Weakly Supervised Attention Networks for Entity Recognition

Barun Patra*, Joel Ruben Antony Moniz*
{barunpatra95, jramoniz}@gmail.com

Abstract

The task of entity recognition has traditionally been modelled as a sequence labelling task. However, this usually requires a large amount of fine-grained data annotated at the token level, which in turn can be expensive and cumbersome to obtain. In this work, we aim to circumvent this requirement of word-level annotated data. To achieve this, we propose a novel architecture for entity recognition from a corpus containing weak binary presence/absence labels, which are relatively easier to obtain. We show that our proposed weakly supervised model, trained solely on a multi-label classification task, performs reasonably well on the task of entity recognition, despite not having access to any token-level ground truth data.

1 Introduction

Entity Recognition frequently finds use as a first step in numerous downstream NLP tasks (Wang and Xue, 2017; Lee et al., 2018; Liang et al., 2018). Traditionally, it has been posed as a sequence labeling task (Lample et al., 2016; Ma and Hovy, 2016), which in turn requires corpora with token-based annotations. A key drawback of this formulation, however, lies in its dependence on corpora annotated at the token-level, which can often be tedious to obtain and expensive to annotate.

One potential way of overcoming this limitation is to move towards a method that utilizes a weaker form of supervision that is easier to obtain. In this work, we focus on one such form of weak supervision: binary labels that indicate the presence of an entity type. The cognitive load of selecting whether an entity type is present or not is usually less than that of actually highlighting and annotating spans with their correct entity types. It also stands to reason that providing these binary labels might be faster. Both these properties are particularly advantageous for a human-in-the-loop setup in a user facing task, since a user is more likely to answer a yes/no question than to provide the annotated entity spans. This, in turn, facilitates cheaper and faster data collection; be it explicitly in the form of feedback questions, or implicitly from mined user logs (for example, clicked search engine results for queries related to ”movies” are likely to contain the entity in question (Xu et al., 2009)).

In this work, we make first steps towards moving away from span-based corpora, relying solely on binary presence/absence classification labels for extracting entities. We propose a novel attention-based model that, though trained on a multi-label classification task, can be used for entity recognition. We show the efficacy of our proposed model on the widely used 2003 CoNLL dataset. Our model achieves reasonable performance without having access to token-level annotations. We thus show that it is possible to extract entities using a weak classification signal. [1]

2 Related Work

Commonly used methods for entity extraction rely on token-level annotated corpora. Conventionally, these supervised methods learn a CRF (Lafferty et al., 2001) or a Seq2Seq (Sutskever et al., 2014) model over either hand-crafter or neural features. More recently, pre-trained embeddings from language models trained on large corpora (Peters et al., 2018; Devlin et al., 2018), when augmented with previous methods, have shown marked improvements.

A contrasting line of work has been to explore unsupervised entity extraction without using

[*] Equal Contribution

[1] All our code can be found at https://github.com/joelmoniz/AttentionSegmentation

*Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing
and the 9th International Joint Conference on Natural Language Processing*, pages 6268–6273,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

