Marginal Likelihood Training of BiLSTM-CRF for Biomedical Named Entity Recognition from Disjoint Label Sets

Nathan Greenberg, Trapit Bansal, Patrick Verga, and Andrew McCallum
College of Information and Computer Sciences
University of Massachusetts Amherst
{ngreenberg, tbansal, pat, mccallum}@cs.umass.edu

Abstract

Extracting typed entity mentions from text is a fundamental component to language understanding and reasoning. While there exist substantial labeled text datasets for multiple *subsets* of biomedical entity types—such as genes and proteins, or chemicals and diseases—it is rare to find large labeled datasets containing labels for all desired entity types together. This paper presents a method for training a single CRF extractor from multiple datasets with disjoint or partially overlapping sets of entity types. Our approach employs marginal likelihood training to insist on labels that are present in the data, while filling in "missing labels". This allows us to leverage all the available data within a single model. In experimental results on the Biocreative V CDR (chemicals/diseases), Biocreative VI ChemProt (chemicals/proteins) and MedMentions (19 entity types) datasets, we show that joint training on multiple datasets improves NER F1 over training in isolation, and our methods achieve state-of-the-art results.

1 Introduction

Identifying entities in text is a vital component in language understanding, facilitating knowledge base construction (Riedel et al., 2013), question answering (Bordes et al., 2015), and search. Identifying these entities are particularly important in biomedical data. While large scale Named Entity Recognition (NER) datasets exist in news and web data (Tjong Kim Sang and De Meulder, 2003; Hovy et al., 2006), biomedical NER datasets are typically smaller and contain only one or two types per dataset. Ultimately, we would like to identify all entity types present across the union of the label sets during inference while leveraging all the available annotations to train our models.

While one may train a single model across the union of all the datasets available, this training procedure assumes that all labels (from the union of the tag set) are correctly annotated in every training instance – which is incorrect. On the other hand, training separate models on each available dataset does not take advantage of shared statistical strength from the multiple sources of information, and requires resolution of the conflicting predictions output by the different models.

To remedy these problems, we propose methods to train a joint model across the multiple tag-sets of the different datasets, sharing statistical strength by using a single feature encoder across datasets while respecting the incompleteness of the labels during training. Thus, our single model can take full advantage of all the available annotated resources and predict the full set of relevant types given a piece of text.

In experiments on three datasets, we show our methods outperform models that do not consider the incomplete annotations. We also show that jointly training on multiple datasets improves performance further and achieves state-of-the-art performance on the Biocreative V CDR dataset.

2 Model

Our models build on state-of-the-art NER systems (Lample et al., 2016) based on bi-directional Long Short Term Memory (BiLSTM) feature extractors fed into a conditional random field (CRF).

The data consists of input sequence of tokens $\mathbf{x} = \{x_1, \ldots, x_T\}$ where each token is a sequence of characters $x_t = \{c_1, \ldots, c_{K_t}\}$. The output consists of labels for each token in the sequence $\mathbf{y} = \{y_1, \ldots, y_T\}$. Labeling is done using the BILOU tagging scheme, following previous observations that it outperforms the BIO tagging scheme (Ratinov and Roth, 2009). We have D such datasets of input tokens and output labels.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2824–2829
Brussels, Belgium, October 31 - November 4, 2018.

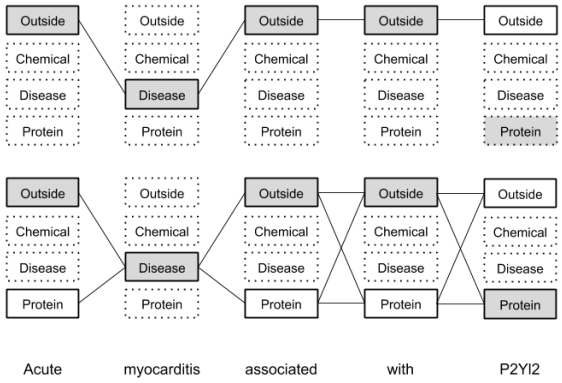

Figure 1: Training example where one label set contains Chemical/Protein and the other contains Chemical/Disease. Here Chemical and Disease annotations are given and Outside is ambiguous. Tokens labeled as Outside could potentially be either Outside or Protein (top). The shaded labels are the gold labels. The EM CRF marginalizes over all potential sequences (bot).

