ACL-IJCNLP 2015

The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing

Proceedings of the Conference
Tutorial Abstracts

July 26-31, 2015

Order print-on-demand copies from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-941643-76-1

Introduction

This volume contains the abstracts of the ACL-IJCNLP 2015 tutorials. This year we had a joint call-for-tutorials, coordinated with the NAACL and EMNLP co-chairs (6 co-chairs in total). We received 32 high-quality proposals, and it was a difficult task to make a final selection. The six co-chairs applied the following criteria for evaluation: relevance to ACL community, quality of instructor, quality of proposal, own estimate of attendance, newly emerging area, and being an introductions into related fields. The tutorials were then assigned to venues, trying to respect proposers' preferences and to balance topics across venues. In the end we accepted eight tutorials for ACL-IJCNLP. All eight of these are organized as half-day tutorials.

We are very grateful to Yang Liu and Thamar Solorio (NAACL tutorial chairs), Maggie Li and Khalil Sima'an (EMNLP tutorial chairs), Le Sun and Yang Liu (local chairs), Wanxiang Che and Guodong Zhou (publication chairs), Yuji Matsumoto (general chair), and of course Priscilla Rasmussen, for various kinds of help, advice and assistance offered during the process of putting the tutorial programme and materials together. Most importantly, we would like to thank the tutorial presenters for the time and effort in preparing and presenting the tutorials.

We hope you will enjoy the tutorials!

ACL-IJCNLP 2015 Tutorial Chairs
Eneko Agirre, University of the Basque Country
Kevin Duh, Nara Institute of Science and Technology

Tutorial chairs:

Eneko Agirre, University of the Basque Country
Kevin Duh, Nara Institute of Science and Technology

Table of Contents

Conference Program

Sunday, July 26, 2015

Morning Session

09:00-12:30 *Successful Data Mining Methods for NLP*
Jiawei Han, Heng Ji and Yizhou Sun

09:00-12:30 *Structured Belief Propagation for NLP*
Matthew R. Gormley and Jason Eisner

09:00-12:30 *Sentiment and Belief: How to Think about, Represent, and Annotate Private States*
Owen Rambow and Janyce Wiebe

09:00-12:30 *Corpus Patterns for Semantic Processing*
Octavian Popescu, Patrick Hanks, Elisabetta Jezek and Daisuke Kawahara

Afternoon Session

14:00-17:30 *Matrix and Tensor Factorization Methods for Natural Language Processing*
Guillaume Bouchard, Jason Naradowsky, Sebastian Riedel, Tim Rocktäschel and Andreas Vlachos

14:00-17:30 *Scalable Large-Margin Structured Learning: Theory and Algorithms*
Liang Huang and Kai Zhao

14:00-17:30 *Detecting Deceptive Opinion Spam using Linguistics, Behavioral and Statistical Modeling*
Arjun Mukherjee

14:00-17:30 *What You Need to Know about Chinese for Chinese Language Processing*
Chu-Ren Huang

Successful Data Mining Methods for NLP

Jiawei Han
Dept. of Computer Science
Univ. of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
hanj@cs.uiuc.edu

Heng Ji
Computer Science Dept.
Rensselaer Polytechnic Institute
Troy, NY 12180, USA
jih@rpi.edu

Yizhou Sun
College of Computer and Information Science
Northeastern University
Boston, MA 02115, USA
yzsun@ccs.neu.edu

1 Overview

Historically Natural Language Processing (NLP) focuses on unstructured data (speech and text) understanding while Data Mining (DM) mainly focuses on massive, structured or semi-structured datasets. The general research directions of these two fields also have followed different philosophies and principles. For example, NLP aims at deep understanding of individual words, phrases and sentences ("micro-level"), whereas DM aims to conduct a high-level understanding, discovery and synthesis of the most salient information from a large set of documents when working on text data ("macro-level"). But they share the same goal of distilling knowledge from data. In the past five years, these two areas have had intensive interactions and thus mutually enhanced each other through many successful text mining tasks. This positive progress mainly benefits from some innovative intermediate representations such as "heterogeneous information networks" [Han et al., 2010, Sun et al., 2012b].