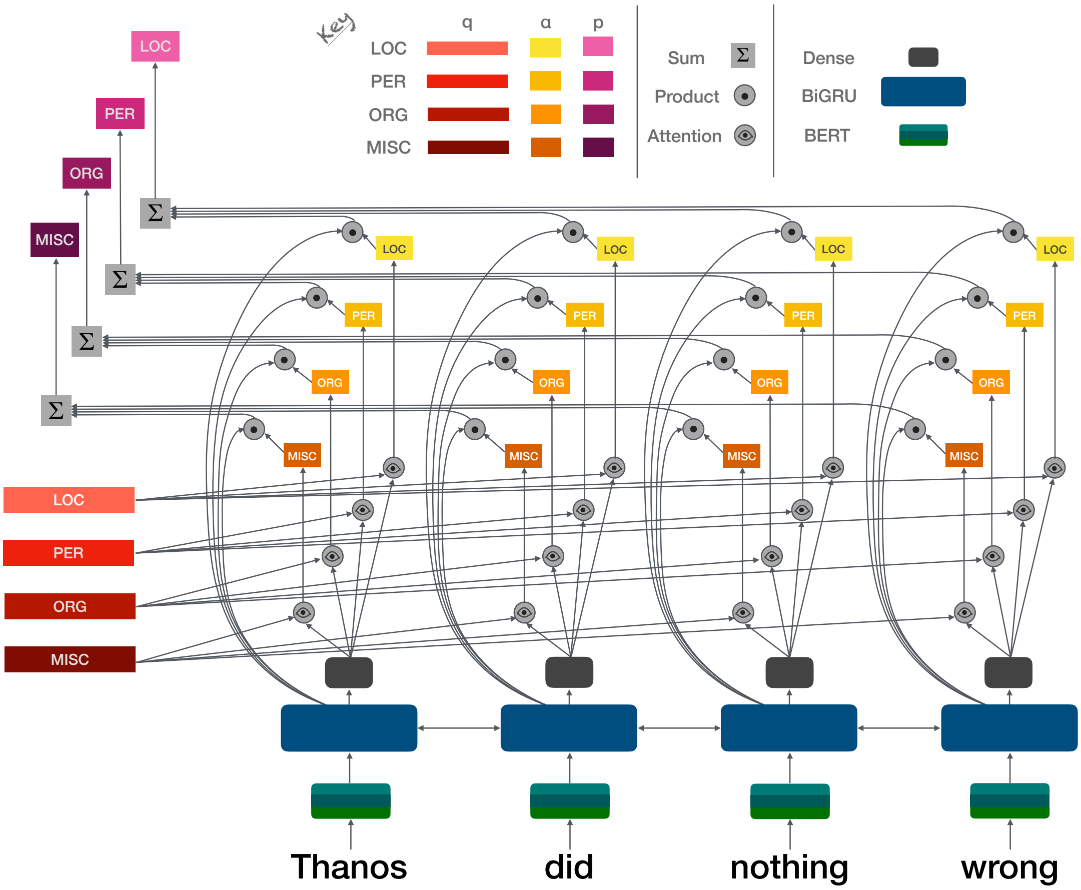

Figure 1: Model Architecture: Sentence-level Classifier. **q** represents the learned query vector per tag, α denotes the attention weights computed per token and **p** denotes the probability of predicting each tag at the sentence level.

ground-truth token or sentence level annotations. For example, a common paradigm for unsupervised Named Entity Recognition involves relying on a seed gazetteer, as in the case of Zhang and Elhadad (2013) and Ghiasvand and Kate (2015) both in the medical domain. In a more general setting, Carlson et al. (2009) use gazetteers to bootstrap training by labelling sequences that can be confidently annotated, and then use this partly labelled data to train their proposed Partial Perceptron algorithm; although they make use of a gazetteer, being the closest in setting to our proposed approach, we use Carlson et al. (2009) as our primary baseline. Another common technique involves bootstrapping the system with a set of rule templates, such as in (Etzioni et al., 2005) and (Collins and Singer, 1999). However, these methods often rely on an initial seed rule-base or on the availability of gazetteers (or the effective generation of these gazetteers using an online source such as Wikipedia).

Aside from directly improving performance on various tasks, attention (Bahdanau et al., 2014; Luong et al., 2015) has proven to be extremely useful when used indirectly in a wide variety of other ways (for example, for segmentation (Tang and Yang, 2018) and unsupervised speech-to-text alignment (Boito et al., 2017; Godard et al., 2018)). In addition, using attention-based models for object segmentation in a weakly supervised setting has been well explored in the vision domain (Teh et al., 2016; Zhang et al., 2018). Inspired by this, we leverage the attention weights of the model to identify entity spans.

3 Method

Figures 1 and 2 describe the different components of our model. Our model comprises of 4 modules: a sentence representation module, an attention module, a token-level tagger and a sentence-level classifier. Concretely, given a sentence $(x_1, \cdots x_l)$, the model predicts whether a tag $t \in \mathcal{T}$ is present in the sentence, as well as an attention distribution over the words for each tag $(\alpha_{1,t} \cdots \alpha_{l,t}) \forall t \in \mathcal{T}$ (where $\mathcal{T}$ is the set of entity tags), as described below:

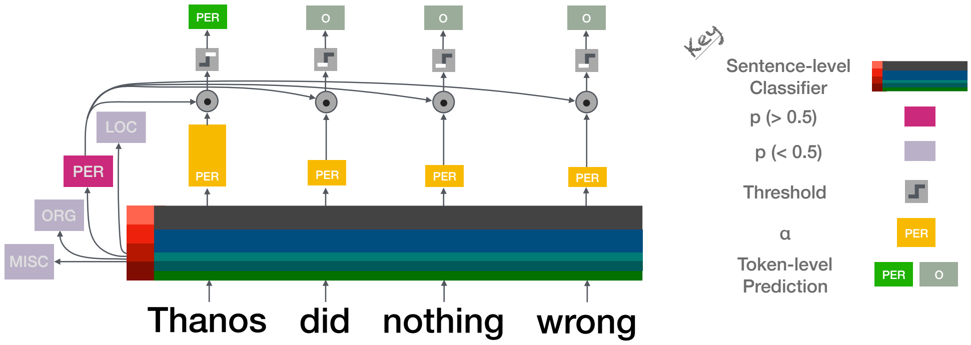

Figure 2: Model Architecture: Token Tagger. The Token Tagger performs tagging using the attention weights. α denotes the attention weights of the predicted tag(s).

Token Representation Module: generates an embedding for each token, representing the word's meaning and its left and right context. We use BERT embeddings (Devlin et al., 2018) for generating a word-level representation. We further use a Bidirectional GRU (Cho et al., 2014) layer to better adapt BERT embeddings to the task to obtain token representations $(e_1 \cdots e_l)$.

Attention Module: consists of an attention mechanism for each tag type. Given the token embeddings, a softmax distribution pertaining to the corresponding tag is generated, modelling the conditional probability of a word in a sentence being of that tag, given that the entire sentence contains the tag. We compute attention as in Bahdanau et al. (2014), using one learned query vector q_t per tag $t \in \mathcal{T}$. The token embeddings $(e_1 \cdots e_l)$ are passed through a dense layer to generate keys $(k_1 \cdots k_l)$ in the query space, which together with query q_t, yield a set of attention weights $(\alpha_{1,t} \cdots \alpha_{l,t})$.

Sentence-level Classifier: generates the probability of the presence of a tag in the sentence. For each tag, the token-level representations $(e_1 \cdots e_l)$ are weighed by the attention distribution $(\alpha_{1,t} \cdots \alpha_{l,t})$ corresponding to the tag. The weighted sum generates a sentence representation s_t per tag, which is passed through a sigmoid layer to generate the probability of the tag being present (p_t), with $p_t > 0.5$ denoting the presence of a tag.

Token Tagger Module: combines the probabilities from the Sentence-level Classifier with the attention weights obtained from the Attention Module to generate BIO tags for each token. Only the attention weights pertaining to the predicted labels $\mathcal{T}'$ are considered (i.e., if no tag is predicted, the entire sentence is marked "O"), with the attention weights being scaled by the probability of the predicted label (i.e $p_t * (\alpha_{1,t} \cdots \alpha_{l,t})$). A word x_i is assigned the label $y_i = \text{argmax}_{t \in \mathcal{T}'}(p_t * \alpha_{i,t})$ if $p_t * \alpha_{i,y_i}$ is greater than a small threshold ϵ and it is neither a punctuation symbol nor a stop-word (Figure 2).

4 Experiments

Dataset: To demonstrate the feasibility of our model, we adapt the commonly used CoNLL 2003 dataset (Ratinov and Roth, 2009). The dataset contains token-level annotations of the Reuter's Corpus (Lewis et al., 2004) for 4 entity types: person (PER), location (LOC), organization (ORG) and miscellaneous (MISC), with each token being tagged in the IOB format (Ramshaw and Marcus, 1999). For training and validation, we strip out the token-level annotations, instead annotating sentences to merely indicate the presence of entity-types. In order to quantify the quality of the extracted entities, we then measure the span level f-score on the test set using the gold entity spans.

Modelling Choices and Hyperparameters: We use the BERT-Base Multilingual Cased model for the Token Representation Module, similar to the NER tagging setup in Devlin et al. (2018). We experiment both with using the embeddings as is, and fine-tuning the top layers of the attention encoder (we only try fine-tuning the top, top-two and top-three layers due to computational constraints). For the Token Tagger module, we find that the attention probabilities usually demarcate the tagged words quite clearly, and that an ϵ value of 0.01 performs reasonably well. We use early stopping based on the (averaged) validation accuracy of the sentence-level predictions. We also observe that fine-tuning the BERT model requires learning rates comparable in order of magnitude to those used in Devlin et al. (2018), and hence use a learning rate of 2e-7 for fine-tuning the transformer layers, and 1e-3 for the rest of the network. More details related to the hyperparameters used are presented in Appendix A. All our models have been implemented using the AllenNLP framework (Gardner et al., 2017).

