Katrin Tomanek

763 Guerrero Street
San Francisco, CA 94110
(415) 316 7845
katrin.tomanek@gmx.de
katrin-tomanek

Personal Information

Dr. Katrin Ruth Sarah Tomanek

Nationality: German

Place and Date of Birth: Heidelberg, 25.06.1979

About Me

Research engineer and software developer with a focus on innovative applications of natural language processing and machine learning. Passionate about solving language-related real-world problems and finding the needle in the haystack of (unstructured) data. Experienced in both adacemic research and industry projects. Ambitious and goal-directed, willing to take responsibility, and a committed team member.

Professional Experience

Since 06/2014 **Senior Data Scientist, NLP Specialist**
VigLink Inc, San Francisco/USA

At VigLink I develop and optimize NLP components for product entity recognition, their linking to offer databases, and the transformation of original offer titles into structured information. This includes amongst others statistical modeling, development of ML and NLP components and their qualitative and quantitative evaluation (including definition of test sets and supervision of annotation endeavors), and A/B testing to determine the impact of improved components on overall business metrics.

02/2011 – 05/2014 **Senior Research Engineer, Software Engineer and Project Manager**
Averbis GmbH, Freiburg/Germany
Selected projects I have been working on:

- **Document Classification and Keywording**
  Several projects with public libraries on multi-class and multi-label document classification (mostly based on Support Vector Machines) as well as automatic keywording (free keywords based on textual context) and automatic indexing (based on terminologies).
  I lead a long-term project with the German National Library (DNB). Within this project, I was also responsible for the developement of the solution for classification and indexing and its optimization to the customer's data.

- **Social Media Analytics**
  Sentiment analysis for user assessments of a major German hotel booking portal. My role was to develop an algorithm for sentiment aspect identification and aspect-based polarity prediction (positive, negative and neutral statements).
  Additionally, I supervised a bachelor thesis on sentiment analysis of German tweets based on Deep Learning.
- **Patent Analytics**
  Planning and design of a sytem for patent analytics including a semi-automatic classification framework, UIMA-based text analysis, and a SOLR-based search engine. Implementation-wise, I was responsible for the classification and the information extraction component and the extension of the system for multilinguality in search and classification.
- **Research Projects**
  I worked as a research engineer in the EU-funded *MANTRA* project where I investigated methods for automatic enrichment of terminologies for multiple languages as well as the application of such terminologies in multilingual information retrieval and document classification.
- **Text Analysis and Machine Learning**
  During my time at Averbis GmbH, I was leading a working group concerned with the maintenance and development of Averbis' text analysis and machine learning core component. This includes developement of new NLP components, extensions and adaptions of the UIMA-based framework, optimization of algorithms and trainig of models.

01/2006 – 06/2010 **Research Assistant**
Friedrich-Schiller-Universität Jena/Germany, Computational Linguistics/Prof. Dr. Udo Hahn
Research done within the following projects:

- **CALBC**: This project aimed at the automatic creation of a diversely annotated corpus with several 100,000 abstracts from life-science publications. I developed ensemble algorithms to combine and merge semantic metadata annotations produced by different NLP-systems so that the resulting annotations best fit application-specific needs.
- **StemNet**: This project aimed at building a large-scale knowledge management system for bio-medical documents. I investigated approaches for effective and practical use of Active Learning to reduce time needed to create training material for named entity recognition systems. I also developed an Active Learning-enabled annotation environment which was then used to accomplish the large-scale annotation of bio-medical entities needed in this project.
- **BOOTStrep**: This project aimed at integrating several bio-medical fact databases and terminologies and setting up a flexible text mining system for the bio-medical domain to automatically update this information. I developed components (mainly solving segmentation tasks such as named entity recognition based on structured learning with Conditional Random Fields) for bio-medical NLP and set up the UIMA-based information extraction system.

08/2003 - 12/2003 **Intern Software Development**, *IndiaNIC Infotech Limited*, Ahmedabad, India.
Development of E-commerce solutions (PHP, JavaScript, MySQL).

10/2000 – 01/2001 **Intern Software Development**, *ORGA GmbH*, Karlsruhe, Germany.
Development of a software system to manage human resource information (Visual Basic).

1999 – 2001 **Freelancer Software Migration**.
Technical support in making software year 2000 compliant (Visual Basic).

