Your Topic Goes Here: An Annotated Bibliography
East Carolina University

Your Name Goes Here

Contents

1 Introduction

This bibliography is intended for beginning students of Natural Language Processing (NLP). Hence, its focus is intentionally broad.

2 Facets of NLP

There are several facets to NLP including those that unify various aspects (general), language modeling, tagging, parsing, information extraction, machine translation, language generation, dialog systems, question-answering systems, discourse understanding, clustering, and summarization.

2.1 General

These are general works that are of interest to beginning students. More to come.

Annotated Bibliography

[1] *EMNLP '11: Proceedings of the First Workshop on Unsupervised Learning in NLP*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011.

[2] Anelia Belogay et al. "Harnessing NLP techniques in the processes of multilingual content management". In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. EACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 6–10.

[3] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, 2009.

[4] The University of Sheffield. *GATE – Java Infrastructure for Human Language Technology*. http://gate.ac.uk/. Sept. 2012.

[5] Sharon Goldwater. "Unsupervised NLP and human language acquisition: making connections to make progress". In: *Proceedings of the First Workshop on Unsupervised Learning in NLP*. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1–1. – Natural language processing and cognitive science are two fields in which unsupervised language learning is an important area of research. Yet there is often little crosstalk between the two fields. In this talk, I will argue that considering the problem of unsupervised language learning from a cognitive perspective can lead to useful insights for the NLP researcher, while also showing how tools and methods from NLP and machine learning can shed light on human language acquisition. I will present two case examples, both of them models inspired by cognitive questions. The first is a model of word segmentation, which introduced new modeling and inference techniques into NLP while also yielding a better fit than previous models to human behavioral data on word segmentation. The second is more recent work on unsupervised grammar induction, in which prosodic cues are used to help identify syntactic boundaries. Preliminary results indicate that such cues can be helpful, but also reveal weaknesses in existing unsupervised grammar induction methods from NLP, suggesting possible directions for future research.

[6] Amit Goyal et al. “Sketching techniques for large scale NLP”. In: *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*. WAC-6 ’10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 17–25. – In this paper, we address the challenges posed by large amounts of text data by exploiting the power of hashing in the context of streaming data. We explore sketch techniques, especially the Count-Min Sketch, which approximates the frequency of a word pair in the corpus without explicitly storing the word pairs themselves. We use the idea of a conservative update with the Count-Min Sketch to reduce the average relative error of its approximate counts by a factor of two. We show that it is possible to store all words and word pairs counts computed from 37 GB of web data in just 2 billion counters (8 GB RAM). The number of these counters is up to 30 times less than the stream size which is a big memory and space gain. In Semantic Orientation experiments, the PMI scores computed from 2 billion counters are as effective as exact PMI scores.

[7] Nitin Indurkhya and Fred J. Damerau. *Handbook of Natural Language Processing*. Chapman & Hall/CRC, 2010. – Features: Explores the practical aspects of building natural language processing systems Covers many well-known and emerging applications, including machine translation, biomedical text mining, and sentiment analysis Provides examples of how to apply the techniques to languages other than English Offers web links, supplementary material, and updates to chapters on a companion wiki: http://handbookofnlp.cse.unsw.edu.au Summary: The Handbook of Natural Language Processing, Second Edition presents practical tools and techniques for implementing natural language processing in computer systems. Along with removing outdated material, this edition updates every chapter and expands the content to include emerging areas, such as sentiment analysis. New to the Second Edition: Greater prominence of statistical approaches New applications section Broader multilingual scope to include Asian and European languages, along with English An actively maintained wiki (http://handbookofnlp.cse.unsw.edu.au) that provides online resources, supplementary information, and up-to-date developments Divided into three sections, the book first surveys classical techniques, including both symbolic and empirical approaches. The second section focuses on statistical approaches in natural language processing. In the final section of the book, each chapter describes a particular class of application, from Chinese machine translation to information visualization to ontology construction to biomedical text mining. Fully updated with the latest developments in the field, this comprehensive, modern handbook emphasizes how to implement practical language processing tools in computational systems.

[8] Grant S. Ingersoll, Thomas S. Morton, and Andrew L. Farris. *Taming Text: How to Find, Organize, and Manipulate It*. Manning Publications Co., 2012. – It is no secret that the world is drowning in text and data. This causes real problems for everyday users who need to make sense of all the information available, and for software engineers who want to make their text-based applications more useful and user-friendly. Whether building a search engine for a corporate website, automatically organizing email, or extracting important nuggets of information from the news, dealing with unstructured text can be daunting. Taming Text is a hands-on, example-driven guide to working with unstructured text in the context of real-world applications. It explores how to automatically organize text, using approaches such as full-text search, proper name recognition, clustering, tagging, information extraction, and summarization. This book gives examples illustrating each of these topics, as well as the foundations upon which they are built.

[9] Nitin Madnani. “Getting started on natural language processing with Python”. In: *Crossroads* 13.4 (Sept. 2007), pp. 5–5.

[10] Christopher Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[11] Peter Norvig. "Natural Language Corpus Data". In: *Beautiful Data: The Stories Behind Elegant Data Solutions*. Ed. by Toby Segaran and Jeff Hammerbacher. O'Reilly Media, 2009. Chap. 14, pp. 219–242.

[12] Jacob Perkins. *Python Text Processing with NLTK 2.0 Cookbook*. Packt Publishing, 2010.

[13] Martin Popel and Zdeněk Žabokrtský. "TectoMT: modular NLP framework". In: *Proceedings of the 7th international conference on Advances in natural language processing*. IceTAL'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 293–304. – In the present paper we describe TectoMT, a multi-purpose open-source NLP framework. It allows for fast and efficient development of NLP applications by exploiting a wide range of software modules already integrated in TectoMT, such as tools for sentence segmentation, tokenization, morphological analysis, POS tagging, shallow and deep syntax parsing, named entity recognition, anaphora resolution, tree-to-tree translation, natural language generation, word-level alignment of parallel corpora, and other tasks. One of the most complex applications of TectoMT is the English-Czech machine translation system with transfer on deep syntactic (tectogrammatical) layer. Several modules are available also for other languages (German, Russian, Arabic).Where possible, modules are implemented in a language-independent way, so they can be reused in many applications.

[14] Stanford NLP Group. *Stanford NLP Tools*. http://nlp.stanford.edu/software/index.shtml. Sept. 2012.

[15] *EANL '08: Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. – The use of NLP in educational applications is becoming increasingly widespread and sophisticated. Such applications are intended to fulfil a variety of needs, from automated scoring of essays and shortanswer responses, to grammatical error detection, to assisting learners in the development of their writing, reading, and speaking skills, in both their native and non-native languages. The rapid growth of this area of research is evidenced by the number of topic-specific workshops in recent years. This workshop is the next in a series which began at ACL 1997 and continued on with HTL/NAACL 2003 and ACL 2005. Since 1997, there have also been other related meetings such as the InSTIL/ICALL Symposium at COLING 2004, and most recently the CALICO 2008 workshop entitled Automatic Analysis of Learner Language: Bridging Foreign Language Teaching Needs and NLP Possibilities. In keeping with previous workshops, our aim is to bring together the ever-growing community of researchers from both academic institutions and industry, and foster communication on issues regarding the broad spectrum of instructional settings, from K-12 to university level to EFL/ESL and professional contexts. In this endeavor, we are assisted by the wide variety of topics and languages covered by the papers presented. For this workshop, we received 18 submissions, and accepted 13 papers: 8 were accepted as long presentations (20 minutes) and 5 as short presentations (15 minutes). All accepted papers are published in these proceedings as full-length papers of up to 9 pages. Each paper was reviewed by two members of the Program Committee.

[16] The Apache Software Foundation. *openNLP*. http://opennlp.apache.org/. Sept. 2012.

2.2 NLP Education

These are general works that are of interest to NLP learning and teaching.