2.1 Feature Encoder BiLSTM

Our model takes a sequence of tokens from a single abstract as input. Tokens are generated using byte-pair encodings (BPE) (Gage, 1994; Sennrich et al., 2016), which have recently been shown to be effective for tokenization of biological texts by addressing the issue of rare or out-of-vocabulary tokens (Verga et al., 2018). BPE starts from white space tokenization and breaks down the tokens further. Because all of the evaluations are on the span level rather than the token level, the use of BPE does not impact any numerical performance. Each token t produced from BPE is mapped to a d dimensional word embedding w.

Character level features have been shown to improve NER accuracy (Lafferty et al., 2001; Lample et al., 2016; Passos et al., 2014). We encode characters in a word using another BiLSTM, similar to Lample et al. (2016), and obtain a character based embedding for every word by concatenating the last hidden state of the forward and backward character LSTM. We concatenate this character based embedding with the d-dimensional word embedding and input it to the word-level BiLSTM. This feature representation is then projected to the label dimension L using a linear layer, giving a matrix of scores $[f_{il}]$ where f_{il} is the score for predicting label $l \in [L]$ for token $i \in [T]$.

2.2 Conditional Random Field (CRF)

BiLSTM-CRF models used for named entity recognition add a CRF layer (Lafferty et al., 2001) on the output representations from the BiLSTM model described. The CRF layer scores all possible labelings to give a probability of the correct label sequence under the model. Given an input sequence of tokens $\mathbf{x} = \{x_1, \ldots, x_T\}$ and the output matrix of scores $[f_{il}]$, the score for an output labeling $\mathbf{y} = \{y_1, \ldots, y_T\}$ is given by: $s(\mathbf{x}, \mathbf{y}) = \sum_{t=1}^{T} \left(A_{y_{t-1}, y_t} + f_{t, y_t}\right)$, where A is an $L \times L$ matrix of parameters for transitioning between output labels. The CRF then generates the likelihood for the correct labeling by normalizing this score over all possible output labelings:

$$\log P(\mathbf{y}|\mathbf{x}) = s(\mathbf{x}, \mathbf{y}) - \operatorname*{logsumexp}_{y'} s(\mathbf{x}, \mathbf{y}') \quad (1)$$

The log normalization term here is:
$\operatorname*{logsumexp}_{y'} s(\mathbf{x}, \mathbf{y}') = \log \sum_{y'} \exp s(\mathbf{x}, \mathbf{y}')$
where the sum goes over all possible labelings y' of the sequence and is computed efficiently using dynamic programming (Lafferty et al., 2001).

2.3 Tagging Multiple Datasets

One way to tag multiple datasets is to concatenate all the datasets with all the output labels and train a single BiLSTM-CRF model. However, this assumes that each text snippet is completely annotated across the label sets, which is not true. We now discuss two models which do not make this assumption.

2.3.1 Multiple CRFs

We first propose one simple method to get around the assumption of complete annotation – train separate CRFs for the label set of each dataset. In particular, to share statistical strengths on the input tokens, we share the BiLSTM feature encoder across the datasets but use separate CRF layers for each of the datasets. This is a multi-task learning model (Caruana, 1998) and is expected to perform better than the naive model as it no longer makes the strict assumption of complete annotation (by using separate CRFs), and shares statistical strength across datasets. However, given a new abstract to tag, this model will generate multiple possible labelings from the different CRFs. Moreover, the labelings output by the different CRFs may be inconsistent, and how to combine these multiple labelings is not obvious. We propose and evaluate a simple heuristic procedure for merging the outputs of the different CRF predictions. Whenever the different CRF predictions disagree on a span

of tokens, we choose the prediction from the CRF that has higher marginal probability of predicting that span of tokens (Alg. 1 in supplementary).