5 Results and Analysis

Type	Model	F-score	Acc
S	(Lample et al., 2016)	90.9	-
	(Peters et al., 2018)	92.2	-
	(Devlin et al., 2018)	92.8	-
US	(Carlson et al., 2009)	55.3	-
WS	BERT + PS	65.8	95.5
	BERT(FT) + PS	68.4	95.5
	BERT + GRU + PS	79.2	95.5
	BERT(FT) + GRU	80.7	95.6
	BERT(FT) + GRU + PS	81.1	95.6

Table 1: Different model performances on the 2003 CoNLL test set. FT denotes fine-tuning, PS denotes probability scaling. S is supervised, US is unsupervised and WS is weakly supervised

Table 1 shows the performance of our model. We observe that our proposed approach significantly outperforms the baseline, and performs reasonably well when compared to various state-of-the-art supervised approaches that use significantly more ground-truth annotation information.

To further investigate the impact of the different components of the model, we ablate our model components (Table 1). We observe that having a contextual GRU layer for adapting to the task has a significant impact on the performance, with the final model performing much better than BERT + PS. Further, fine-tuning and probability scaling also help improve model performance.

5.1 Impact of Stop-Word Removal

We find that the model learns to focus on indicator words that frequently occur in sentences where a particular tag is present, and tends to use them to identify the presence of entities at the sentence level. For example, the word ”at” is often indicative of the existence of a location in a sentence, the name of the location itself aside, and the model tends to focus on both. This behaviour is unsurprising, given that the model is trained purely using a signal of whether or not a sentence contains an entity of that type, with no idea about the entity boundaries themselves. Based on what we commonly observe, a few of these indicator words include: prepositions such as “at” and “in” when a LOC entity is present; common titles preceding PER names (such as “President” X or “Miss” Y); conjunctions that might separate two or more entities (such as X “and” Y).

This results in the model picking out spurious words alongside the actual entity, which necessitates the use of stop-word and symbol removal. However, this removal also results in the model not being able to pick out words when they occur within an entity span (for example, ”Republic of Iceland”). This is particularly problematic for ORG (4.85% spans have stop-words) and MISC (3.15% spans have stop-words), compared to LOC(0.6%) and PER(0.76%). This issue can potentially be mitigated with either a more sophisticated tagger module or a better stop-word/symbol removal mechanism, which we leave to future work.

5.2 Errors due to Incorrect Entity Boundaries

Model	Text	Type	Micro
(Peters et al., 2018)	95.9	93.8	94.8
BERT(FT) + GRU + PS	84.0	89.0	86.5

Table 2: MUC f-scores for the model

Since our model does not have access to annotated training data, it has no direct supervision for learning entity boundaries. This particularly hurts the model in the CoNLL task, since precision, recall and f-score are measured based on an exact string match. In order to investigate this, we use the metric used in MUC events (Grishman and Sundheim, 1996; Chinchor, 1998), wherein a system is scored on two axes: finding the correct text and the correct type. A text is correct if the entity boundaries are correct, regardless of their type, while a type is correct if an entity is tagged with a correct type, regardless of the boundaries, as long as there is an overlap with the gold type. The final score is a micro average f-measure (see Nadeau and Sekine (2007) for more details).

Table 2 shows the performance of the model along the two axes, as well as the MUC score. We also report the same metrics for the supervised model proposed in Peters et al. (2018). We

Idx	Sentence	Predictions	Gold
0	@0@ Neal Lancaster , Dave Barr (Canada) , Mike Sullivan , Willie	LOC PER	LOC PER
1	@0@. Fatima Yusuf (Nigeria) @0@.@0@	LOC PER	LOC PER
2	Russian peacemaker Alexander Lebed said he and rebel military leader Aslan Maskhadov agreed after overnight talks to defer the decision on whether Chechnya should be independent until December @0@ , @0@ .	LOC MISC PER	LOC MISC PER
3	" The world community should not be indifferent to the fact that President Lukashenko , who leads this European state of @0@ million people , is trying to establish a dictatorship with his new constitution , " Sharetsky said .	MISC PER	MISC PER
4	" I would like to thank the Eagles organisation and the wonderful fans of Philadelphia for supporting me throughout my career , " Cunningham said .	LOC ORG PER	ORG PER
5	June @0@-@0@ v Hampshire (three days)	LOC	ORG

Figure 3: Qualitative Analysis (Best viewed in color) PER, LOC, MISC, ORG. The Gold column shows the gold tags contained in the sentence while the Predictions column contains the tags predicted by the model. Tags in green are correctly predicted, while those in red are incorrect.

see that the model primarily loses out on the text based measure, and performs quite well on the type based one. This is in accordance with our hypothesis that the model identifies the correct entities, but fails at finding the exact entity boundaries. A better boundary detection method can consequently be used alongside our model to improve the entity retrievals in a downstream task.