Annotated Bibliography

[1] Steven Bird et al. “Multidisciplinary instruction with the Natural Language Toolkit”. In: *Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics*. TeachCL ’08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 62–70.

[2] *EdAppsNLP 05: Proceedings of the second workshop on Building Educational Applications Using NLP*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005.

[3] Ken Church et al. “Report on the first summer school on NLP and IR in Beijing”. In: *SIGIR Forum* 45.2 (Jan. 2012), pp. 41–42.

[4] Judy Cushing and Rachel Hastings. “Introducing computational linguistics with NLTK (Natural Language Toolkit)”. In: *J. Comput. Sci. Coll.* 25.1 (Oct. 2009), pp. 167–169.

[5] Reva Freedman. “Concrete assignments for teaching NLP in an M.S. program”. In: *Proceedings of the Second ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. TeachNLP ’05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 37–42. – The professionally oriented computer science M.S. students at Northern Illinois University are intelligent, interested in new ideas, and have good programming skills and a good math background. However, they have no linguistics background, find traditional academic prose difficult and uninteresting, and have had no exposure to research. Given this population, the assignments I have found most successful in teaching Introduction to NLP involve concrete projects where students could see for themselves the phenomena discussed in class. This paper describes three of my most successful assignments: duplicating Kernighan et al.’s Bayesian approach to spelling correction, a study of Greenberg’s universals in the student’s native language, and a dialogue generation project. For each assignment I discuss what the students learned and why the assignment was successful.

[6] Reva Freedman. “Teaching NLP to computer science majors via applications and experiments”. In: *Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics*. TeachCL ’08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 114–119. – Most computer science majors at Northern Illinois University, whether at the B.S. or M.S. level, are professionally oriented. However, some of the best students are willing to try something completely different. NLP is a challenge for them because most have no background in linguistics or artificial intelligence, have little experience in reading traditional academic prose, and are unused to open-ended assignments with gray areas. In this paper I describe a syllabus for Introduction to NLP that concentrates on applications and motivates concepts through student experiments. Core materials include an introductory linguistics textbook, the Jurafsky and Martin textbook, the NLTK book, and a Python textbook.

[7] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Second. Prentice Hall, 2009. – ... ideal for ... linguists who want to learn more about computational modeling and techniques in language processing; computer scientists building language applications who want to learn more about the linguistic underpinnings of the field; speech technologists who want to learn more about language understanding, semantics and discourse; and all those wanting to learn more about speech processing. For instructors ... this book is a dream. It covers virtually every aspect of NLP... What’s truly astounding is that the book covers such a broad range of topics, while giving the reader the depth to understand and make use of the concepts, algorithms and techniques that are presented... ideal as a course textbook for advanced undergraduates, as well as graduate students and researchers in the field. – Jo-

hanna Moore, University of Edinburgh Speech and Language Processing is a comprehensive, reader-friendly, and up-to-date guide to computational linguistics, covering both statistical and symbolic methods and their application. It will appeal both to senior undergraduate students, who will find it neither too technical nor too simplistic, and to researchers, who will find it to be a helpful guide to the newly established techniques of a rapidly growing research field. – Graeme Hirst, University of Toronto This book is an absolute necessity for instructors at all levels, as well as an indispensable reference for researchers. Introducing NLP, computational linguistics, and speech recognition comprehensively in a single book is an ambitious enterprise. The authors have managed it admirably, paying careful attention to traditional foundations, relating recent developments and trends to those foundations, and tying it all together with insight and humor. Remarkable. – Philip Resnik, University of Maryland This is quite simply the most complete introduction to natural language and speech technology ever written. Virtually every topic in the field is covered, in a prose style that is both clear and engaging. The discussion is linguistically informed, and strikes a nice balance between theoretical computational models, and practical applications. It is an extremely impressive achievement. – Richard Sproat, AT&T Labs – Research – This text refers to an out of print or unavailable edition of this title.

[8] Yoshinobu Kano et al. “Integrated NLP evaluation system for pluggable evaluation metrics with extensive interoperable toolkit”. In: *Proceedings of the Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing*. SETQA-NLP ’09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 22–30. – To understand the key characteristics of NLP tools, evaluation and comparison against different tools is important. And as NLP applications tend to consist of multiple semi-independent sub-components, it is not always enough to just evaluate complete systems, a fine grained evaluation of underlying components is also often worthwhile. Standardization of NLP components and resources is not only significant for reusability, but also in that it allows the comparison of individual components in terms of reliability and robustness in a wider range of target domains. But as many evaluation metrics exist in even a single domain, any system seeking to aid inter-domain evaluation needs not just predefined metrics, but must also support pluggable user-defined metrics. Such a system would of course need to be based on an open standard to allow a large number of components to be compared, and would ideally include visualization of the differences between components. We have developed a pluggable evaluation system based on the UIMA framework, which provides visualization useful in error analysis. It is a single integrated system which includes a large ready-to-use, fully interoperable library of NLP tools.

[9] Dan Klein. “A core-tools statistical NLP course”. In: *Proceedings of the Second ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. TeachNLP ’05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 23–27. – In the fall term of 2004, I taught a new statistical NLP course focusing on core tools and machine-learning algorithms. The course work was organized around four substantial programming assignments in which the students implemented the important parts of several core tools, including language models (for speech reranking), a maximum entropy classifier, a part-of-speech tagger, a PCFG parser, and a word-alignment system. Using provided scaffolding, students built realistic tools with nearly state-of-the-art performance in most cases. This paper briefly outlines the coverage of the course, the scope of the assignments, and some of the lessons learned in teaching the course in this way. More to come.

[10] Lillian Lee. “A non-programming introduction to computer science via NLP, IR, and AI”. In: *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1*. ETMTNLP ’02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 33–38. – This paper describes a new Cornell University course serving as a non-programming introduction to computer science, with natural language processing and information retrieval forming a crucial part of the syllabus. Material was drawn from a wide variety of topics (such as theories of discourse structure and random graph models of the World Wide Web) and presented at some technical depth, but was massaged to make it suitable for a freshman-level course. Student feedback from the first running of the class was overall quite positive, and a grant from the GE Fund has been awarded to further support the course’s development and goals.

[11] Elizabeth D. Liddy and Nancy J. McCracken. “Hands-on NLP for an interdisciplinary audience”. In: *Proceedings of the Second ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. TeachNLP ’05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 62–68. – The need for a single NLP offering for a diverse mix of graduate students (including computer scientists, information scientists, and linguists) has motivated us to develop a course that provides students with a breadth of understanding of the scope of real world applications, as well as depth of knowledge of the computational techniques on which to build in later experiences. We describe the three hands-on tasks for the course that have proven successful, namely: 1) in-class group simulations of computational processes; 2) team posters and public presentations on state-of-the-art commercial NLP applications, and; 3) team projects implementing various levels of human language processing using open-source software on large textual collections. Methods of evaluation and indicators of success are also described.

[12] Nitin Madnani and Bonnie J. Dorr. “Combining open-source with research to re-engineer a hands-on introductory NLP course”. In: *Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics*. TeachCL ’08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 71–79. – We describe our first attempts to re-engineer the curriculum of our introductory NLP course by using two important building blocks: (1) Access to an easy-to-learn programming lan- guage and framework to build hands-on pro- gramming assignments with real-world data and corpora and, (2) Incorporation of interest- ing ideas from recent NLP research publica- tions into assignment and examination prob- lems. We believe that these are extremely im- portant components of a curriculum aimed at a diverse audience consisting primarily of first- year graduate students from both linguistics and computer science. Based on overwhelm- ingly positive student feedback, we find that our attempts were hugely successful.