However, successful collaborations between any two fields require substantial mutual understanding, patience and passion among researchers. Similar to the applications of machine learning techniques in NLP, there is usually a gap of at least several years between the creation of a new DM approach and its first successful application in NLP. More importantly, many DM approaches such as gSpan [Yan and Han, 2002] and RankClus [Sun et al., 2009a] have demonstrated their power on structured data. But they remain relatively unknown in the NLP community, even though there are many obvious potential applications. On the other hand, compared to DM, the NLP community has paid more attention to developing large-scale data annotations, resources, shared tasks which cover a wide range of multiple genres and multiple domains. NLP can also provide the basic building blocks for many DM tasks such as text cube construction [Tao et al., 2014]. Therefore in many scenarios, for the same approach the NLP experiment setting is often much closer to real-world applications than its DM counterpart.

We would like to share the experiences and lessons from our extensive inter-disciplinary collaborations in the past five years. The primary goal of this tutorial is to bridge the knowledge gap between these two fields and speed up the transition process. We will introduce two types of DM methods: (1). those state-of-the-art DM methods that have already been proven effective for NLP; and (2). some newly developed DM methods that we believe will fit into some specific NLP problems. In addition, we aim to suggest some new research directions in order to better marry these two areas and lead to more fruitful outcomes. The tutorial will thus be useful for researchers from both communities. We will try to provide a concise roadmap of recent perspectives and results, as well as point to the related DM software and resources, and NLP data sets that are available to both research communities.

2 Outline

We will focus on the following three perspectives.

2.1 Where do NLP and DM Meet

We will first pick up the tasks shown in Table 1 that have attracted interests from both NLP and DM, and give an overview of different solutions to these problems. We will compare their fundamental differences in terms of goals, theories, principles and methodologies.

Proceedings of the Tutorials of the 53rd Annual Meeting of the ACL and the 7th IJCNLP, pages 1–4,
Beijing, China, July 26-31, 2015.

Tasks	DM Methods	NLP Methods
Phrase mining / Chunking	Statistical pattern mining [El-Kishky et al., 2015; Danilevsky et al., 2014; Han et al., 2014]	Supervised chunking trained from Penn Treebank
Topic hierarchy / Taxonomy construction	Combine statistical pattern mining with information networks [Wang et al., 2014]	Lexical/Syntactic patterns (e.g., COLING2014 workshop on taxonomy construction)
Entity Linking	Graph alignment [Li et al., 2013]	TAC-KBP Entity Linking methods and Wikification
Relation discovery	Hierarchical clustering [Wang et al., 2012]	ACE relation extraction, bootstrapping
Sentiment Analysis	Pseudo-friendship network analysis [Deng et al., 2014]	Supervised methods based on linguistic resources

Table 1. Examples for Tasks Solved by Different NLP and DM Methods

2.2 Successful DM Methods Applied for NLP

Then we will focus on introducing a series of effective DM methods which have already been adopted for NLP applications. The most fruitful research line exploited Heterogeneous Information Networks [Tao et al., 2014; Sun et al., 2009ab, 2011, 2012ab, 2013, 2015]. For example, the meta-path concept and methodology [Sun et al., 2011] has been successfully used to address morph entity discovery and resolution [Huang et al., 2013] and Wikification [Huang et al., 2014]; the Co-HITS algorithm [Deng et al., 2009] was applied to solve multiple NLP problems including tweet ranking [Huang et al., 2012] and slot filling validation [Yu et al., 2014]. We will synthesize the important aspects learned from these successes.

