Awesomeness with job ads

Clemens Westrup

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 16.1.2015

Thesis supervisor:

Michael Mathioudakis, Ph.D.

Thesis advisor:

Prof. Aristides Gionis

Aalto University
School of Science

aalto university
school of science

abstract of the
master's thesis

Author: Clemens Westrup

Title: Awesomeness with job ads

Date: 16.1.2015 Language: English Number of pages: 0+0

Department of Information and Computer Science

Professorship: Machine Learning, Data Mining, and Probabilistic Modeling

Supervisor: Michael Mathioudakis, Ph.D.

Advisor: Prof. Aristides Gionis

Your abstract in English. Try to keep the abstract short; approximately 100 words should be enough. The abstract explains your research topic, the methods you have used, and the results you obtained. Your abstract in English. Try to keep the abstract short; approximately 100 words should be enough. The abstract explains your research topic, the methods you have used, and the results you obtained.
Your abstract in English. Try to keep the abstract short; approximately 100 words should be enough. The abstract explains your research topic, the methods you have used, and the results you obtained. Your abstract in English. Try to keep the abstract short; approximately 100 words should be enough. The abstract explains your research topic, the methods you have used, and the results you obtained.

Keywords: NLP, bla bla, keyword

Preface

I want to thank bla bla bla

Otaniemi, 16.1.2015

Clemens Westrup

Contents

Symbols and abbreviations

Symbols

$\mathbf{B}$	magnetic flux density
c	speed of light in vacuum $\approx 3 \times 10^8$ [m/s]
ω_{D}	Debye frequency
ω_{latt}	average phonon frequency of lattice
$\uparrow$	electron spin direction up
$\downarrow$	electron spin direction down

Operators

$\nabla \times \mathbf{A}$	curl of vectorin $\mathbf{A}$
$\frac{\mathrm{d}}{\mathrm{d}t}$	derivative with respect to variable t
$\frac{\partial}{\partial t}$	partial derivative with respect to variable t
$\sum_i$	sum over index i
$\mathbf{A} \cdot \mathbf{B}$	dot product of vectors $\mathbf{A}$ and $\mathbf{B}$

Abbreviations

AC	alternating current
APLAC	an object-oriented analog circuit simulator and design tool (originally Analysis Program for Linear Active Circuits)
BCS	Bardeen-Cooper-Schrieffer
DC	direct current
TEM	transverse eletromagnetic

1 Background / Context

1.1 Research objectives

main objective interim objectives

1.2 Related work

1.2.1 Feature extraction from Text

1.2.2 Text classification

References

[Alfonseca and Manandhar, 2002] Alfonseca, E. and Manandhar, S. (2002). An unsupervised method for general named entity recognition and automated concept discovery. In *Proceedings of the 1st international conference on general WordNet, Mysore, India*, page 34.

[Baluja et al., 2000] Baluja, S., Mittal, V. O., and Sukthankar, R. (2000). Applying Machine Learning for High-Performance Named-Entity Extraction. *Computational Intelligence*, 16(4):586–595.

[Bastian et al., 2014] Bastian, M., Hayes, M., Vaughan, W., Shah, S., Skomoroch, P., Kim, H., Uryasev, S., and Lloyd, C. (2014). LinkedIn skills: large-scale topic extraction and inference. pages 1–8. ACM Press.

[Bengio et al., 2006] Bengio, Y., Schwenk, H., Senécal, J.-S., Morin, F., and Gauvain, J.-L. (2006). Neural Probabilistic Language Models. In Holmes, P. D. E. and Jain, P. L. C., editors, *Innovations in Machine Learning*, number 194 in Studies in Fuzziness and Soft Computing, pages 137–186. Springer Berlin Heidelberg. DOI: 10.1007/3-540-33486-6_6.

[Bernstein et al., 2011] Bernstein, M. S., Brandt, J., Miller, R. C., and Karger, D. R. (2011). Crowds in Two Seconds: Enabling Realtime Crowd-powered Interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 33–42, New York, NY, USA. ACM.

[Blei et al., 2010] Blei, D., Carin, L., and Dunson, D. (2010). Probabilistic Topic Models. *IEEE Signal Processing Magazine*, 27(6):55–65.

