the expertise required and the limited access to some medical text, such as electronic health records.

  Active learning aims to reduce the cost of constructing a labeled dataset by allowing a human-in-the-loop (Settles and Craven, 2008; Stanovsky et al., 2017). The model selects one or several most informative instances and presents these instances to the annotators. Since only these most informative instances need to be manually annotated by human experts, it can reduce the need for human effort and therefore the cost of constructing large labeled dataset. We are exploring this method to relieve the pain of lacking training data.

- Finally, we are going to utilize recent advances in NER domain to improve the effectiveness of complex entity mentions recognizers, such as character-level embedding (Kuru et al., 2016) and joint models (Luo et al., 2015).

Besides employing active learning to create specific data set with nested, overlapping, discontinuous entity mentions, we notice that there are some off-the-shelf corpora in biomedical domain that we can use to evaluate our proposed methods, although, to our knowledge, none of these data sets

contains all three kinds of complex mentions, e.g., GENIA (Kim et al., 2003) only contains nested entity mentions, and CADEC (Karimi et al., 2015) and SemEval2014 (Pradhan et al., 2014) contain overlapping and discontinuous mentions. In addition, ACE [1] and NNE (Ringland, 2016) are newswire corpora with nested entity mentions.

On these data sets, we will use standard evaluation metrics for NER tasks, namely micro-average precision, recall and f1-score, to evaluate the effectiveness of proposed methods in recognizing both complex and simple mentions. However, due to the complexity of complex NE mentions, we will include different boundary matching relaxation, such as partial match and approximate match (Tsai et al., 2006), to measure the proposed methods in identifying these complex mentions.

5 Summary

We reviewed the existing methods of recognizing nested, overlapping and discontinuous entity mentions, categorizing them as token-level and sentence-level approaches, and discussed their strengths and limitations. We also identified the research gap and introduce some directions we are exploring.

Acknowledgments The author thanks Sarvnaz Karimi, Ben Hachey, Cecile Paris and Data61's Language and Social Computing team for their support and helpful discussions. The author also thanks three anonymous reviewers for their insightful comments.

References

Beatrice Alex, Barry Haddow, and Claire Grover. 2007. Recognising nested named entities in biomedical text. In *ACL Workshop on Biological, Translational, and Clinical Language Processing*, pages 65–72, Prague, Czech Republic.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing*, pages 267–292.

Kate Byrne. 2007. Nested named entity recognition in historical archive text. In *International Conference on Semantic Computing*, pages 589–596, Irvine, California.

Jason Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, Doha, Qatar.

James Curran and Stephen Clark. 2003. Language independent ner using a maximum entropy tagger. In *Conference on Natural Language Learning*, pages 164–167.

Xiang Dai, Sarvnaz Karimi, and Cecile Paris. 2017. Medication and adverse event extraction from noisy text. In *Australasian Language Technology Workshop*, pages 79–87, Brisbane, Australia.

Jenny Rose Finkel and Christopher D. Manning. 2009. Nested named entity recognition. In *Conference on Empirical Methods in Natural Language Processing*, pages 141–150, Suntec, Singapore.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. CADEC: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73–81.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. 2003. Genia corpusa semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180i182.

Alexandre Klementiev and Dan Roth. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics*, pages 817–824, Sydney, Australia.

Onur Kuru, Ozan Arkan Can, and Deniz Yuret. 2016. Charner: Character-level named entity recognition. In *International Conference on Computational Linguistics*, pages 911–921, Osaka, Japan.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California.

Peng-Hsuan Li, Ruo-Ping Dong, Yu-Siang Wang, Ju-Chieh Chou, and Wei-Yun Ma. 2017. Leveraging linguistic structures for named entity recognition with bidirectional recursive neural networks. In *Conference on Empirical Methods in Natural Language Processing*, pages 2654–2659, Copenhagen, Denmark.

[1] https://www.ldc.upenn.edu/collaborations/past-projects/ace

Nut Limsopatham and Nigel Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. In *Annual Meeting of the Association for Computational Linguistics*, pages 1014–1023, Berlin, Germany.

Wei Lu and Dan Roth. 2015. Joint mention extraction and classification with mention hypergraphs. In *Conference on Empirical Methods in Natural Language Processing*, pages 857–867, Lisbon, Portugal.

Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. Joint entity recognition and disambiguation. In *Conference on Empirical Methods in Natural Language Processing*, pages 879–888, Lisbon, Portugal.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Annual Meeting of the Association for Computational Linguistics*, pages 1064–1074, Berlin, Germany.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Flexible text segmentation with structured multilabel classification. In *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 987–994, Vancouver, Canada.

Alejandro Metke-Jimenez and Sarvnaz Karimi. 2015. Concept extraction to identify adverse drug reactions in medical forums: A comparison of algorithms. *CoRR abs/1504.06936*.

Diego Mollá, Menno Zaanen, and Steve Cassidy. 2007. Named entity recognition in question answering of speech data. In *Australasian Language Technology Workshop*, pages 57–65, Melbourne, Australia.

Aldrian Obaja Muis and Wei Lu. 2016. Learning to recognize discontiguous entities. In *Conference on Empirical Methods in Natural Language Processing*, pages 75–84, Austin, Texas.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Annual Meeting of the Association for Computational Linguistics*, pages 1946–1958, Vancouver, Canada.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. Semeval-2014 task 7: Analysis of clinical text. In *International Workshop on Semantic Evaluation*, pages 54–62, Dublin, Ireland.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Conference on Natural Language Learning*, pages 147–155, Boulder, Colorado.

Nicky Ringland. 2016. *Structured Named Entities*. Ph.D. thesis, University of Sydney.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the mwe gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, Honolulu, Hawaii.

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P. Xing. 2017. Towards automated icd coding using deep learning. *CoRR abs/1711.04075*.

Gabriel Stanovsky, Daniel Gruhl, and Pablo N Mendes. 2017. Recognizing mentions of adverse drug reaction in social media using knowledge-infused recurrent models. In *Conference of the European Chapter of the Association for Computational Linguistics*, pages 142–151, Valencia, Spain.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Conference on Neural Information Processing Systems*, pages 3104–3112, Montreal, Canada.

Richard Tzong-Han Tsai, Shih-Hung Wu, Wen-Chi Chou, Yu-Chun Lin, Ding He, Jieh Hsiang, Ting-Yi Sung, and Wen-Lian Hsu. 2006. Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics*, 7(1):92.

Chang Xu, Dacheng Tao, and Chao Xu. 2016. Robust extreme multi-label learning. In *Conference on Knowledge Discovery and Data Mining*, pages 1275–1284, San Francisco, California.

Mingbin Xu, Hui Jiang, and Sedtawut Watcharawittayakul. 2017. A local detection approach for named entity recognition and mention detection. In *Annual Meeting of the Association for Computational Linguistics*, pages 1237–1247, Vancouver, Canada.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *International Conference on Computational Linguistics*, pages 2335–2344, Dublin, Ireland.

Ce Zhang. 2015. *DeepDive: A Data Management System for Automatic Knowledge Base Construction*. Ph.D. thesis, University of Wisconsin Madison.