5.3 Qualitative Analysis

Fig 3 shows examples of where our proposed model performs well, and where it fails. Examples 0, 1 and 2 show when the model is correctly able to identify the entity classes persons, locations and miscellaneous. Examples 3, 4 and 5 show some of the shortcomings. As described in 5.1, indicator tokens like President (Example 3) are usually marked as the PER entity type, which gets penalized when compared to the gold labels. Another common failure case we observe is when outside domain knowledge is necessary to disambiguate the entity type. For example, in Example 4, the model predicts Philadelphia as LOC, while the correct tag is ORG (referring to the Philadelphia Eagles). Similarly, in Example 5, the model classifies Hampshire as LOC, while the correct tag is ORG (the cricket club).

6 Conclusion

We present a novel method for entity recognition using a relatively weak supervision signal. Our proposed model, trained on a multi-label classification task, achieves reasonable entity recognition performance.

While our proposed method is simple, we demonstrate that it works surprisingly well. Various other formulations for this task are possible: for example, one might involve marginalizing over the tags and using this for predicting a label; another could perform attention over spans instead of tokens. We plan on investigating these alternate approaches in future work.

Acknowledgments

We would like to thank the anonymous reviewers for their invaluable feedback, which helped shaped the paper into its current form. We would also like to thank Matthew R. Gormley for the helpful discussions on the topic.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Marcely Zanon Boito, Alexandre Bérard, Aline Villavicencio, and Laurent Besacier. 2017. Unwritten languages demand attention too! word discovery with encoder-decoder models. In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*, pages 458–465. IEEE.

Andrew Carlson, Scott Gaffney, and Flavian Vasile. 2009. Learning a named entity tagger from gazetteers with the partial perceptron. In *AAAI Spring Symposium: Learning by Reading and Learning to Read*, pages 7–13.

Nancy Chinchor. 1998. Overview of muc-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*.

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 *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165(1):91–134.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Omid Ghiasvand and Rohit J Kate. 2015. Biomedical named entity recognition with less supervision. In *Healthcare Informatics (ICHI), 2015 International Conference on*, pages 495–495. IEEE.

Pierre Godard, Marcely Zanon-Boito, Lucas Ondel, Alexandre Berard, François Yvon, Aline Villavicencio, and Laurent Besacier. 2018. Unsupervised word segmentation from speech with attention. *arXiv preprint arXiv:1806.06734*.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, volume 1.

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. In *Proceedings of NAACL-HLT*, pages 260–270.

Jihwan Lee, Dongchan Kim, Ruhi Sarikaya, and Young-Bum Kim. 2018. Coupled representation learning for domains, intents and slots in spoken language understanding. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 714–719. IEEE.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

Chen Liang, Mohammad Norouzi, Jonathan Berant, Quoc V Le, and Ni Lao. 2018. Memory augmented policy optimization for program synthesis and semantic parsing. In *Advances in Neural Information Processing Systems*, pages 9994–10006.

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

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *CoNLL*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Zhiwen Tang and Grace Hui Yang. 2018. Deeptilebars: Visualizing term distribution for neural information retrieval. *arXiv preprint arXiv:1811.00606*.

Eu Wern Teh, Mrigank Rochan, and Yang Wang. 2016. Attention networks for weakly supervised object localization. In *BMVC*, pages 1–11.

Chuan Wang and Nianwen Xue. 2017. Getting the most out of amr parsing. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1257–1268.

Gu Xu, Shuang-Hong Yang, and Hang Li. 2009. Named entity mining from click-through data using weakly supervised latent dirichlet allocation. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1365–1374. ACM.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. 2018. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102.

Shaodian Zhang and Noémie Elhadad. 2013. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of biomedical informatics*, 46(6):1088–1098.