Table 5: Regexes created by C, J and A after the first 5 minutes for the DATETIME task.

Source	Regex and their properties
C	`\d{4}-\d{1,2}-\d{1,2}T\d{1,2}:\d{1,4}-\d{1,4}`
J	`([0-9][0-9][0-9][0-9]|[0-9][0-9])\/([0-9][0-9]\/[0-9][0-9])`
A	`(Monday|Tuesday|Wednesday|Thursday|Friday|Saturday|Sunday){0,1}\s*[0-9]{1,2}`

	No. of matches in S_0	EntPrec	EntRecall	EntF1
C	34	1.0	0.001	0.001
J	161,299	0.0	0.0	0.0
A	3,795,149	0.001	0.027	0.002

needed to train and fine-tune an NN. Our assumption is that the training is instantaneous. We use a standard PC with a single GeForce GTX 1080 Graphics Card in our actual experiments. For all tasks, excluding **DATETIME**, pretraining an NN on S_{wl} took in the range of 20 minutes to an hour, and it took almost 2 hours for **DATETIME**. Each round of fine-tuning on all data sets ranged from 2 to 20 minutes. Thus, it appears that the user would waste time waiting for an NN to be pretrained and fine-tuned. However, this limitation is not necessarily a fatal flaw of our study due to several reasons: (1) a user could proceed with manual labeling and regex construction while waiting for NN training, (2) a user could switch to some other task while waiting, (3) the training time could be significantly improved if it were implemented on a more powerful computer system, (4) our study did not focus on training speedups, and it is possible that with some tuning the training time could be further reduced.

7 CONCLUSIONS

We investigated the problem of entity extraction, where entities follow or closely resemble patterns described using regular expressions. Industrial strength entity recognizers for this class of entities employ regex. Regex is either manually crafted or learned. The main drawbacks of regexes are that they tend to be complex to achieve high coverage, are difficult to maintain, and are not resilient to noise, such as typos. In the wake of data deluge, deep learning algorithms are an attractive alternative, but they require large amount of human annotated data. We propose a framework that combines the advantages of regexes and deep learning, coupled with weak supervision and active learning.

We conducted extensive experiments with data from 5 application domains: email, course number, phone number, datetime, and bill date. We also conducted a user study with 4 volunteers. The experiments showed that we can build ML models that are regex-oblivious, achieve high accuracy, and are resilient to small noise. The user study provided interesting insights about the trade-offs between constructing regexes and manually labeling the unlabeled text.

ACKNOWLEDGMENTS

This work was supported in part by the following grants: U.S. NSF BigData 1546480 and U.S. NIH R21CA202130.

REFERENCES

[1] Alberto Bartoli, Giorgio Davanzo, Andrea De Lorenzo, Marco Mauri, Eric Medvet, and Enrico Sorio. 2012. Automatic generation of regular expressions from examples with genetic programming. In *GECCO*. 1477–1478.
[2] Alberto Bartoli, Giorgio Davanzo, Andrea De Lorenzo, Eric Medvet, and Enrico Sorio. 2014. Automatic synthesis of regular expressions from examples. *Computer* 47, 12 (2014), 72–80.
[3] Alberto Bartoli, Andrea De Lorenzo, Eric Medvet, and Fabiano Tarlao. 2016. Inference of regular expressions for text extraction from examples. *IEEE Transactions on Knowledge and Data Engineering* 28, 5 (2016), 1217–1230.
[4] Alberto Bartoli, Andrea De Lorenzo, Eric Medvet, and Fabiano Tarlao. 2018. Active learning of regular expressions for entity extraction. *IEEE Transactions on cybernetics* 48, 3 (2018), 1067–1080.
[5] James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *JMLR* 13, Feb (2012), 281–305.
[6] Falk Brauer, Robert Rieger, Adrian Mocan, and Wojciech M Barczynski. 2011. Enabling information extraction by inference of regular expressions from sample entities. In *CIKM*. ACM, 1285–1294.
[7] Yukun Chen, Thomas A Lasko, Qiaozhu Mei, Joshua C Denny, and Hua Xu. 2015. A study of active learning methods for named entity recognition in clinical text. *Journal of biomedical informatics* 58 (2015), 11–18.
[8] Jason PC Chiu and Eric Nichols. 2015. Named entity recognition with bidirectional LSTM-CNNs. *arXiv preprint arXiv:1511.08308* (2015).
[9] Robert A Cochran, Loris D'Antoni, Benjamin Livshits, David Molnar, and Margus Veanes. 2015. Program boosting: Program synthesis via crowd-sourcing. In *ACM SIGPLAN Notices*, Vol. 50. ACM, 677–688.
[10] François Denis. 2001. Learning regular languages from simple positive examples. *Machine Learning* 44, 1-2 (2001), 37–66.
[11] Henning Fernau. 2009. Algorithms for learning regular expressions from positive data. *Information and Computation* 207, 4 (2009), 521–541.
[12] 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* (2017).
[13] Athanasios Giannakopoulos, Claudiu Musat, Andreea Hossmann, and Michael Baeriswyl. 2017. Unsupervised aspect term extraction with b-lstm & crf using automatically labelled datasets. *arXiv preprint arXiv:1709.05094* (2017).
[14] IQAndreas. 2014. What is meant by "Now you have two problems"? `https://softwareengineering.stackexchange.com/questions/223634/what-is-meant-by-now-you-have-two-problems`
[15] Anita Krishnakumar. 2007. *Active learning literature survey.* Technical Report. Technical reports, University of California, Santa Cruz. 42.
[16] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360* (2016).
[17] Yunyao Li, Rajasekar Krishnamurthy, Sriram Raghavan, Shivakumar Vaithyanathan, and HV Jagadish. 2008. Regular expression learning for information extraction. In *EMNLP*. 21–30.
[18] Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354* (2016).
[19] Rachel Millner. 2008. Four Regular Expressions to Check Email Addresses. `https://tinyurl.com/y93bomtv`
[20] Maureen A. Murtaugh, Bryan Smith Gibson, Doug Redd, and Qing Zeng-Treitler. 2015. Regular expression-based learning to extract bodyweight values from clinical notes. *Journal of Biomedical Informatics* 54 (2015), 186 – 190.
[21] Karin Murthy, P Deepak, and Prasad M Deshpande. 2012. Improving recall of regular expressions for information extraction. In *WISE*. Springer, 455–467.
[22] Hieu T Nguyen and Arnold Smeulders. 2004. Active learning using pre-clustering. In *ICML*. ACM, 79.
[23] Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. In *Advances in neural information processing systems*. 3567–3575.
[24] Jingbo Shang, Liyuan Liu, Xiang Ren, Xiaotao Gu, Teng Ren, and Jiawei Han. 2018. Learning Named Entity Tagger using Domain-Specific Dictionary. *arXiv preprint arXiv:1809.03599* (2018).
[25] Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep Active Learning for Named Entity Recognition. *arXiv preprint arXiv:1707.05928* (2017).
[26] Stanley Simoes, P Deepak, Munu Sairamesh, Deepak Khemani, and Sameep Mehta. 2018. Content and Context: Two-Pronged Bootstrapped Learning for Regex-Formatted Entity Extraction.. In *AAAI*.
[27] Qing T. Zeng, Sergey Goryachev, Scott Weiss, Margarita Sordo, Shawn N. Murphy, and Ross Lazarus. 2006. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making* 6, 1 (26 Jul 2006), 30.
[28] Shanshan Zhang, Lihong He, Slobodan Vucetic, and Eduard Dragut. 2018. Regular Expression Guided Entity Mention Mining from Noisy Web Data. In *EMNLP*. 1991–2000.