[Blei, 2012] Blei, D. M. (2012). Probabilistic Topic Models. *Commun. ACM*, 55(4):77–84.

[Brooke, 2014] Brooke, J. (2014). *Computational Approaches to Style and the Lexicon*. PhD thesis, University of Toronto.

[Callison-Burch and Dredze, 2010] Callison-Burch, C. and Dredze, M. (2010). Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12. Association for Computational Linguistics.

[Chen and Goodman, 1996] Chen, S. F. and Goodman, J. (1996). An Empirical Study of Smoothing Techniques for Language Modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Cheng et al., 2014] Cheng, J., Kartsaklis, D., and Grefenstette, E. (2014). Investigating the role of prior disambiguation in deep-learning compositional models of meaning. *arXiv preprint arXiv:1411.4116*.

[Chien and Chen, 2008] Chien, C.-F. and Chen, L.-F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with Applications*, 34(1):280–290.

[Clark et al., 2013] Clark, A., Fox, C., and Lappin, S. (2013). *The Handbook of Computational Linguistics and Natural Language Processing.* John Wiley & Sons.

[Collobert and Weston, 2008] Collobert, R. and Weston, J. (2008). A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167, New York, NY, USA. ACM.

[Domingos, 2012] Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. *Commun. ACM*, 55(10):78–87.

[Echarte et al., 2007] Echarte, F., Astrain, J. J., Córdoba, A., and Villadangos, J. E. (2007). Ontology of Folksonomy: A New Modelling Method. *SAAKM*, 289:36.

[Faruqui and Padó, 2011] Faruqui, M. and Padó, S. (2011). "I Thou Thee, Thou Traitor": Predicting Formal vs. Informal Address in English Literature. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 467–472, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Gers et al., 1999] Gers, F., Schmidhuber, J., and Cummins, F. (1999). Learning to forget: continual prediction with LSTM. In *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*, volume 2, pages 850–855 vol.2.

[Ginter and Kanerva, 2014] Ginter, F. and Kanerva, J. (2014). Fast Training of word2vec Representations Using N-gram Corpora.

[Guillory and Hancock, 2012] Guillory, J. and Hancock, J. T. (2012). The effect of Linkedin on deception in resumes. *Cyberpsychology, Behavior and Social Networking*, 15(3):135–140.

[Hokey Min and Ahmed Emam, 2003] Hokey Min and Ahmed Emam (2003). Developing the profiles of truck drivers for their successful recruitment and retention: A data mining approach. *International Journal of Physical Distribution & Logistics Management*, 33(2):149–162.

[Hussain et al., 2007] Hussain, Z., Wallace, J., and Cornelius, N. E. (2007). The use and impact of human resource information systems on human resource management professionals. *Information & Management*, 44(1):74–89.

[Ipeirotis, 2010] Ipeirotis, P. G. (2010). Demographics of mechanical turk.

[Jagadish et al., 1998] Jagadish, H. V., Koudas, N., Muthukrishnan, S., Poosala, V., Sevcik, K. C., and Suel, T. (1998). Optimal histograms with quality guarantees. In *VLDB*, volume 98, pages 24–27.

[Johnson and Zhang, 2014] Johnson, R. and Zhang, T. (2014). Effective Use of Word Order for Text Categorization with Convolutional Neural Networks. *arXiv:1412.1058 [cs, stat]*. arXiv: 1412.1058.

[Kalchbrenner et al., 2014] Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

[Kim, 2014] Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

[Kotzias et al., 2015] Kotzias, D., Denil, M., de Freitas, N., and Smyth, P. (2015). From Group to Individual Labels Using Deep Features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 597–606, New York, NY, USA. ACM.

[Lafferty et al., 2001] Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Departmental Papers (CIS)*.

[Lange and Zilles, 2004] Lange, S. and Zilles, S. (2004). Formal language identification: query learning vs. Gold-style learning. *Information Processing Letters*, 91(6):285–292.

[Le and Mikolov, 2014] Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

[Lee et al., 2015] Lee, Y.-C., Hong, J., Kim, S.-W., Gao, S., and Hwang, J.-Y. (2015). On recommending job openings. In *Proceedings of the 26th ACM Conference on Hypertext &#38; Social Media*, HT '15, pages 331–332, New York, NY, USA. ACM.