2.3 New DM Methods Promising for NLP

Then we will introduce a wide range of new DM methods which we believe are promising to NLP. We will align the problems and solutions by categorizing their special characteristics from both the linguistic perspective and the mining perspective. One thread we will focus on is graph mining. We will recommend some effective graph pattern mining methods [Yan and Han, 2002&2003; Yan et al., 2008; Chen et al., 2010] and their potential applications in cross-document entity clustering and slot filling. Some recent DM methods can also be used to capture implicit textual cues which might be difficult to generalize using traditional syntactic analysis. For example, [Kim et al., 2011] developed a syntactic tree mining approach to predict authors from papers, which can be extended to more general stylistic analysis. We will carefully survey the major challenges and solutions that address these adoptions.

2.4 New Research Directions to Integrate NLP and DM

We will conclude with a discussion of some key new research directions to better integrate DM and NLP. What is the best framework for integration and joint inference? Is there an ideal common representation, or a layer between these two fields? Is Information Networks still the best intermediate step to accomplish the Language-to-Networks-to-Knowledge paradigm?

2.5 Resources

We will present an overview of related systems, demos, resources and data sets.

3 Tutorial Instructors

Jiawei Han is Abel Bliss Professor in the Department of Computer Science at the University of Illinois. He has been researching into data mining, information network analysis, and database systems, with over 600 publications. He served as the founding Editor-in-Chief of ACM Transactions on Knowledge Discovery from Data (TKDD). He has received ACM SIGKDD Innovation Award (2004), IEEE Computer Society Technical Achievement Award (2005), IEEE Computer Society W. Wallace McDowell Award (2009), and Daniel C. Drucker Eminent Faculty Award at UIUC (2011). He is a Fellow of ACM and a Fellow of IEEE. He is currently the Director of Information Network Academic Research Center (INARC) supported by the Network Science-Collaborative Technology Alliance (NS-CTA) program of U.S. Army Research Lab and

also the Director of KnowEnG, an NIH Center of Excellence in big data computing as part of NIH Big Data to Knowledge (BD2K) initiative. His co-authored textbook "Data Mining: Concepts and Techniques" (Morgan Kaufmann) has been adopted worldwide. He has delivered tutorials in many reputed international conferences, including WWW'14, SIGMOD'14 and KDD'14.

Heng Ji is Edward H. Hamilton Development Chair Associate Professor in Computer Science Department of Rensselaer Polytechnic Institute. She received "AI's 10 to Watch" Award in 2013, NSF CAREER award in 2009, Google Research Awards in 2009 and 2014 and IBM Watson Faculty Awards in 2012 and 2014. In the past five years she has done extensive collaborations with Prof. Jiawei Han and Prof. Yizhou Sun on applying data mining techniques to NLP problems and jointly published 15 papers, including a "Best of SDM2013" paper and a "Best of ICDM2013" paper. She has delivered tutorials at COLING2012, ACL2014 and NLPCC2014.

Yizhou Sun is an assistant professor in the College of Computer and Information Science of Northeastern University. She received her Ph.D. in Computer Science from the University of Illinois at Urbana Champaign in 2012. Her principal research interest is in mining information and social networks, and more generally in data mining, database systems, statistics, machine learning, information retrieval, and network science, with a focus on modeling novel problems and proposing scalable algorithms for large scale, real-world applications. Yizhou has over 60 publications in books, journals, and major conferences. Tutorials based on her thesis work on mining heterogeneous information networks have been given in several premier conferences, including EDBT 2009, SIGMOD 2010, KDD 2010, ICDE 2012, VLDB 2012, and ASONAM 2012. She received 2012 ACM SIGKDD Best Student Paper Award, 2013 ACM SIGKDD Doctoral Dissertation Award, and 2013 Yahoo ACE (Academic Career Enhancement) Award.

Reference

[Chen et al., 2010] Chen Chen, Xifeng Yan, Feida Zhu, Jiawei Han, and Philip S. Yu. 2010. Graph OLAP: A Multi-Dimensional Framework for Graph Data Analysis. Knowledge and Information Systems (KAIS).

[Danilevsky et al., 2014] Marina Danilevsky, Chi Wang, Nihit Desai, Xiang Ren, Jingyi Guo, and Jiawei Han. 2014. Automatic Construction and Ranking of Topical Keyphrases on Collections of Short Documents. Proc. 2014 SIAM Int. Conf. on Data Mining (SDM'14).