Education

Technische Universität Dortmund, Germany

01/2006 – 04/2010 **PhD**, *Computer Science*,
Thesis: "Resource-Aware Annotation through Active Learning"
Supervisors: Prof. Dr. Katharina Morik, Prof. Dr. Udo Hahn
Graduated summa cum laude

My thesis centers around the question how Active Learning (AL) – a selective sampling strategy where only examples of high utility for classifier training are selected for manual annotation – can be applied as resource-aware strategy for linguistic annotation. A set of requirements is defined and several approaches and adaptations to the standard form of AL are proposed to meet these requirements. This includes: (1) a novel method to monitor and stop the AL-driven annotation process; (2) an approach to semi-supervised AL where only highly critical tokens have to actually be manually annotated while the rest is automatically tagged; (3) a discussion and empirical investigation of the reusability of actively drawn samples; (4) a comparative study how class imbalance can be reduced right upfront during AL-driven data acquisition; (5) two methods for selective sampling of examples which are useful for multiple learning problems; (6) an extensive evaluation of the proposed approaches to AL for Named Entity Recognition with respect to both savings in corpus size and actual annotation time; and finally (7) three methods how these approaches can be made cost-conscious so as to reduce annotation time even more.

During my PhD I organized two workshops on Active Learning for Natural Language Processing, held in conjunctions with the NAACL (North American Chapter of the ACL, an internationally renowned conference for computational linguistics). I published several papers on Active Learning in peer-reviewed, internationally renowned conferences including ACL, EMNLP, NAACL, and K-CAP.

I have also been involved (and still am) into many annotations projects (both with and without Active Learning) and am a member of the organizing committee of the Linguistic Annotation Workshop (LAW), a workshop which is run every year in conjunction with one of the major international conferences on computational linguistics..

Universität Karlsruhe (TH), Germany

10/1998 – 09/2005 **Diploma (equivalent to M.Sc.)**, *Industrial Engineering*,
Specializations: computer science, data mining, operations research
Overall grade: "sehr gut" (1.4)
Thesis: "Ontology-driven classification of named entities based on a machine learning approach"
Supervisors: Prof. Dr. Rudi Studer, Philipp Cimiano.

Leipniz-Gymnasium Östringen, Germany

1998 **Abitur (university-entrance diploma in German school system)**, grade: "very good" (1.4).

Teaching Experience

09/2008	Tutorial on the Unstructured Information Management Architecture (UIMA) at the 3rd International Symposium on Semantic Mining in Biomedicine (SMBM).
Winter 2007	Teaching Assistant (lab group): natural language processing and UIMA.
Summer 2006	Teaching Assistant (lab group): mapping of biomedical named entities to databases.
Winter 2001	Teaching Assistant: statistics for social sciences.

Awards

2009	Best paper award at the International Conference on Knowledge Capture (KCAP'09)

Research Community Service

Co-Chair	NAACL Workshops on Active Learning for Natural Language Processing 2009 and 2010; several workshops on the Unstructured Information Management Architecture (UIMA) including LREC 2008, GSCL 2009, GSCL 2013; 5th Linguistic Annotation Workshop (LAW-V) 2011.
Programme Committee/ Reviewing	Coling 2010; LREC Workshop on New Challenges for NLP Frameworks 2010; NAACL Workshop on Active Learning for Natural Language Processing 2009 and 2010; ECML/PKDD Workshop on High-level Information Extraction 2008; multiple workshops on the Unstructured Information Management Architecture (UIMA) including LREC 2008, GSCL 2009, GSCL 2013; Linguistic Annotation Workshops (LAW) 2011, 2012, 2013, 2014, 2015; KDD 2015
Scientific Memberships	SIGANN (ACL Special Interest Group for Annotation)

Technical Skills

Natural Language Processing: UIMA, GATE, openNLP, Stanford NLP, LingPipe, NTLK

Information Retrieval: Apache SOLR/Lucene, Elasticsearch

ML & Data Mining: Apache Spark, MLLib, Mallet, Weka, Rapidminer, R, LibSVM, scikit-learn, scipy, pandas, Hadoop

Databases: MySQL, PostgreSQL, MongoDB, Neo4J, Cassandra, Hive

Programming Languages: Java, Python, Scala

Software Development: Eclipse, Subversion, Mercurial, git, Maven, ANT, JUnit

Operating Systems: Linux/Unix, Windows, Mac

Language Skills

German	**native**
English	**fluent**
French	**good**
Spanish	**good**

Publications

Kornel Marko, **Katrin Tomanek**, and Oliver Juwig. Social Media Analytics: Automatische Analyse von Hotelbewertungen. In *KnowTech 2013 - 15. Kongress für Wissensmanagement und Social Media*, pages 317–324, 2013.