2.3.2 EM Marginal CRF

We also propose an alternative principled approach that does not require a heuristic merging process. In order to label D datasets with some disjoint labels, we only consider the probability of the "observed labels" and allow the "unobserved" tokens to be free. Thus, when tagging dataset $i \in [D]$, we treat the non-entity tokens as potentially taking any entity type label from any of the other datasets as well as the 'O' label.

For a particular input $\mathbf{x}$ of length T from a dataset $i \in [D]$ with label set S_i, let $\mathbf{y}$ be the gold output label. Let $E \subset [T]$ be the index of tokens with any entity type label in S_i and $N \subset [T]$ be the index of tokens with 'O' label, and let $\mathbf{y}_E$ be the output sequence corresponding to indices in E, and similarly $\mathbf{y}_N$ be the output sequence for indices in N. Then, from (1), we get the likelihood $P_i(\mathbf{y_E} \cup \mathbf{y_N}|\mathbf{x})$, and a naive CRF trained on the concatenation of all the data will maximize this probability. However, since we cannot make the complete annotation assumption, we should instead maximize only the marginal probability of the observed entities on the dataset i, $P_i(\mathbf{y_E}|\mathbf{x})$, allowing $\mathbf{y_N}$ to take any values from the labels of the *other* datasets: $\cup_{j\neq i}^{D} S_j$. Thus,

$$\log P_i(\mathbf{y_E}|\mathbf{x}) = \log \sum_{\mathbf{y_N} \in \cup_{j\neq i} S_j} P_i(\mathbf{y_E}, \mathbf{y_N}|\mathbf{x})$$

$$\log P_i(\mathbf{y_E}|\mathbf{x}) = \underset{\mathbf{y_N} \in \cup_{j\neq i} S_j}{\text{logsumexp}}\, s(\mathbf{x}, \mathbf{y}_E, \mathbf{y}_N) - \log Z$$

where $\log Z$ is the log normalization term which is the same as in (1). Note that since the normalization term is the same here as for a standard CRF, we can still use the same dynamic programming algorithm as for a regular CRF to compute this $\log Z$. Now, in order to compute the first term, we note that it is similar to the computation required to compute $\log Z$ – whereas $\log Z$ is obtained by summing over all possible output sequences, this term is obtained by summing over all possible output sequences which have indices in E fixed to the correct label and indices in N taking values from $\cup_{j\neq i} S_j$. Thus, this can be computed using the same dynamic programming algorithm (Tsuboi et al., 2008), and the implementation of training this model is compatible with modern automatic differentiation libraries.

3 Experimental Results

We perform experiments on two benchmark Biocreative datasets as well as the recently introduced MedMentions data (Murty et al., 2018). Our experiments consider three types of models. The **single CRF** model naively concatenates all training datasets together and assumes complete labeling, **multi CRF** has a single Bi-LSTM feature encoder with a separate CRF for each dataset (Section 2.3.1), and **EM CRF** has a single feature encoder and a single CRF trained with EM marginalization (Section 2.3.2). For full dataset statistics and specific implementation details see supplementary material.

3.1 Biocreative V / VI

Biocreative V Chemical Disease Relation (CDR): consists of 1,500 titles and abstracts from PubMed, human annotated with chemical and disease mentions (Li et al., 2016), and has been used in previous NER evluations (Fries et al., 2017; Leaman and Lu, 2016). **Biocreative VI ChemProt (CP)**: consists of 2,432 PubMed titles and abstracts, and contains human annotated mentions of both chemicals and proteins (Krallinger et al., 2017)[1].

Our results are shown in Table 1. The top portion of the table shows models trained on single datasets, and the bottom portion shows models trained on both CDR and CP. Comparing the top and bottom portions of the table, we can see that models trained on both CP and CDR outperform training on either in isolation. Further, we see in the bottom section that our EM CRF outperforms the single CRF model and is generally better than the multi CRF model.

3.2 Adding Additional Data

Weakly Labeled data The addition of weakly labeled data has been used recently to improve the performance of relation extraction systems (Peng et al., 2016; Verga et al., 2018). In these approaches, titles and abstracts from PubMed are annotated using Pubtator, a state of the art entity tagging and linking/normalization system (Wei et al., 2013). We use the same weakly labeled data from Verga et al. (2018).