[Deng et al., 2009] Hongbo Deng. Michael R. Lyu and Irwin King. 2009. A Generalized Co-HITS algorithm and its Application to Bipartite Graphs. Proc. KDD2009.

[Deng et al., 2014] Hongbo Deng, Jiawei Han, Hao Li, Heng Ji, Hongning Wang, and Yue Lu. 2014. Exploring and Inferring User-User Pseudo-Friendship for Sentiment Analysis with Heterogeneous Networks. Statistical Analysis and Data Mining, Feb. 2014.

[El-Kishky et al., 2015] Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, Jiawei Han. 2015. Scalable Topical Phrase Mining from Text Corpora. Proc. PVLDB 8(3): 305 – 316.

[Han et al., 2010] Jiawei Han, Yizhou Sun, Xifeng Yan, and Philip S. Yu. 2010. Mining Heterogeneous Information Networks. Tutorial at the 2010 ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD'10), Washington, D.C., July 2010.

[Han et al., 2014] Jiawei Han, Chi Wang, Ahmed El-Kishky. 2014. Bringing Structure to Text: Mining Phrases, Entity Concepts, Topics, and Hierarchies. KDD2014 conference tutorial.

[Huang et al., 2013] Hongzhao Huang, Zhen Wen, Dian Yu, Heng Ji, Yizhou Sun, Jiawei Han and He Li. 2013. Resolving Entity Morphs in Censored Data. Proc. the 51st Annual Meeting of the Association for Computational Linguistics (ACL2013).

[Huang et al., 2014] Hongzhao Huang, Yunbo Cao, Xiaojiang Huang, Heng Ji and Chin-Yew Lin. 2014. Collective Tweet Wikification based on Semi-supervised Graph Regularization. Proc. the 52nd Annual Meeting of the Association for Computational Linguistics (ACL2014).

[Kim et al., 2011] Sangkyum Kim, Hyungsul Kim, Tim Weninger, Jiawei Han, Hyun Duk Kim, "Authorship Classification: A Discriminative Syntactic Tree Mining Approach", in Proc. of 2011 Int. ACM SIGIR Conf. on Research & Development in Information Retrieval (SIGIR'11), Beijing, China, July 2011.

[Li et al., 2013] Yang Li, Chi Wang, Fangqiu Han, Jiawei Han, Dan Roth, Xifeng Yan. 2013. Mining Evidences for Named Entity Disambiguation. Proc. of 2013 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'13). pp. 1070-1078.

[Sun et al., 2009a] Yizhou Sun, Jiawei Han, Peixiang Zhao, Zhijun Yin, Hong Cheng and Tianyi Wu. 2009. RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis. Proc. the 12th International Conference on Extending Database Technology: Advances in Database Technology.

[Sun et al., 2009b] Yizhou Sun, Yintao Yu, and Jiawei Han. 2009. Ranking-Based Clustering of Heterogeneous Information Networks with Star Network Schema. Proc. 2009 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'09).

[Sun et al., 2011] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu and Tianyi Wu. 2011. PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks. Proc. International Conference on Very Large Data Bases (VLDB2011).

[Sun et al., 2012a] Yizhou Sun, Brandon Norick, Jiawei Han, Xifeng Yan, Philip S. Yu, and Xiao Yu. Integrating Meta-Path Selection with User Guided Object Clustering in Heterogeneous Information Networks. Proc. of 2012 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'12).

[Sun et al., 2012b] Yizhou Sun and Jiawei Han. 2012. Mining Heterogeneous Information Networks: Principles and Methodologies, Morgan & Claypool Publishers.

[Sun et al., 2013] Yizhou Sun, Brandon Norick, Jiawei Han, Xifeng Yan, Philip S. Yu, Xiao Yu. 2013. PathSelClus: Integrating Meta-Path Selection with User-Guided Object Clustering in Heterogeneous Information Networks. ACM Transactions on Knowledge Discovery from Data (TKDD), 7(3): 11.