Peter Klügl, Richard Eckart de Castilho, and **Katrin Tomanek**, editors. *Proceedings of the 3rd Workshop on Unstructured Information Management Architecture, Darmstadt, Germany, September 23, 2013*, volume 1038 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

Katrin Tomanek, Philipp Daumke, Frank Enders, Jens Huber, Katharina Theres, and Marcel Müller. An interactive de-identification-system. In *Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine (SMBM 2012)*, 2012.

Katrin Tomanek, Frank Enders, Philipp Daumke, Marcel Müller, Martin Sedlmayr, and Hans-Ulrich Prokosch. Ein System zur De-identifikation medizinischer Rohdaten. In *GMDS 2012 - Tagungsband der 57. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie*, 2012.

Katrin Tomanek and Katharina Morik. Inspecting sample reusability for active learning. In *JMLR: Workshop and Conference Proceedings; Workshop on Active Learning and Experimental Design*, pages 169–181, 2011.

Katrin Tomanek. *Resource-Aware Annotation through Active Learning*. PhD thesis, Technical University of Dortmund, 2010.

Katrin Tomanek, Udo Hahn, Steffen Lohmann, and Jürgen Ziegler. A cognitive cost model of annotations based on eye-tracking data. In *ACL'10: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1158–1167, 2010.

Katrin Tomanek and Udo Hahn. A comparison of approaches to cost-sensitive active learning. In *COLING'01: Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1247–1255, 2010.

Steffen Lohmann, **Katrin Tomanek**, Jürgen Ziegler, and Udo Hahn. Getting at the cognitive complexity of linguistic metadata annotation – a pilot study using eye-tracking. In *COGSCI'10: Proceedings of the Annual Cognitive Science Conference*, pages 2146–2151, 2010.

Udo Hahn, **Katrin Tomanek**, Elena Beisswanger, and Erik Faessler. A proposal for a configurable silver standard. In *The LAW at ACL 2010: Proceedings of the 4th Linguistic Annotation Workshop*, pages 235–242. Association for Computational Linguistics, 2010.

Katrin Tomanek and Udo Hahn. Annotation time stamps — temporal metadata from the linguistic annotation process. In *LREC'10: Proceedings of the 7th International Conference on Language Resources and Evaluation*. European Language Resources Association, 2010.

Erik Faessler, Rico Landefeld, **Katrin Tomanek**, and Udo Hahn. LuCas - A Lucene CAS Indexer. In *Proceedings of the Biennial GSCL Conference 2009*, pages 217–224. Gunter Narr, 2009.

Katrin Tomanek and Udo Hahn. Reducing class imbalance during active learning for Named Entity annotation. In *K-CAP'09 — Proceedings of the 5th International Conference on Knowledge Capture*, pages 105–112. ACM, 2009.

Katrin Tomanek and Udo Hahn. Semi-supervised active learning for sequence labeling. In *ACL/IJCNLP'09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1039–1047. Association for Computational Linguistics, 2009.

Katrin Tomanek and Udo Hahn. Timed Annotations — Enhancing MUC7 metadata by the time it takes to annotate Named Entities. In *Proceedings of the 3rd Linguistic Annotation Workshop*, pages 112–115. Association for Computational Linguistics, 2009.

Fredrik Olsson and **Katrin Tomanek**. An intrinsic stopping criterion for committee-based active learning. In *CoNLL'09: Proceedings of the 13th Conference on Computational Natural Language Learning*, pages 138–146. Association for Computational Linguistics, 2009.

Katrin Tomanek and Fredrik Olsson. A web survey on the use of active learning to support annotation of text data. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 45–48. Association for Computational Linguistics, 2009.

Katrin Tomanek, Florian Laws, Udo Hahn, and Hinrich Schütze. On proper unit selection in active learning: Co-selection effects for Named Entity Recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 9–17. Association for Computational Linguistics, 2009.

Udo Hahn, **Katrin Tomanek**, Ekaterina Buyko, Jung Jae Kim, and Dietrich Rebholz-Schuhmann. How feasible and robust is the automatic extraction of gene regulation events? A cross-method evaluation under lab and real-life conditions. In *Proceedings of the NAACL workshop on BioNLP 2009*, pages 37–45. Association for Computational Linguistics, 2009.