[13] *EdAppsNLP ’09: Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. – NLP researchers are now building educational applications across a number of areas, including automated evaluation of student writing and speaking, rich grammatical error detection with an increasing focus on English language learning, tools to support student reading, and intelligent tutoring. This workshop is the fourth in a series, specifically related to "Building NLP Applications for Education", that began at NAACL/HLT (2003), and continued at ACL 2005 (Ann Arbor), ACL-HLT 2008 (Columbus), and now, at NAACL-HLT 2009 (Boulder). Research in this area continues to grow, and there is ever-increasing interest which was evidenced this year by the fact that we had the largest number of submissions. For this workshop, we received 25 submissions and accepted 12 papers. All of these papers are published

in these proceedings. Each paper was reviewed by at least two members of the Program Committee.

[14] *IUNLPBEA '10: Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. – NLP researchers are now building educational applications across a number of areas, including automated evaluation of student writing and speaking, rich grammatical error detection with an increasing focus on English language learning, tools to support student reading, and intelligent tutoring. This workshop is the fifth in a series, specifically related to Building NLP Applications for Education, that began at NAACL/HLT (2003), and continued at ACL 2005 (Ann Arbor), ACL/HLT 2008 (Columbus), NAACL/HLT 2009 (Boulder), and now at NAACL/HLT 2010 (Los Angeles). Research in this area continues to grow, and there is ever-increasing interest and practical application which was evidenced this year, again, by an even larger number of submissions. We received a record 28 submissions and accepted 13 papers, two of which include demos. All of the papers are published in these proceedings. Each paper was carefully reviewed by two members of the Program Committee. We selected reviewers most appropriate for each paper so as to give more helpful feedback and comments. This workshop offers an opportunity to present and publish work that is highly relevant to NAACL, but is also highly specialized, and so this workshop is often a more appropriate venue for such work. While the field is growing, we do recognize that there is a core group of institutions and researchers who work in this area. That said, we continue to have a very strong policy to deal with conflicts of interest. First, reviewers were not assigned any papers to evaluate if the paper had an author from their institution. Second, with respect to the organizing committee, authors of papers where there was a conflict of interest recused themselves from the discussion.

[15] *IUNLPBEA '11: Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. – Research in NLP applications for education continues to progress using innovative NLP techniques. New technologies have made it possible to include speech in both assessment and in Intelligent Tutoring Systems (ITS). NLP techniques are also being used to generate assessments and tools for curriculum development of reading materials, as well as tools to support assessment and test development. As a community, we continue to improve existing capabilities and to identify and generate innovative and creative ways to use NLP in applications for writing, reading, speaking, critical thinking, and assessment. In this workshop, we focus on contributions to core educational problem spaces: development of curriculum and assessment (e.g., applications that help teachers develop reading materials), delivery of curriculum and assessments (e.g., applications where the student receives instruction and interacts with the system), and reporting of assessment outcomes (e.g., automated essay scoring). The need for, and the rapid development of, language-based capabilities have been driven by increased requirements for state and national assessments and a growing population of foreign and second language learners. This is the sixth in a series of workshops on Building NLP Applications for Education that began at NAACL/HLT 2003 (Edmonton), and continued at ACL 2005 (Ann Arbor), ACL/HLT 2008 (Columbus), NAACL/HLT 2009 (Boulder), NAACL/HLT 2010 (Los Angeles), and now ACL/HLT 2011 (Portland). Research in this area continues to grow, and there is ever-increasing interest and practical application that was evidenced this year, again, by the large number of submissions.

[16] Fei Xia. "The evolution of a statistical NLP course". In: *Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics*. TeachCL '08. Stroudsburg, PA, USA: Association

for Computational Linguistics, 2008, pp. 45–53. – This paper describes the evolution of a statistical NLP course, which I have been teaching every year for the past three years. The paper will focus on major changes made to the course (including the course design, assignments, and the use of discussion board) and highlight the lessons learned from this experience.

2.3 Language Modeling

These works discuss approaches to modeling and evaluating natural languages.

Annotated Bibliography

[1] Anja Belz and Eric Kow. "Discrete vs. continuous rating scales for language evaluation in NLP". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 230–235.

[2] Thorsten Brants and Alex Franz. *Web 1T 5-gram Version 1*. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13, Linguistic Data Consortium. 2012.

[3] Alexis Dimitriadis. "Matching needs and resources: how NLP can help theoretical linguistics". In: *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*. NLPLING '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 22–24. – While some linguistic questions pose challenges that could be met by developing and applying NLP techniques, other problems can best be approached with a blend of old-fashioned linguistic investigation and the use of simple, well-established NLP tools. Unfortunately, this means that the NLP component is too simple to be of interest to the computationally-minded, while existing tools are often difficult for the programming novice to use. For NLP to come to the aid of research in theoretical linguistics, a continuing investment of effort is required to bridge the gap. This investment can be made from both sides.

[4] Amit Goyal, Hal Daumé III, and Suresh Venkatasubramanian. "Streaming for large scale NLP: language modeling". In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. NAACL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 512–520. – In this paper, we explore a streaming algorithm paradigm to handle large amounts of data for NLP problems. We present an efficient low-memory method for constructing high-order approximate n-gram frequency counts. The method is based on a deterministic streaming algorithm which efficiently computes approximate frequency counts over a stream of data while employing a small memory footprint. We show that this method easily scales to billion-word monolingual corpora using a conventional (8 GB RAM) desktop machine. Statistical machine translation experimental results corroborate that the resulting high-n approximate small language model is as effective as models obtained from other count pruning methods.

[5] Serguei A. Mokhov. "Evolution of MARF and its NLP framework". In: *Proceedings of the Third C* Conference on Computer Science and Software Engineering*. C3S2E '10. New York, NY, USA: ACM, 2010, pp. 118–122. – We review the evolution, challenges, and the future of the open-source MARF framework and its applications from being an audio recognition system into a general recognition pipeline for voice, speech, natural language, forensics, security applications, and other classification tasks, including from becoming a single-threaded pipeline into an autonomic distributed system designed to work in heterogeneous environments and being

a research platform for comparative studies of algorithms while applying software engineering methodology to the framework's design to remain flexible and extensible.

[6] Alejandro Rago et al. "Early aspect identification from use cases using NLP and WSD techniques". In: *Proceedings of the 15th workshop on Early aspects*. EA '09. New York, NY, USA: ACM, 2009, pp. 19–24. – In this article, we present a semi-automated approach for identifying candidate early aspects in requirements specifications. This approach aims at improving the precision of the aspect identification process in use cases, and also solving some problems of existing aspect mining techniques caused by the vagueness and ambiguity of text in natural language. To do so, we apply a combination of text analysis techniques such as: natural language processing (NLP) and word sense disambiguation (WSD). As a result, our approach is able to generate a graph of candidate concerns that crosscut the use cases, as well as a ranking of these concerns according to their importance. The developer then selects which concerns are relevant for his/her domain. Although there are still some challenges, we argue that this approach can be easily integrated into a UML development methodology, leading to improved requirements elicitation.

[7] Davy Weissenbacher. "Bayesian network, a model for NLP?" In: *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*. EACL '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 195–198. – The NLP systems often have low performances because they rely on unreliable and heterogeneous knowledge. We show on the task of non-anaphoric it identification how to overcome these handicaps with the Bayesian Network (BN) formalism. The first results are very encouraging compared with the state-of-the-art systems.

2.4 Tagging and Annotation

These works discuss approaches to tagging natural language text.