[Sun et al., 2015] Yizhou Sun, Jie Tang, Jiawei Han, Cheng Chen, and Manish Gupta. 2015. Co-Evolution of Multi-Typed Objects in Dynamic Heterogeneous Information Networks. IEEE Trans. on Knowledge and Data Engineering.

[Tao et al., 2014] Fangbo Tao, Jiawei Han, Heng Ji, George Brova, Chi Wang, Brandon Norick, Ahmed El-Kishky, Jialu Liu, Xiang Ren, Yizhou Sun. 2014. NewsNetExplorer: Automatic Construction and Exploration of News Information Networks. Proc. of 2014 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'14).

[Wang et al., 2012] Chi Wang, Jiawei Han, Qi Li, Xiang Li, Wen-Pin Lin and Heng Ji. 2012. Learning Hierarchical Relationships among Partially Ordered Objects with Heterogeneous Attributes and Links. Proc. 2012 SIAM International Conference on Data Mining.

[Wang et al., 2014] Chi Wang, Jialu Liu, Nihit Desai, Marina Danilevsky, and Jiawei Han. 2014. Constructing Topical Hierarchies in Heterogeneous Information Networks. Proc. Knowledge and Information Systems (KAIS).

[Yan et al., 2008] Xifeng Yan, Hong Cheng, Jiawei Han, and Philip S. Yu. 2008. Mining Significant Graph Patterns by Scalable Leap Search. Proc. 2008 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'08).

[Yan and Han, 2002] Xifeng Yan and Jiawei Han. 2002. gSpan: Graph-Based Substructure Pattern Mining. Proc. 2002 of Int. Conf. on Data Mining (ICDM'02).

[Yan and Han, 2003] Xifeng Yan and Jiawei Han. 2003. CloseGraph: Mining Closed Frequent Graph Patterns. Proc. 2003 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'03), Washington, D.C., Aug. 2003.

[Yu et al., 2014] Dian Yu, Hongzhao Huang, Taylor Cassidy, Heng Ji, Chi Wang, Shi Zhi, Jiawei Han, Clare Voss and Malik Magdon-Ismail. 2014. The Wisdom of Minority: Unsupervised Slot Filling Validation based on Multi-dimensional Truth-Finding. Proc. The 25th International Conference on Computational Linguistics (COLING2014).

Structured Belief Propagation for NLP

Matthew R. Gormley **Jason Eisner**
Department of Computer Science
Johns Hopkins University, Baltimore, MD
{mrg,jason}@cs.jhu.edu

1 Tutorial Overview

Statistical natural language processing relies on probabilistic models of linguistic structure. More complex models can help capture our intuitions about language, by adding linguistically meaningful interactions and latent variables. However, inference and learning in the models we *want* often poses a serious computational challenge.

Belief propagation (BP) and its variants provide an attractive approximate solution, especially using recent training methods. These approaches can handle joint models of interacting components, are computationally efficient, and have extended the state-of-the-art on a number of common NLP tasks, including dependency parsing, modeling of morphological paradigms, CCG parsing, phrase extraction, semantic role labeling, and information extraction (Smith and Eisner, 2008; Dreyer and Eisner, 2009; Auli and Lopez, 2011; Burkett and Klein, 2012; Naradowsky et al., 2012; Stoyanov and Eisner, 2012).

This tutorial delves into BP with an emphasis on recent advances that enable state-of-the-art performance in a variety of tasks. Our goal is to elucidate how these approaches can easily be applied to new problems. We also cover the theory underlying them. Our target audience is researchers in human language technologies; we do not assume familiarity with BP.

In the first three sections, we discuss applications of BP to NLP problems, the basics of modeling with factor graphs and message passing, and the theoretical underpinnings of "what BP is doing" and how it relates to other inference techniques. In the second three sections, we cover key extensions to the standard BP algorithm to enable modeling of linguistic structure, efficient inference, and approximation-aware training. We survey a variety of software tools and introduce a new software framework that incorporates many of the modern approaches covered in this tutorial.