Joachim Wermter, **Katrin Tomanek**, and Udo Hahn. High-performance gene name normalization with GeNo. *Bioinformatics*, 25(6):815–821, 2009.

Roi Reichart, **Katrin Tomanek**, Udo Hahn, and Ari Rappoport. Multi-task active learning for linguistic annotations. In *ACL'08: Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*, pages 861–869. Association for Computational Linguistics, 2008.

Udo Hahn, Elena Beisswanger, Ekaterina Buyko, Michael Poprat, **Katrin Tomanek**, and Joachim Wermter. Semantic annotations for biology: A corpus development initiative at the Jena University Language & Information Engineering (JULIE) Lab. In *LREC'08: Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 2257–2261. European Language Resources Association, 2008.

Katrin Tomanek and Udo Hahn. Approximating learning curves for active-learning-driven annotation. In *LREC'08: Proceedings of the 6th International Language Resources and Evaluation*, pages 1319–1324. European Language Resources Association, 2008.

Udo Hahn, Ekaterina Buyko, Rico Landefeld, Matthias Mühlhausen, Michael Poprat, **Katrin Tomanek**, and Joachim Wermter. An overview of JCoRe, the JULIE Lab UIMA component repository. In *Proceedings of the LREC'08 Workshop 'Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP'*, pages 1–7, 2008.

Roman Klinger and **Katrin Tomanek**. Classical Probabilistic Models and Conditional Random Fields. Algorithm Engineering Report TR07-2-013, Department of Computer Science, Dortmund University of Technology, 2007.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. A reappraisal of sentence and token splitting for life sciences documents. In *MEDINFO'07: Proceedings of the 12th World Congress on Medical Informatics. Building Sustainable Health Systems*, pages 524–528. IOS Press, 2007.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. Sentence and token splitting based on conditional random fields. In *PACLING'07: Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 49–57. Pacific Association for Computational Linguistics, 2007.

Ekaterina Buyko, **Katrin Tomanek**, and Udo Hahn. Resolution of coordination ellipses in biological named entities using conditional random fields. In *PACLING'07: Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 163–171. Pacific Association for Computational Linguistics, 2007.

Ekaterina Buyko, **Katrin Tomanek**, and Udo Hahn. Resolution of coordination ellipses in complex biological named entity mentions using conditional random fields. In *BioLINK'07: Proceedings of the BioLINK SIG 2007. The Annual Meeting of the ISMB BioLINK Special Interest Group on Text Data Mining, in Association with ISMB 2007*, pages 9–12, 2007.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. Efficient annotation with the Jena ANnotation Environment. In *BioLINK'07: Proceedings of the BioLINK SIG 2007. The Annual Meeting of the ISMB BioLINK Special Interest Group on Text Data Mining, in Association with ISMB 2007*, pages 43–46, 2007.

Ekaterina Buyko, Scott Piao, Yoshimasa Tsuruoka, **Katrin Tomanek**, Jin-Dong Kim, John McNaught, Udo Hahn, Jian Su, and Sophia Ananiadou. Bootstrep annotation scheme: Encoding information for text mining. In *Proceedings of the 4th Corpus Linguistics Conference*, 2007.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. An approach to text corpus construction which cuts annotation costs and maintains corpus reusability of annotated data. In *EMNLP-CoNLL'07: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning*, pages 486–495. Association for Computational Linguistics, 2007.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. Efficient annotation with the Jena ANnotation Environment (JANE). In *Proceedings of the 1st Linguistic Annotation Workshop*, pages 9–16. Association for Computational Linguistics, 2007.

Udo Hahn, Ekaterina Buyko, **KatrinTomanek** , Scott Piao, John McNaught, Yoshimasa Tsuruoka, and Sophia Ananiadou. An annotation type system for a data-driven NLP pipeline. In *The LAW at ACL 2007: Proceedings of the 1st Linguistic Annotation Workshop*, pages 33–40. Association for Computational Linguistics, 2007.

Joachim Wermter, **Katrin Tomanek**, and Felix Balzer. Automatische Erkennung und effiziente Annotation von anonymisierungsrelevanten Begriffen in klinischen Freitexten. In *GMDS 2006 - Tagungsband der 51. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie*, pages 151–152. Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V., 2006.