Annotated Bibliography

[1] Asif Ekbal and Sriparna Saha. "Weighted Vote-Based Classifier Ensemble for Named Entity Recognition: A Genetic Algorithm-Based Approach". In: 10.2 (June 2011), 9:1–9:37. – In this article, we report the search capability of Genetic Algorithm (GA) to construct a weighted vote-based classifier ensemble for Named Entity Recognition (NER). Our underlying assumption is that the reliability of predictions of each classifier differs among the various named entity (NE) classes. Thus, it is necessary to quantify the amount of voting of a particular classifier for a particular output class. Here, an attempt is made to determine the appropriate weights of voting for each class in each classifier using GA. The proposed technique is evaluated for four leading Indian languages, namely Bengali, Hindi, Telugu, and Oriya, which are all resource-poor in nature. Evaluation results yield the recall, precision and F-measure values of 92.08%, 92.22%, and 92.15%, respectively for Bengali; 96.07%, 88.63%, and 92.20%, respectively for Hindi; 78.82%, 91.26%, and 84.59%, respectively for Telugu; and 88.56%, 89.98%, and 89.26%, respectively for Oriya. Finally, we evaluate our proposed approach with the benchmark dataset of CoNLL-2003 shared task that yields the overall recall, precision, and F-measure values of 88.72%, 88.64%, and 88.68%, respectively. Results also show that the vote based classifier ensemble identified by the GA-based approach outperforms all the individual classifiers, three conventional baseline ensembles, and some other existing ensemble techniques. In a part of the article, we formulate the problem of feature selection in any classifier under the single

objective optimization framework and show that our proposed classifier ensemble attains superior performance to it.

[2] Eduard Hovy. "Injecting linguistics into NLP through annotation". In: *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*. NLPLING '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 79–79. – Over the past 20 years, the size of the L in Computatioonal Linguistics has been shrinking relative to the size of the C. The result is that we are incresingly becoming a community of uniformed but sophisticated engineers, applying to problems very complex machine learning techninques that use very simple (simplistic?) analyses/theories. (Try finding a theoretical account of subjectivity, opinion, entailment, or inference in publications surrounding the associated competitions of the past few years.)

[3] Pontus Stenetorp et al. "BRAT: a web-based tool for NLP-assisted text annotation". In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. EACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 102–107. – We introduce the brat rapid annotation tool (BRAT), an intuitive web-based tool for text annotation supported by Natural Language Processing (NLP) technology. BRAT has been developed for rich structured annotation for a variety of NLP tasks and aims to support manual curation efforts and increase annotator productivity using NLP techniques. We discuss several case studies of real-world annotation projects using pre-release versions of BRAT and present an evaluation of annotation assisted by semantic class disambiguation on a multicategory entity mention annotation task, showing a 15% decrease in total annotation time. BRAT is available under an open-source license from: http://brat.nlplab.org.

[4] Yue Zhang and Stephen Clark. "A fast decoder for joint word segmentation and POS-tagging using a single discriminative model". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. EMNLP '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 843–852. – We show that the standard beam-search algorithm can be used as an efficient decoder for the global linear model of Zhang and Clark (2008) for joint word segmentation and POS-tagging, achieving a significant speed improvement. Such decoding is enabled by: (1) separating full word features from partial word features so that feature templates can be instantiated incrementally, according to whether the current character is separated or appended; (2) deciding the POS-tag of a potential word when its first character is processed. Early-update is used with perceptron training so that the linear model gives a high score to a correct partial candidate as well as a full output. Effective scoring of partial structures allows the decoder to give high accuracy with a small beam-size of 16. In our 10-fold cross-validation experiments with the Chinese Tree-bank, our system performed over 10 times as fast as Zhang and Clark (2008) with little accuracy loss. The accuracy of our system on the standard CTB 5 test was competitive with the best in the literature.

2.5 Parsing

These works discuss approaches and algorithms for parsing natural language text.

Annotated Bibliography

[1] Khaled Abdalgader and Andrew Skabar. "Unsupervised similarity-based word sense disambiguation using context vectors and sentential word importance". In: *ACM Trans. Speech Lang. Process.* 9.1 (May 2012), 2:1–2:21.

[2] Dmitry Davidov, Roi Reichart, and Ari Rappoport. “Superior and efficient fully unsupervised pattern-based concept acquisition using an unsupervised parser”. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*. CoNLL ’09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 48–56. – Sets of lexical items sharing a significant aspect of their meaning (concepts) are fundamental for linguistics and NLP. Unsupervised concept acquisition algorithms have been shown to produce good results, and are preferable over manual preparation of concept resources, which is labor intensive, error prone and somewhat arbitrary. Some existing concept mining methods utilize supervised language-specific modules such as POS taggers and computationally intensive parsers. In this paper we present an efficient fully unsupervised concept acquisition algorithm that uses syntactic information obtained from a fully unsupervised parser. Our algorithm incorporates the bracketings induced by the parser into the meta-patterns used by a symmetric patterns and graph-based concept discovery algorithm. We evaluate our algorithm on very large corpora in English and Russian, using both human judgments and WordNet-based evaluation. Using similar settings as the leading fully unsupervised previous work, we show a significant improvement in concept quality and in the extraction of multiword expressions. Our method is the first to use fully unsupervised parsing for unsupervised concept discovery, and requires no language-specific tools or pattern/word seeds.

[3] Dan Klein. “The Unsupervised Learning of Natural Language Structure”. PhD thesis. Stanford University, Mar. 2005. – There is precisely one complete language processing system to date: the human brain. Though there is debate on how much built-in bias human learners might have, we definitely acquire language in a primarily unsupervised fashion. On the other hand, computational approaches to language processing are almost exclusively supervised, relying on hand-labeled corpora for training. This reliance is largely due to unsupervised approaches having repeatedly exhibited discouraging performance. In particular, the problem of learning syntax (grammar) from completely unannotated text has received a great deal of attention for well over a decade, with little in the way of positive results. We argue that previous methods for this task have generally underperformed because of the representations they used. Overly complex models are easily distracted by non-syntactic correlations (such as topical associations), while overly simple models aren’t rich enough to capture important first-order properties of language (such as directionality, adjacency, and valence). In this work, we describe several syntactic representations and associated probabilistic models which are designed to capture the basic character of natural language syntax as directly as possible. First, we examine a nested, distributionalmethod which induces bracketed tree structures. Second, we examine a dependency model which induces word-to-word dependency structures. Finally, we demonstrate that these two models perform better in combination than they do alone. With these representations, high-quality analyses can be learned from surprisingly little text, with no labeled examples, in several languages (we show experiments with English, German, and Chinese). Our results show above-baseline performance in unsupervised parsing in each of these languages. Grammar induction methods are useful since parsed corpora exist for only a small number of languages. More generally, most high-level NLP tasks, such as machine translation and question-answering, lack richly annotated corpora, making unsupervised methods extremely appealing even for common languages like English. Finally, while the models in this work are not intended to be cognitively plausible, their effectiveness can inform the investigation of what biases are or are not needed in the human acquisition of language.

[4] Roberto Navigli. “Word sense disambiguation: A survey”. In: *ACM Comput. Surv.* 41.2 (Feb. 2009), 10:1–10:69. – Word sense disambiguation (WSD) is the ability to identify the meaning

of words in context in a computational manner. WSD is considered an AI-complete problem, that is, a task whose solution is at least as hard as the most difficult problems in artificial intelligence. We introduce the reader to the motivations for solving the ambiguity of words and provide a description of the task. We overview supervised, unsupervised, and knowledge-based approaches. The assessment of WSD systems is discussed in the context of the Senseval/Semeval campaigns, aiming at the objective evaluation of systems participating in several different disambiguation tasks. Finally, applications, open problems, and future directions are discussed.

[5] L. Venkata Subramaniam et al. “A survey of types of text noise and techniques to handle noisy text”. In: *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*. AND ’09. New York, NY, USA: ACM, 2009, pp. 115–122. – Often, in the real world noise is ubiquitous in text communications. Text produced by processing signals intended for human use are often noisy for automated computer processing. Automatic speech recognition, optical character recognition and machine translation all introduce processing noise. Also digital text produced in informal settings such as online chat, SMS, emails, message boards, newsgroups, blogs, wikis and web pages contain considerable noise. In this paper, we present a survey of the existing measures for noise in text. We also cover application areas that ingest this noisy text for various tasks like Information Retrieval and Information Extraction.