2 Outline

1. Probabilistic Modeling [15 min., Eisner]
   - Intro: Modeling with factor graphs
   - Constituency and dependency parsing
   - Joint CCG Parsing and supertagging
   - Transliteration; Morphology
   - Alignment; Phrase extraction
   - Joint models for NLP; Semantic role labeling; Targeted sentiment
   - Variable-centric view of the world
2. Belief Propagation Basics [40 min., Eisner]
   - Messages and beliefs
   - Sum-product algorithm
   - Relation to the forward-backward and Viterbi algorithms
   - BP as dynamic programming
   - Acyclic vs. loopy graphs
3. Theory [25 min., Gormley]
   - From sum-product to max-product
   - From arc consistency to BP
   - From Gibbs sampling to particle BP to BP
   - Convergence properties
   - Bethe free energy
4. Incorporating Structure into Factors and Variables [30 min., Gormley]
   - Embedding dynamic programs (e.g. inside-outside) within factors
   - String-valued variables and finite state machines
5. Message approximation and scheduling [20 min., Eisner]
   - Computing fewer messages
   - Pruning messages
   - Expectation Propagation and Penalized EP
6. Approximation-aware Training [30 min., Gormley]
   - Empirical risk minimization under approximations (ERMA)
   - BP as a computational expression graph
   - Automatic differentiation (AD)
7. Software [10 min., Gormley]

Proceedings of the Tutorials of the 53rd Annual Meeting of the ACL and the 7th IJCNLP, pages 5–6,
Beijing, China, July 26-31, 2015.

3 Instructors

Matt Gormley is a PhD student at Johns Hopkins University working with Mark Dredze and Jason Eisner. His current research focuses on joint modeling of multiple linguistic strata in learning settings where supervised resources are scarce. He has authored papers in a variety of areas including topic modeling, global optimization, semantic role labeling, relation extraction, and grammar induction.

Jason Eisner is a Professor in Computer Science and Cognitive Science at Johns Hopkins University, where he has received two school-wide awards for excellence in teaching. His 90+ papers have presented many models and algorithms spanning numerous areas of NLP. His goal is to develop the probabilistic modeling, inference, and learning techniques needed for a unified model of all kinds of linguistic structure. In particular, he and his students introduced structured belief propagation (which incorporates classical NLP models and their associated dynamic programming algorithms), as well as loss-calibrated training for use with belief propagation.

References

Michael Auli and Adam Lopez. 2011. A comparison of loopy belief propagation and dual decomposition for integrated CCG supertagging and parsing. In *Proceedings of ACL*.

David Burkett and Dan Klein. 2012. Fast inference in phrase extraction models with belief propagation. In *Proceedings of NAACL*.

Markus Dreyer and Jason Eisner. 2009. Graphical models over multiple strings. In *Proceedings of EMNLP*.

Jason Naradowsky, Sebastian Riedel, and David Smith. 2012. Improving NLP through marginalization of hidden syntactic structure. In *Proceedings of EMNLP 2012*.

David A. Smith and Jason Eisner. 2008. Dependency parsing by belief propagation. In *Proceedings of EMNLP*.

Veselin Stoyanov and Jason Eisner. 2012. Minimum-risk training of approximate CRF-Based NLP systems. In *Proceedings of NAACL-HLT*.

Sentiment and Belief: How to Think about, Represent, and Annotate Private States

A Tutorial

Owen Rambow
Columbia University
rambow@ccls.columbia.edu

Janyce Wiebe
University of Pittsburgh
wiebe@cs.pitt.edu

1 Tutotial Description

1.1 Introduction

Over the last ten years, there has been an explosion in interest in sentiment analysis, with many interesting and impressive results. For example, the first twenty publications on Google Scholar returned for the Query "sentiment analysis" all date from 2003 or later, and have a total citation count of 12,140. The total number of publications is in the thousands. Partly, this interest is driven by the immediate commercial applications of sentiment analysis.