Results when adding in the additional weakly labeled data is shown in Table 2. Our models

[1]To the best of our knowledge, there is no benchmark result for this dataset

	CDR						ChemProt					
	Chemical			Disease			Chemical			Protein		
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CDR single CRF	91.8	84.7	88.1	80.7	75.0	77.8	-	-	-	-	-	-
CP single CRF	-	-	-	-	-	-	85.9	84.2	85.0	83.3	81.3	82.3
CDR+CP												
single CRF	93.3	91.9	92.6	82.0	75.5	78.6	87.8	86.8	**87.3**	82.9	83.2	83.0
multi CRF	94.1	91.8	**92.9**	82.7	76.8	**79.6**	84.8	88.4	86.6	83.7	81.6	82.6
EM CRF	94.0	91.8	**92.9**	81.1	77.7	79.4	87.1	87.6	**87.3**	83.2	83.6	**83.4**

Table 1: Precision, recall, and F1 for Biocreative V CDR and Biocreative VI ChemProt(CP) Datasets. The top portion of the table shows models trained on single datasets, the bottom portion trains on both CDR and CP, and the bottom portion trains on CDR and CP. Highest F1 scores in each section are bolded.

	CDR						ChemProt					
	Chemical			Disease			Chemical			Protein		
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1
TaggerOne	92.4	84.7	88.4	83.1	76.4	79.6	-	-	-	-	-	-
TaggerOne †	94.2	88.8	91.4	85.2	80.2	82.6	-	-	-	-	-	-
WLD single CRF	97.5	85.2	91.0	84.4	83.0	83.7	86.3	77.5	81.6	80.0	63.7	70.9
CDR+CP+WLD												
single CRF	95.7	92.4	94.0	84.5	82.9	83.7	87.6	87.2	87.4	82.0	84.7	83.3
multi CRF	95.6	93.2	**94.4**	85.7	84.0	**84.8**	85.6	90.0	87.7	87.3	81.8	84.5
EM CRF	96.6	92.1	94.3	84.9	83.6	84.2	88.9	87.5	**88.2**	84.0	86.1	**85.0**

Table 2: WLD trains with weakly labeled data. Highest F1 scores in each section are bolded. † jointly performs NER and entity linking.

improve further, outperforming the state-of-the-art TaggerOne model (Leaman and Lu, 2016).

3.3 MedMentions

MedMentions (Murty et al., 2018) is a recently introduced large dataset of PubMed abstracts containing entity linked mentions of many different semantic types. We used this data to create an artificially extreme example where two training sets contain 9 and 10 entity types each. The two type sets are fully disjoint (further details in supplementary).

In Table 3, we see that the single CRF model performs very poorly in this extreme setting due to the large amount of missing annotations. The multi CRF and EM CRF both perform well and come close to the performance of a single CRF trained on the full data, which is approximately twice as much annotated data.

Model	P	R	F1
single CRF	65.0	24.3	35.3
multi CRF	62.5	50.9	56.1
EM CRF	59.7	54.2	**56.8**
Full single CRF	60.5	58.3	59.4

Table 3: MedMentions results. Full single CRF is trained on the full set of annotations. Other models are trained on the two disjoint training sets.

4 Related Work

Until recently, feature engineered machine learning models were the highest performing approaches to NER (Ratinov and Roth, 2009; Passos et al., 2014). More recently, neural network based approaches have become state-of-the-art (Lample et al., 2016; Strubell et al., 2017; Peters et al., 2017). In BioNLP, many highest performing systems still use engineered features fed into a CRF (Wei et al., 2015; Leaman et al., 2015; Leaman and Lu, 2016). In addition to the two datasets we explored in this work, there are several other popular bio NER datasets for chemicals (Krallinger et al., 2015), species (Wang et al., 2010), diseases

(Doğan et al., 2014), and genes (Tanabe et al., 2005).