[6] Kuansan Wang, Christopher Thrasher, and Bo-June Paul Hsu. “Web scale NLP: a case study on url word breaking”. In: *Proceedings of the 20th international conference on World wide web*. WWW ’11. New York, NY, USA: ACM, 2011, pp. 357–366. – This paper uses the URL word breaking task as an example to elaborate what we identify as crucial in designing statistical natural language processing (NLP) algorithms for Web scale applications: (1) rudimentary multilingual capabilities to cope with the global nature of the Web, (2) multi-style modeling to handle diverse language styles seen in the Web contents, (3) fast adaptation to keep pace with the dynamic changes of the Web, (4) minimal heuristic assumptions for generalizability and robustness, and (5) possibilities of efficient implementations and minimal manual efforts for processing massive amount of data at a reasonable cost. We first show that the state-of-the-art word breaking techniques can be unified and generalized under the Bayesian minimum risk (BMR) framework that, using a Web scale N-gram, can meet the first three requirements. We discuss how the existing techniques can be viewed as introducing additional assumptions to the basic BMR framework, and describe a generic yet efficient implementation called word synchronous beam search. Testing the framework and its implementation on a series of large scale experiments reveals the following. First, the language style used to build the model plays a critical role in the word breaking task, and the most suitable for the URL word breaking task appears to be that of the document title where the best performance is obtained. Models created from other language styles, such as from document body, anchor text, and even queries, exhibit varying degrees of mismatch. Although all styles benefit from increasing modeling power which, in our experiments, corresponds to the use of a higher order N-gram, the gain is most recognizable for the title model. The heuristics proposed by the prior arts do contribute to the word breaking performance for mismatched or less powerful models, but are less effective and, in many cases, lead to poorer performance than the matched model with minimal assumptions. For the matched model based on document titles, an accuracy rate of 97.18% can already be achieved using simple trigram without any heuristics.

2.6 Information Extraction

These works discuss approaches and algorithms for extracting information from natural language.

Annotated Bibliography

[1] Hisham Assal et al. “Partnering enhanced-NLP with semantic analysis in support of information extraction”. In: *Ontology-Driven Software Engineering*. ODiSE’10. New York, NY, USA: ACM, 2010, 9:1–9:7.

[2] Peter Clark and Phil Harrison. “Boeing’s NLP system and the challenges of semantic representation”. In: *Proceedings of the 2008 Conference on Semantics in Text Processing*. STEP ’08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 263–276. – We describe Boeing’s NLP system, BLUE, comprising a pipeline of a parser, a logical form (LF) generator, an initial logic generator, and further processing modules. The initial logic generator produces logic whose structure closely mirrors the structure of the original text. The subsequent processing modules then perform, with somewhat limited scope, additional transformations to convert this into a more usable representation with respect to a specific target ontology, better able to support inference. Generating a semantic representation is challenging, due to the wide variety of semantic phenomena which can occur in text. We identify seventeen such phenomena which occurred in the STEP 2008 "shared task" texts, comment on BLUE’s ability to handle them or otherwise, and discuss the more general question of what exactly constitutes a "semantic representation", arguing that a spectrum of interpretations exist.

[3] Lise Getoor and Ashwin Machanavajjhala. “Entity resolution: theory, practice & open challenges”. In: *Proc. VLDB Endow.* 5.12 (Aug. 2012), pp. 2018–2019. – This tutorial brings together perspectives on ER from a variety of fields, including databases, machine learning, natural language processing and information retrieval, to provide, in one setting, a survey of a large body of work. We discuss both the practical aspects and theoretical underpinnings of ER. We describe existing solutions, current challenges, and open research problems.

[4] Jiafeng Guo et al. “Named entity recognition in query”. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. SIGIR ’09. New York, NY, USA: ACM, 2009, pp. 267–274. – This paper addresses the problem of Named Entity Recognition in Query (NERQ), which involves detection of the named entity in a given query and classification of the named entity into predefined classes. NERQ is potentially useful in many applications in web search. The paper proposes taking a probabilistic approach to the task using query log data and Latent Dirichlet Allocation. We consider contexts of a named entity (i.e., the remainders of the named entity in queries) as words of a document, and classes of the named entity as topics. The topic model is constructed by a novel and general learning method referred to as WS-LDA (Weakly Supervised Latent Dirichlet Allocation), which employs weakly supervised learning (rather than unsupervised learning) using partially labeled seed entities. Experimental results show that the proposed method based on WS-LDA can accurately perform NERQ, and outperform the baseline methods.

[5] Alpa Jain and Panagiotis G. Ipeirotis. “A quality-aware optimizer for information extraction”. In: *ACM Trans. Database Syst.* 34.1 (Apr. 2009), 5:1–5:48. – A large amount of structured information is buried in unstructured text. Information extraction systems can extract structured relations from the documents and enable sophisticated, SQL-like queries over unstructured text. Information extraction systems are not perfect and their output has imperfect preci-

sion and recall (i.e., contains spurious tuples and misses good tuples). Typically, an extraction system has a set of parameters that can be used as "knobs" to tune the system to be either precision- or recall-oriented. Furthermore, the choice of documents processed by the extraction system also affects the quality of the extracted relation. So far, estimating the output quality of an information extraction task has been an ad hoc procedure, based mainly on heuristics. In this article, we show how to use Receiver Operating Characteristic (ROC) curves to estimate the extraction quality in a statistically robust way and show how to use ROC analysis to select the extraction parameters in a principled manner. Furthermore, we present analytic models that reveal how different document retrieval strategies affect the quality of the extracted relation. Finally, we present our maximum likelihood approach for estimating, on the fly, the parameters required by our analytic models to predict the runtime and the output quality of each execution plan. Our experimental evaluation demonstrates that our optimization approach predicts accurately the output quality and selects the fastest execution plan that satisfies the output quality restrictions.

[6] Patrick Watrin and Thomas François. "An n-gram frequency database reference to handle MWE extraction in NLP applications". In: *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*. MWE '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 83–91. – The identification and extraction of Multiword Expressions (MWEs) currently deliver satisfactory results. However, the integration of these results into a wider application remains an issue. This is mainly due to the fact that the association measures (AMs) used to detect MWEs require a critical amount of data and that the MWE dictionaries cannot account for all the lexical and syntactic variations inherent in MWEs. In this study, we use an alternative technique to overcome these limitations. It consists in defining an n-gram frequency data-base that can be used to compute AMs on-the-fly, allowing the extraction procedure to efficiently process all the MWEs in a text, even if they have not been previously observed.

[7] Daya C. Wimalasuriya and Dejing Dou. "Components for information extraction: ontology-based information extractors and generic platforms". In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. CIKM '10. New York, NY, USA: ACM, 2010, pp. 9–18. – Information Extraction (IE) has existed as a field for several decades and has produced some impressive systems in the recent past. Despite its success, widespread usage and commercialization remain elusive goals for this field. We identify the lack of effective mechanisms for reuse as one major reason behind this situation. Here, we mean not only the reuse of the same IE technique in different situations but also the reuse of information related to the application of IE techniques (e.g., features used for classification). We have developed a comprehensive component-based approach for information extraction that promotes reuse to address this situation. We designed this approach starting from our previous work on the use of multiple ontologies in information extraction [24]. The key ideas of our approach are "information extractors," which are components of an IE system that make extractions with respect to particular components of an ontology and "platforms for IE," which are domain and corpus independent implementations of IE techniques. A case study has shown that this component-based approach can be successfully applied in practical situations.

2.7 Machine Translation

These works discuss approaches and algorithms for machine translation of natural languages.

Annotated Bibliography

[1] Chris Callison-Burch et al. “Findings of the 2009 workshop on statistical machine translation”. In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*. StatMT ’09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 1–28.

