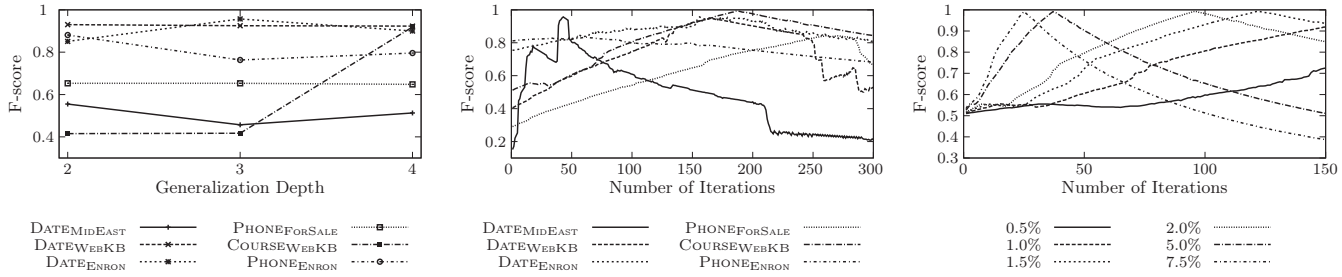

(a) (b) (c)

Figure 3: Effect of (a) generalization depth, (b) number of iterations, and (c) expansion rate on the F-score of $\mathcal{M}_{exp}$

TASK	DATE$_{\text{MIDEAST}}$	DATE$_{\text{WEBKB}}$	DATE$_{\text{ENRON}}$
r_{seed}	\d{2}/\d{2}/\d{2}	\d{2}/\d{2}/\d{2}	\d{2}/\d{2}/\d{2}
$\mathcal{R}$	\d{**1**,2}/**?**\d{**1**,2}/**?**\d{2} \d{**1**,2}/**?**\d{2}**[-.\:/*|]**\d{2,**3**} \d{2}/**?**\d{2}**\W**\d{2}	\d{**1**,2}/\d{**1**,2}/\d{2} \d{2}/\d{**1**,2}/\d{2} \d{**1**,2}/\d{2}/\d{2}	\d{**1**,2}**[-.\:/*|]**\d{**1**,2}**[-.\:/*|]**\d{2} \d{**1**,2}/\d**.**/**?**\d{2} \d{**1**,2}/\d{**1**,2}/\d{**1**,2}
TASK	PHONE$_{\text{FORSALE}}$	COURSE$_{\text{WEBKB}}$	PHONE$_{\text{ENRON}}$
r_{seed}	\(\d{3}\)\d{3}-\d{4}	CS\d{3}	\d{3}-\d{3}-\d{4}
$\mathcal{R}$	\(**?**\d{3}**\W**\d{3}**\W**\d{4} \(\d{3}\)**\W?**\d{3}**\W**\d{4} \(**?**\d{**1**,3}**\W**\d{3}-\d{4}	C**?[a-zA-Z]**{1,**2**}\d{3} **[a-zA-Z]**{1,**2**}S**?**\d{3} CS**\w**{1,**3**}\d**\w**	\d{3}**\W**{1,**2**}\d{3}**[-.\:/*|]**\d{4} \d{**1**,3}-**?**\d{3}-**?**\d{4} \d{3}**\W**{1,**2**}\d{3}-\d{**3**,4}

Table 4: Top-3 recommended regexes. The generalized units are boldfaced and underlined.

cal Methods in Natural Language Processing, EMNLP 2013, 827–832. Association for Computational Linguistics.

Gupta, S., and Manning, C. D. 2014. Improved pattern learning for bootstrapped entity extraction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, CoNLL 2014, 98–108. Association for Computational Linguistics.

Gupta, S., and Manning, C. D. 2015. Distributed representation of words to guide bootstrapped entity classifiers. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT 2015, 1215–1220. Association for Computational Linguistics.

Hosmer, Jr., D. W.; Lemeshow, S.; and Sturdivant, R. X. 2013. *Applied Logistic Regression*. Wiley Series in Probability and Statistics. Hoboken, New Jersey: John Wiley & Sons, Inc., third edition.

Jurafsky, D., and Martin, J. H. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall Series in Artificial Intelligence. Upper Saddle River, New Jersey: Prentice Hall, 1st edition.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics – Doklady* 10(8):707–710.

Li, Y.; Krishnamurthy, R.; Raghavan, S.; Vaithyanathan, S.; and Jagadish, H. V. 2008. Regular expression learning for information extraction. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2008, 21–30. Association for Computational Linguistics.

Martınez, C. 2005. Forty years of quicksort and quickselect: a personal view. In *Algorithms Seminar, 2002–2004*, 101–104. INRIA.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, NIPS 2013, 3111–3119. Curran Associates, Inc.

Murthy, K.; P., D.; and Deshpande, P. M. 2012. Improving recall of regular expressions for information extraction. In *Proceedings of the 13th International Conference on Web Information Systems Engineering*, WISE 2012, 455–467. Springer-Verlag.

Qadir, A.; Mendes, P. N.; Gruhl, D.; and Lewis, N. 2015. Semantic lexicon induction from twitter with pattern relatedness and flexible term length. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI 2015, 2432–2439. AAAI Press.

Riloff, E., and Jones, R. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and Eleventh Conference on Innovative Applications of Artificial Intelligence*, AAAI/IAAI 1999, 474–479. AAAI Press / The MIT Press.

Sarmento, L.; Jijkoun, V.; de Rijke, M.; and Oliveira, E. 2007. "more like these": Growing entity classes from seeds. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, CIKM 2007, 959–962. Association for Computing Machinery.

Schulz, K. U., and Mihov, S. 2002. Fast string correction with levenshtein automata. *International Journal on Document Analysis and Recognition* 5(1):67–85.

Thelen, M., and Riloff, E. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2002, 214–221. Association for Computational Linguistics.