In concurrent work, Wang et al. (2018) train a model very similar to our multi-CRF model on multiple biological NER datasets with non-fully overlapping labels. Additionally, they experiment with different ways of sharing the parameters of the BiLSTM encoder. We believe this work is complementary to ours, and in many ways deals with a simpler subset of the tasks we address. Wang et al. assumes complete labeling in each of their datasets, and does not attempt to merge the final results of the multiple CRFS. On the other hand, we focus on the problem of cohesively labeling a dataset with the joint set of the different label sets, either directly through the EM model or by the merging process of the multi-CRF model.

Our method of training via marginal likelihood is the same as Tsuboi et al. (2008), who trained CRF models for Japanese word segmentation and POS tagging where only partial annotations of sentences are available. In comparison, we use the marginal likelihood training in conjunction with state-of-the art deep learning models for NER and use it to tag across multiple disjoint labels sets.

5 Conclusions and Future Work

We've introduced a method for training NER models on multiple datasets containing disjoint label sets. We show experimentally that this joint training improves performance and that our EM CRF methods outperform models using a single CRF.

One interesting problem that our models do not account for is the existence of overlapping and non-continuous entity spans. Particularly when annotating using disjoint label sets, a token could belong to multiple entity spans from different label sets. We are interested in investigating this problem in future work.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and the Center for Data Science, in part by the Chan Zuckerberg Initiative under the project Scientific Knowledge Base Construction., and in part by the National Science Foundation under Grant No. IIS-1514053. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*.

Rich Caruana. 1998. Multitask learning. In *Learning to learn*, pages 95–133. Springer.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Jason Fries, Sen Wu, Alex Ratner, and Christopher Ré. 2017. Swellshark: A generative model for biomedical named entity recognition without labeled data. *arXiv preprint arXiv:1704.06360*.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics.

Martin Krallinger, Obdulia Rabal, Saber A. Akhondi, Martn Prez Prez, Jess Santamara, Prez Gael Rodrguez, Georgios Tsatsaronis, Ander Intxaurrondo, Jos Antonio Lpez, Umesh Nandal, Erin Van Buel, Akileshwari Chandrasekhar, Marleen Rodenburg, Astrid Laegreid, Marius Doornenbal, Julen Oyarzabal, Analia Loureno, and Alfonso Valencia. 2017. Overview of the biocreative vi chemical-protein interaction track. *Proceedings of the BioCreative VI Workshop*, page 140.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(S1):S2.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Robert Leaman and Zhiyong Lu. 2016. Taggerone: joint named entity recognition and normalization with semi-markov models. *Bioinformatics*, 32(18):2839–2846.

Robert Leaman, Chih-Hsuan Wei, and Zhiyong Lu. 2015. tmchem: a high performance approach for chemical named entity recognition and normalization. *Journal of cheminformatics*, 7(1):S3.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Shikhar Murty, Patrick Verga, Luke Vilnis, Irena Radovanovic, and Andrew McCallum. 2018. Hierarchical Losses and New Resources for Fine-grained Entity Typing and Linking. In *The 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. *CoNLL-2014*, page 78.

Yifan Peng, Chih-Hsuan Wei, and Zhiyong Lu. 2016. Improving chemical disease relation extraction with rich features and weakly labeled data. *Journal of cheminformatics*, 8(1):53.

Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. 2017. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark.

Lorraine Tanabe, Natalie Xie, Lynne H Thom, Wayne Matten, and W John Wilbur. 2005. Genetag: a tagged corpus for gene/protein named entity recognition. *BMC bioinformatics*, 6(1):S3.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Yuta Tsuboi, Hisashi Kashima, Hiroki Oda, Shinsuke Mori, and Yuji Matsumoto. 2008. Training conditional random fields using incomplete annotations. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 897–904. Association for Computational Linguistics.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously Self-attending to All Mentions for Full-Abstract Biological Relation Extraction. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, New Orleans, Louisiana.

Xinglong Wang, Jun'ichi Tsujii, and Sophia Ananiadou. 2010. Disambiguating the species of biomedical named entities using natural language parsers. *Bioinformatics*, 26(5):661–667.

Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. 2018. Cross-type biomedical named entity recognition with deep multi-task learning. *arXiv preprint arXiv:1801.09851*.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2013. Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Research*, 41.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2015. Gnormplus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed research international*, 2015.