[2] Carmen Heger et al. “The RWTH Aachen machine translation system for WMT 2010”. In: *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. WMT ’10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 93–97. – In this paper we describe the statistical machine translation system of the RWTH Aachen University developed for the translation task of the Fifth Workshop on Statistical Machine Translation. State-of-the-art phrase-based and hierarchical statistical MT systems are augmented with appropriate morpho-syntactic enhancements, as well as alternative phrase training methods and extended lexicon models. For some tasks, a system combination of the best systems was used to generate a final hypothesis. We participated in the constrained condition of German-English and French-English in each translation direction.

[3] Rabih Zbib et al. “Decision trees for lexical smoothing in statistical machine translation”. In: *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. WMT ’10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 428–437. – We present a method for incorporating arbitrary context-informed word attributes into statistical machine translation by clustering attribute-qualified source words, and smoothing their word translation probabilities using binary decision trees. We describe two ways in which the decision trees are used in machine translation: by using the attribute-qualified source word clusters directly, or by using attribute-dependent lexical translation probabilities that are obtained from the trees, as a lexical smoothing feature in the decoder model. We present experiments using Arabic-to-English newswire data, and using Arabic diacritics and part-of-speech as source word attributes, and show that the proposed method improves on a state-of-the-art translation system.

2.8 Language Generation

These works discuss approaches and algorithms for generating natural language.

Annotated Bibliography

[1] Dana Dannélls. “Applying Semantic Frame Theory to Automate Natural Language Template Generation from Ontology Statements”. In: *Proceedings of the 6th International Natural Language Generation Conference*. INLG ’10. Trim, Co. Meath, Ireland: Association for Computational Linguistics, 2010, pp. 179–183. URL: http://dl.acm.org/citation.cfm?id=1873738.1873762. – Today there exist a growing number of framenet-like resources offering semantic and syntactic phrase specifications that can be exploited by natural language generation systems. In this paper we present on-going work that provides a starting point for exploiting framenet information for multilingual natural language generation. We describe the kind of information offered by modern computational lexical resources and discuss how template-based generation systems can benefit from them.

[2] Verena Rieser and Oliver Lemon. “Empirical Methods in Natural Language Generation”. In: ed. by Emiel Krahmer and Mariët Theune. Berlin, Heidelberg: Springer-Verlag, 2010. Chap. Natural Language Generation As Planning Under Uncertainty for Spoken Dialogue Systems, pp. 105–120. ISBN: 3-642-15572-3, 978-3-642-15572-7. URL: http://dl.acm.org/citation.cfm?id=1880370.1880378. – We present and evaluate a new model for Natural Language Generation (NLG) in Spoken Dialogue Systems, based on statistical planning, given noisy feedback from the current generation context (e.g. a user and a surface realiser). The model is adaptive and incremental at the turn level, and optimises NLG actions with respect to a data-driven objective function. We study its use in a standard NLG problem: how to present information (in this case a set of search results) to users, given the complex trade-offs between utterance length, amount of information conveyed, and cognitive load. We set these trade-offs in an objective function by analysing existing match data. We then train a NLG policy using Reinforcement Learning (RL), which adapts its behaviour to noisy feedback from the current generation context. This policy is compared to several baselines derived from previous work in this area. The learned policy significantly outperforms all the prior approaches.

2.9 Dialog Systems

These works discuss approaches and algorithms for developing dialog systems. More to come.

Annotated Bibliography

[1] David Griol et al. “A Domain-independent Statistical Methodology for Dialog Management in Spoken Dialog Systems”. In: *Comput. Speech Lang.* 28.3 (May 2014), pp. 743–768. ISSN: 0885-2308. DOI: 10.1016/j.csl.2013.09.002. URL: http://dx.doi.org/10.1016/j.csl.2013.09.002. – This paper proposes a domain-independent statistical methodology to develop dialog managers for spoken dialog systems. Our methodology employs a data-driven classification procedure to generate abstract representations of system turns taking into account the previous history of the dialog. A statistical framework is also introduced for the development and evaluation of dialog systems created using the methodology, which is based on a dialog simulation technique. The benefits and flexibility of the proposed methodology have been validated by developing statistical dialog managers for four spoken dialog systems of different complexity, designed for different languages (English, Italian, and Spanish) and application domains (from transactional to problem-solving tasks). The evaluation results show that the proposed methodology allows rapid development of new dialog managers as well as to explore new dialog strategies, which permit developing new enhanced versions of already existing systems.

2.10 Question-answering Systems

These works discuss approaches and algorithms for question-answering systems.

Annotated Bibliography

[1] Ashton Anderson et al. “Discovering value from community activity on focused question answering sites: a case study of stack overflow”. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD ’12. New York, NY, USA: ACM, 2012, pp. 850–858.

[2] Andrea Andrenucci and Eriks Sneiders. “Automated Question Answering: Review of the Main Approaches”. In: *International Conference on Information Technology and Applications* 1 (2005), pp. 514–519. – Automated Question-Answering aims at delivering concise information that contains answers to user questions. This paper reviews and compares three main question-answering approaches based on Natural Language Processing, Information Retrieval, and question templates, eliciting their differences and the context of application that best suits each of them.

[3] Anna Shtok et al. “Learning from the past: answering new questions with past answers”. In: *Proceedings of the 21st international conference on World Wide Web*. WWW ’12. New York, NY, USA: ACM, 2012, pp. 759–768. – Community-based Question Answering sites, such as Yahoo! Answers or Baidu Zhidao, allow users to get answers to complex, detailed and personal questions from other users. However, since answering a question depends on the ability and willingness of users to address the asker’s needs, a significant fraction of the questions remain unanswered. We measured that in Yahoo! Answers, this fraction represents 15% of all incoming English questions. At the same time, we discovered that around 25% of questions in certain categories are recurrent, at least at the question-title level, over a period of one year. We attempt to reduce the rate of unanswered questions in Yahoo! Answers by reusing the large repository of past resolved questions, openly available on the site. More specifically, we estimate the probability whether certain new questions can be satisfactorily answered by a best answer from the past, using a statistical model specifically trained for this task. We leverage concepts and methods from query-performance prediction and natural language processing in order to extract a wide range of features for our model. The key challenge here is to achieve a level of quality similar to the one provided by the best human answerers. We evaluated our algorithm on offline data extracted from Yahoo! Answers, but more interestingly, also on online data by using three “live” answering robots that automatically provide past answers to new questions when a certain degree of confidence is reached. We report the success rate of these robots in three active Yahoo! Answers categories in terms of both accuracy, coverage and askers’ satisfaction. This work presents a first attempt, to the best of our knowledge, of automatic question answering to questions of social nature, by reusing past answers of high quality.

2.11 Question-generation Systems

These works discuss approaches and algorithms for automatic generation of test questions for learning and assessment.

Annotated Bibliography

[1] Manish Agarwal and Prashanth Mannem. “Automatic gap-fill question generation from text books”. In: *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*. IUNLPBEA ’11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 56–64. – In this paper, we present an automatic question generation system that can generate gap-fill questions for content in a document. Gap-fill questions are fill-in-the-blank questions with multiple choices (one correct answer and three distractors) provided. The system finds the informative sentences from the document and generates gap-fill questions from them by first blanking keys from the sentences and then determining the distractors for these keys. Syntactic and lexical features are used in this process without relying on any external re-

source apart from the information in the document. We evaluated our system on two chapters of a standard biology textbook and presented the results.

[2] Kristy Elizabeth Boyer and Paul Piwek. *Proceedings of QG2010: The Third Workshop on Question Generation*. http://questiongeneration.org. 2010.