Sentiment is a "private state" (Wiebe, 1990). However, it is not the only private state that has received attention in the computational literature; others include belief and intention. In this tutorial, we propose to provide a deeper understanding of what a private state is. We will concentrate on sentiment and belief. We will provide background that will allow the tutorial participants to understand the notion of a private state as a cognitive phenomenon, which can be manifested linguistically in communication in various ways. We will explain the formalization in terms of a triple of state, source, and target. We will discuss how to model the source and the target. We will then explain in some detail the annotations that have been made. The issue of annotation is crucial for private states: while the MPQA corpus (Wiebe et al., 2005; Wilson, 2007) has been around for some time, most research using it does not make use of many of its features. We believe this is because the MPQA annotation is quite complex and requires a deeper understanding of the phenomenon of "private state", which is what the annotation is getting at. Furthermore, there are currently several efforts underway of creating new versions of annotations, which we will also present.

The larger goal of this tutorial is to allow the tutorial participants to gain a deeper understanding of the role of private states in human communication, and to encourage them to use this deeper understanding in their computational work. The immediate goal of this tutorial is to allow the participants to make more complete use of available annotated resources. We propose to achieve these goals by concentrating on annotated corpora, since this will allow participants to both understand the underlying content (achieving the larger goal) and the technical details of the annotations (achieving the immediate goal).

1.2 Current Work on Annotating Sentiment

To date, the computational analyses of sentiment are often fairly superficial. Much work in sentiment analysis and opinion mining is at the document level (Pang et al., 2002). There is increasing interest in more fine-grained levels: sentence-level (McDonald et al., 2007), phrase-level (Choi and Cardie, 2008; Agarwal et al., 2009), aspect-level (Hu and Liu, 2004; Titov and McDonald, 2008), etc. Sentiments toward entities and events ("eTargets") expressed in blogs, newswire, editorials, etc. are particularly important. A system that could recognize sentiments toward entities and events would be valuable in an application such as Automatic Question Answering, to support answering questions such as "Toward whom/what is X negative/positive?" "Who is negative/positive toward X?" (Stoyanov et al., 2005).

Proceedings of the Tutorials of the 53rd Annual Meeting of the ACL and the 7th IJCNLP, pages 7–11,
Beijing, China, July 26-31, 2015.

Or, to augment an automatic wikification system (Ratinov et al., 2011), which could include information about whom or what the subject supports or opposes. A recent NIST evaluation – The Knowledge Base Population (KBP) Sentiment track[1] — aims at using corpora to collect information regarding sentiments expressed toward or by named entities. Annotated corpora of reviews (Hu and Liu, 2004; Titov and McDonald, 2008), widely used in NLP, often include annotations of targets that are aspects of products or services. As such, they are somewhat limited, excluding, e.g., events or agents of events.

A widely used corpus is Version 2 of the MPQA opinion annotated corpus (Wiebe et al., 2005; Wilson, 2007). It is entirely span-based, and contains no eTarget annotations. However, it provides an infrastructure for sentiment annotation that is not provided by other sentiment NLP corpora, and is much more varied in topic, genre, and publication source. MPQA 3.0 (Deng and Wiebe, 2015), which was recently created, adds entity- and event-target (*eTarget*) annotations to the MPQA 2.0 annotations (Wilson, 2007).[2] The MPQA annotations consist of *private states*, states of a *source* holding an *attitude*, optionally toward a *target*. An important property of sources is that they are *nested*, reflecting the fact that private states and speech events are often embedded in one another. There are several types of attitudes included in MPQA 2.0, including sentiment, arguing, and intention. The tutorial will focus on sentiments (while also discussing the others), which are defined in (Wilson, 2007) as positive and negative evaluations, emotions, and judgements. In the future, eTargets may be added to private states with other types of attitudes.

This tutorial will present the original MPQA annotation scheme (V2) and its recent extension to include eTarget annotations (V3), which we believe is a valuable new resource for the community.