[Leskovec et al., 2014] Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). *Mining of massive datasets*. Cambridge University Press.

[Lewis, 1992] Lewis, D. D. (1992). Feature selection and feature extraction for text categorization. In *Proceedings of the workshop on Speech and Natural Language*, pages 212–217. Association for Computational Linguistics.

[Little et al., 2010] Little, G., Chilton, L. B., Goldman, M., and Miller, R. C. (2010). Exploring Iterative and Parallel Human Computation Processes. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 68–76, New York, NY, USA. ACM.

[Luhn, 1958] Luhn, H. (1958). A Business Intelligence System. *IBM Journal of Research and Development*, 2(4):314–319.

[Mason and Watts, 2010] Mason, W. and Watts, D. J. (2010). Financial Incentives and the "Performance of Crowds". *SIGKDD Explor. Newsl.*, 11(2):100–108.

[Mikolov, 2012] Mikolov, T. (2012). *Statistical Language Models Based on Neural Networks*. PhD thesis, Ph. D. thesis, Brno University of Technology.

[Mikolov et al., 2013] Mikolov, T., Le, Q. V., and Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

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

[Navigli and Velardi, 2004] Navigli, R. and Velardi, P. (2004). Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics*, 30(2):151–179.

[Pavlick et al., 2014] Pavlick, E., Post, M., Irvine, A., Kachaev, D., and Callison-Burch, C. (2014). The Language Demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2(0):79–92.

[Powell et al., 2011] Powell, D. M., Goffin, R. D., and Gellatly, I. R. (2011). Gender differences in personality scores: Implications for differential hiring rates. *Personality and Individual Differences*, 50(1):106–110.

[Rothstein and Goffin, 2006] Rothstein, M. G. and Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review*, 16(2):155–180.

[Saidi Mehrabad and Fathian Brojeny, 2007] Saidi Mehrabad, M. and Fathian Brojeny, M. (2007). The development of an expert system for effective selection and appointment of the jobs applicants in human resource management. *Computers & Industrial Engineering*, 53(2):306–312.

[Shahaf et al., 2012] Shahaf, D., Guestrin, C., and Horvitz, E. (2012). Metro maps of science. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1122–1130. ACM.

[Sivaram and Ramar, 2010] Sivaram, N. and Ramar, K. (2010). Applicability of clustering and classification algorithms for recruitment data mining. *International Journal of Computer Applications*, 4(5):23–28.

[Stavrou et al., 2007] Stavrou, E. T., Charalambous, C., and Spiliotis, S. (2007). Human resource management and performance: A neural network analysis. *European Journal of Operational Research*, 181(1):453–467.

[Strzalkowski and Wang, 1996] Strzalkowski, T. and Wang, J. (1996). A Self-learning Universal Concept Spotter. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 931–936, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Taylor, 2006] Taylor, S. (2006). Acquaintance, meritocracy and critical realism: Researching recruitment and selection processes in smaller and growth organizations. *Human Resource Management Review*, 16(4):478–489.

[von Ahn and Dabbish, 2004] von Ahn, L. and Dabbish, L. (2004). Labeling Images with a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 319–326, New York, NY, USA. ACM.

[Waltz, 2003] Waltz, E. (2003). *Knowledge Management in the Intelligence Enterprise.* Artech House.

[Youyou et al., 2015] Youyou, W., Kosinski, M., and Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040.

[Zhang et al., 2004] Zhang, L., Pan, Y., and Zhang, T. (2004). Focused Named Entity Recognition Using Machine Learning. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 281–288, New York, NY, USA. ACM.

[Zhang et al., 2006] Zhang, L., Wu, X., and Yu, Y. (2006). Emergent Semantics from Folksonomies: A Quantitative Study. In Spaccapietra, S., Aberer, K., and Cudré-Mauroux, P., editors, *Journal on Data Semantics VI*, number 4090 in Lecture Notes in Computer Science, pages 168–186. Springer Berlin Heidelberg. DOI: 10.1007/11803034_8.