[3] Michael Heilman and Noah A. Smith. "Good Question! Statistical Ranking for Question Generation". In: *Annual Conference of the North American Chapter of the ACL*. Association for Computational Linguistics. 2010, pp. 609–617. – We address the challenge of automatically generating questions from reading materials for educational practice and assessment. Our approach is to overgenerate questions, then rank them. We use manually written rules to perform a sequence of general purpose syntactic transformations (e.g., subject-auxiliary inversion) to turn declarative sentences into questions. These questions are then ranked by a logistic regression model trained on a small, tailored dataset consisting of labeled output from our system. Experimental results show that ranking nearly doubles the percentage of questions rated as acceptable by annotators, from 27% of all questions to 52% of the top ranked 20% of questions.

[4] William J. Therrien and Charles Hughes. "Comparison of Repeated Reading and Question Generation on Students' Reading Fluency and Comprehension". In: *Learning Disabilities: A Contemporary Journal* 6.1 (2008), pp. 1–16. – This study was conducted to ascertain if repeated reading or question generation was more effective at improving reading fluency and comprehension of fourth- through sixth-grade students with learning disabilities or reading problems. Adult tutors trained by the investigator conducted the interventions. Instructional components and training within each of the interventions were based on best practices reported in the literature. Repeated reading consisted of students rereading passages aloud until they reached a performance criterion. Question generation consisted of students reading passages purposefully in an attempt to adapt and answer story structure prompts. The results of the study indicate that (a) repeated reading improves students' fluency on passages that are reread and (b) when reading instructional-level material, repeated reading is more effective at improving factual comprehension than question generation.

2.12 Discourse Understanding

These works discuss approaches and algorithms for developing discourse understanding systems.

Annotated Bibliography

[1] Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. "Towards semi-supervised classification of discourse relations using feature correlations". In: *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. SIGDIAL '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 55–58. – Two of the main corpora available for training discourse relation classifiers are the RST Discourse Treebank (RST-DT) and the Penn Discourse Treebank (PDTB), which are both based on the Wall Street Journal corpus. Most recent work using discourse relation classifiers have employed fully-supervised methods on these corpora. However, certain discourse relations have little labeled data, causing low classification performance for their associated classes. In this paper, we attempt to tackle this problem by employing a semi-supervised method for discourse relation classification. The proposed method is based on the analysis of feature cooccurrences in unlabeled data. This information is then used as a basis to extend the feature vectors during training.

The proposed method is evaluated on both RST-DT and PDTB, where it significantly outperformed baseline classifiers. We believe that the proposed method is a first step towards improving classification performance, particularly for discourse relations lacking annotated data.

2.13 Clustering

These works discuss approaches and algorithms for clustering natural language text.

Annotated Bibliography

[1] Roberto Navigli et al. "Two birds with one stone: learning semantic models for text categorization and word sense disambiguation". In: *Proceedings of the 20th ACM international conference on Information and knowledge management*. CIKM '11. New York, NY, USA: ACM, 2011, pp. 2317–2320. – In this paper we present a novel approach to learning semantic models for multiple domains, which we use to categorize Wikipedia pages and to perform domain Word Sense Disambiguation (WSD). In order to learn a semantic model for each domain we first extract relevant terms from the texts in the domain and then use these terms to initialize a random walk over the WordNet graph. Given an input text, we check the semantic models, choose the appropriate domain for that text and use the best-matching model to perform WSD. Our results show considerable improvements on text categorization and domain WSD tasks.

2.14 Summarization

These works discuss approaches and algorithms for summarizing single and multiple documents.

Annotated Bibliography

[1] Annie Louis, Aravind Joshi, and Ani Nenkova. "Discourse indicators for content selection in summarization". In: *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. SIGDIAL '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 147–156. – We present analyses aimed at eliciting which specific aspects of discourse provide the strongest indication for text importance. In the context of content selection for single document summarization of news, we examine the benefits of both the graph structure of text provided by discourse relations and the semantic sense of these relations. We find that structure information is the most robust indicator of importance. Semantic sense only provides constraints on content selection but is not indicative of important content by itself. However, sense features complement structure information and lead to improved performance. Further, both types of discourse information prove complementary to non-discourse features. While our results establish the usefulness of discourse features, we also find that lexical overlap provides a simple and cheap alternative to discourse for computing text structure with comparable performance for the task of content selection.

[2] Nadav Rotem. *Open Text Summarizer*. http://libots.sourceforge.net/. Sept. 2012.

2.15 NLP Applications

These works discuss practical applications of NLP.

Annotated Bibliography

[1] Ravi Kant, Srinivasan H. Sengamedu, and Krishnan S. Kumar. “Comment spam detection by sequence mining”. In: *Proceedings of the fifth ACM international conference on Web search and data mining*. WSDM ’12. New York, NY, USA: ACM, 2012, pp. 183–192. – Comments are supported by several web sites to increase user participation. Users can usually comment on a variety of media types - photos, videos, news articles, blogs, etc. Comment spam is one of the biggest challenges facing this feature. The traditional approach to combat spam is to train classifiers using various machine learning techniques. Since the commonly used classifiers work on the entire comment text, it is easy to mislead them by embedding spam content in good content. In this paper, we make several contributions towards comment spam detection. (1) We propose a new framework for spam detection that is immune to embed attacks. We characterize spam by a set of frequently occurring sequential patterns. (2) We introduce a variant (called min-closed) of the frequent closed sequence mining problem that succinctly captures all the frequently occurring patterns. We prove as well as experimentally show that the set of min-closed sequences is an order of magnitude smaller than the set of closed sequences and yet has exactly the same coverage. (3) We describe MCPRISM, extension of the recently published PRISM algorithm that effectively mines min-closed sequences, using prime encoding. In the process, we solve the open problem of using the prime-encoding technique to speed up traditional closed sequence mining. (4) We finally need to whittle down the set of frequent subsequences to a small set without sacrificing coverage. This problem is NP-Hard but we show that the coverage function is submodular and hence the greedy heuristic gives a fast algorithm that is close to optimal. We then describe the experiments that were carried out on a large real world comment data and the publicly available Gazelle dataset. (1) We show that nearly 80% of spam on real world data can be effectively captured by the mined sequences at very low false positive rates. (2) The sequences mined are highly discriminative. (3) On Gazelle data, the proposed algorithmic enhancements are faster by at least by a factor and by an order of magnitude on the larger comment dataset.

[2] Raymond Y. K. Lau et al. “Text mining and probabilistic language modeling for online review spam detection”. In: *ACM Trans. Manage. Inf. Syst.* 2.4 (Jan. 2012), 25:1–25:30. – In the era of Web 2.0, huge volumes of consumer reviews are posted to the Internet every day. Manual approaches to detecting and analyzing fake reviews (i.e., spam) are not practical due to the problem of information overload. However, the design and development of automated methods of detecting fake reviews is a challenging research problem. The main reason is that fake reviews are specifically composed to mislead readers, so they may appear the same as legitimate reviews (i.e., ham). As a result, discriminatory features that would enable individual reviews to be classified as spam or ham may not be available. Guided by the design science research methodology, the main contribution of this study is the design and instantiation of novel computational models for detecting fake reviews. In particular, a novel text mining model is developed and integrated into a semantic language model for the detection of untruthful reviews. The models are then evaluated based on a real-world dataset collected from amazon.com. The results of our experiments confirm that the proposed models outperform other well-known baseline models in detecting fake reviews. To the best of our knowledge, the work discussed in this article represents the first successful attempt to apply text mining methods and semantic language models to the detection of fake consumer reviews. A managerial implication of our research is that firms can apply our design artifacts to monitor online consumer reviews to develop effective marketing or product design strategies based on genuine consumer feedback posted to the Internet.

[3] Florian Laws, Christian Scheible, and Hinrich Schütze. “Active learning with Amazon Mechanical Turk”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP ’11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1546–1556. – Supervised classification needs large amounts of annotated training data that is expensive to create. Two approaches that reduce the cost of annotation are active learning and crowdsourcing. However, these two approaches have not been combined successfully to date. We evaluate the utility of active learning in crowdsourcing on two tasks, named entity recognition and sentiment detection, and show that active learning outperforms random selection of annotation examples in a noisy crowdsourcing scenario.

[4] Gondy Leroy and James E. Endicott. “Combining NLP with evidence-based methods to find text metrics related to perceived and actual text difficulty”. In: *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. IHI ’12. New York, NY, USA: ACM, 2012, pp. 749–754. – Measuring text difficulty is prevalent in health informatics since it is useful for information personalization and optimization. Unfortunately, it is uncertain how best to compute difficulty so that it relates to reader understanding. We aim to create computational, evidence-based metrics of perceived and actual text difficulty. We start with a corpus analysis to identify candidate metrics which are further tested in user studies. Our corpus contains blogs and journal articles (N=1,073) representing easy and difficult text. Using natural language processing, we calculated base grammatical and semantic metrics, constructed new composite metrics (noun phrase complexity and semantic familiarity), and measured the commonly used Flesch-Kincaid grade level. The metrics differed significantly between document types. Nouns were more prevalent but less familiar in difficult text; verbs and function words were more prevalent in easy text. Noun phrase complexity was lower, semantic familiarity was higher and grade levels were lower in easy text. Then, all metrics were tested for their relation to perceived and actual difficulty using follow-up analyses of two user studies conducted earlier. Base metrics and noun phrase complexity correlated significantly with perceived difficulty and could help explain actual difficulty.

[5] alias-i.com. *LingPipe – A toolkit for processing text using computational linguistics*. http://alias-i.com/lingpipe/. Sept. 2012.

[6] Amber McKenzie et al. “Information extraction from helicopter maintenance records as a springboard for the future of maintenance text analysis”. In: *Proceedings of the 23rd international conference on Industrial engineering and other applications of applied intelligent systems - Volume Part I*. IEA/AIE’10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 590–600. – This paper introduces a novel application of information extraction techniques to extract data from helicopter maintenance records to populate a database. The goals of the research are to preprocess the text-based data for further use in data mining efforts and to develop a system to provide a rough analysis of generic maintenance records to facilitate in the development of training corpora for use in machine-learning for more refined information extraction system design. The Natural Language Toolkit was used to implement partial parsing of text by way of hierarchical chunking of the text. The system was targeted towards inspection descriptions and succeeded in extracting the inspection code, description of the part/action, and date/time information with 80.7% recall and 89.9% precision.

[7] Rada Mihalcea and Dragomir Radev. *Graph-based Natural Language Processing and Information Retrieval*. Cambridge University Press, 2011. – Graph theory and the fields of natural language processing and information retrieval are well-studied disciplines. Traditionally, these areas have been perceived as distinct, with different algorithms, different applications, and

different potential end-users. However, recent research has shown that these disciplines are intimately connected, with a large variety of natural language processing and information retrieval applications finding efficient solutions within graph-theoretical frameworks. This book extensively covers the use of graph-based algorithms for natural language processing and information retrieval. It brings together topics as diverse as lexical semantics, text summarization, text mining, ontology construction, text classification, and information retrieval, which are connected by the common underlying theme of the use of graph-theoretical methods for text and information processing tasks. Readers will come away with a firm understanding of the major methods and applications in natural language processing and information retrieval that rely on graph-based representations and algorithms.

[8] Georgios Paltoglou and Mike Thelwall. "Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media". In: *ACM Trans. Intell. Syst. Technol.* 3.4 (Sept. 2012), 66:1–66:19. – Sentiment analysis is a growing area of research with significant applications in both industry and academia. Most of the proposed solutions are centered around supervised, machine learning approaches and review-oriented datasets. In this article, we focus on the more common informal textual communication on the Web, such as online discussions, tweets and social network comments and propose an intuitive, less domain-specific, unsupervised, lexicon-based approach that estimates the level of emotional intensity contained in text in order to make a prediction. Our approach can be applied to, and is tested in, two different but complementary contexts: subjectivity detection and polarity classification. Extensive experiments were carried on three real-world datasets, extracted from online social Web sites and annotated by human evaluators, against state-of-the-art supervised approaches. The results demonstrate that the proposed algorithm, even though unsupervised, outperforms machine learning solutions in the majority of cases, overall presenting a very robust and reliable solution for sentiment analysis of informal communication on the Web.

[9] Matthew A. Russell. *Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites*. O'Reilly Media, 2011. – Want to tap the tremendous amount of valuable social data in Facebook, Twitter, LinkedIn, and Google+? This refreshed edition helps you discover who's making connections with social media, what they're talking about, and where they're located. You'll learn how to combine social web data, analysis techniques, and visualization to find what you've been looking for in the social haystack — as well as useful information you didn't know existed. Each standalone chapter introduces techniques for mining data in different areas of the social Web, including blogs and email. All you need to get started is a programming background and a willingness to learn basic Python tools.

[10] Marta Sabou, Kalina Bontcheva, and Arno Scharl. "Crowdsourcing research opportunities: lessons from natural language processing". In: *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*. i-KNOW '12. New York, NY, USA: ACM, 2012, 17:1–17:8. – Although the field has led to promising early results, the use of crowdsourcing as an integral part of science projects is still regarded with skepticism by some, largely due to a lack of awareness of the opportunities and implications of utilizing these new techniques. We address this lack of awareness, firstly by highlighting the positive impacts that crowdsourcing has had on Natural Language Processing research. Secondly, we discuss the challenges of more complex methodologies, quality control, and the necessity to deal with ethical issues. We conclude with future trends and opportunities of crowdsourcing for science, including its potential for disseminating results, making science more accessible, and enriching educational programs.

[11] Dipti Misra Sharma. “On the role of NLP in linguistics”. In: *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*. NLPLING ’10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 18–21. – This paper summarizes some of the applications of NLP techniques in various linguistic sub-fields, and presents a few examples that call for a deeper engagement between the two fields.

3 Peer Feedback and Revision

Remember that the goal of peer feedback is not only to point out shortcomings in the paper, but more importantly to suggest ways to improving the quality of the paper (writing style, flow, grammar, technical accuracy).
Name of the person who provided feedback: Peer’s name goes here.

Following are the specific suggestions made by the peer:

1. suggestion 1 description
2. suggestion 2 description
3. suggestion n description

In the following, discuss how the peer’s suggestions for improvement are incorporated into the document.

1. Explain how peer’s suggestion 1 is addressed
2. Explain how peer’s suggestion 2 is addressed
3. Explain how peer’s suggestion n is addressed

4 Instructor Feedback and Revision

This section should list feedback provided by the instructor and discuss how you have incorporated the feedback in improving the paper.

Following are the specific suggestions made by the instructor:

1. suggestion 1 description
2. suggestion 2 description
3. suggestion n description

In the following, discuss how the instructor’s suggestions for improvement are incorporated into the document.

1. Explain how instructor’s suggestion 1 is addressed
2. Explain how instructor’s suggestion 2 is addressed
3. Explain how instructor’s suggestion n is addressed

5 Self-assessment

This assignment will receive zero points if the required LaTeX template is not used.

Rubric line item	Max possible points	Earned points
Read the paper titled "Scaffolding Beginning Research Students Using Open Source Tools" and used its ideas in completing this assignment.	10	nn
Consulted ACM Digital Library, IEEE Computer Society Digital Library, SiteSeerX, MENDELEY, Google Scholar, Ultimate Research Assistant, and other online sources for identifying and collecting bibliography sources.	10	nn
Several facets of the assigned topic has been identified and the annotated bibliography is organized according to these facets.	20	nn
For each bibliography source, included are: summary in the form of abstract, assessment, and reflection.	30	nn
LaTeX and BibTeX files follow standard conventions and are easy to read and maintain.	5	nn
Peer feedback is solicited and incorporated.	5	nn
Instructor feedback is solicited and incorporated.	5	nn
Writing is of professional quality and is free from grammatical and syntactic errors	10	nn
Self-assessment is performed.	5	nn
Total points	100	nn