1.3 Belief Annotations

Compared to sentiment, belief has received far less attention in the computational community. There have been several efforts at annotating belief recently. The most complete is FactBank (Saurí and Pustejovsky, 2009), which represents the source of the belief, the target, the strength, and the polarity (using a system of 10 tags which cover strength and polarity). Following (Wiebe et al., 2005), the sources are nested, reflecting the same nesting of private states we also observe for sentiment. FactBank is a rich and complex annotation; the so-called LU corpus of Diab et al. (2009) was created independently, and represents a subset of the annotations of FactBank. The LU corpus annotates only the writer's belief in the propositions in the text, only distinguishes 3 types of belief, but does clearly represent the target. Unlike FactBank, which is annotated on top of the Penn Treebank, the LU corpus represents a diverse set of texts. The recent annotations at the LDC for the DARPA DEFT project follow the simplicity of the LU corpus annotations, but extend the tagset of the LU corpus to four tags. An annotation effort in the spring of 2015 will include the source of the belief. The LDC effort is important since it covers a new domain – web discussion forums. Its size is an order of magnitude larger than that of FactBank or the LU corpus (about 800,000 words). This tutorial will discuss these resources and compare the annotations.

1.4 Integration Issues

Sentiment and belief are very similar: most importantly, they are both private states. They both involve a holder and a target, and within the broad categories of sentiment and belief there are subdivisions, which can affect the strength of the private state. There is an important difference though: while the target of a sentiment can be an entity or an event (state of affairs), belief can only target a state of affairs. In addition to being similar types of phenomena, the same linguistic means can convey sentiment and belief at the same time: the utterance *I regret that I am leaving tomorrow* reveals both the utterer's sentiment and belief towards the leaving event. Despite these interactions between sentiment and belief, there has been no attempt to jointly annotate or predict sentiment and belief. The tutorial will use examples to show the interaction between sentiment and belief, and discuss some issues that arise in joint annotation and tagging.

[1] http://www.nist.gov/tac/2014/KBP/Sentiment/index.html

[2] Available at http://mpqa.cs.pitt.edu

2 Tutorial Contents

1. Introduction: an overview over the issue of private states, and how they relate to other well-known concepts such as the BDI (belief-desire-intention) model (Bratman, 1999 1987), related work in NLP (such as RST (Mann and Thompson, 1987) and dialog act tagging), linguistic semantics (for example, the notion of veridicity (Karttunen, 1971) and modality), and cognitive science. (45 minutes)

2. Representing sentiment: a presentation of early work, of MPQA V2 (with nested sources, and attitude, expressive-subjective element, and target span annotations), and of MPQA Version 3 (extension of MPQA V2 to eTargets). (45 minutes)

3. Break (15 minutes)

4. Representing belief: a presentation of FactBank, the LU corpus, and the ongoing LDC annotation under the DARPA DEFT program. (30 minutes)

5. Integration and looking forward: a discussion of how sentiment and belief interact, and how we can integrate their annotations, including a discussion of a General Modality Annotation Scheme. (45 minutes)

3 Tutorial Instructors

3.1 Owen Rambow

Owen Rambow is a Senior Research Scientist at the Center for Computational Learning Systems at Columbia University. He is also the co-chair of the Center for New Media at the Data Science Institute at Columbia University. He has been interested in modeling cognitive states in relation to language for a long time, initially in the context of natural language generation (Rambow, 1993; Walker and Rambow, 1994). More recently, he has studied belief in the context of recognizing beliefs in language (Diab et al., 2009; Prabhakaran et al., 2010; Danlos and Rambow, 2011; Prabhakaran et al., 2012). He is currently involved in the DARPA DEFT Belief group, working with other researchers and with the LDC to define annotation standards and evaluations. He has recently led the pilot evaluation for belief recognition (in English) in the DARPA DEFT program.

He has been the PI or co-PI on many other Government grants from the NSF, DARPA, and IARPA. He has been the Chair of the North American Chapter of the Association for Computational Linguistics. He has been on the editorial board of *Computational Linguistics*, and has served as chair or area chair for several major conferences. `http://www.cs.columbia.edu/~rambow`

3.2 Janyce Wiebe

Janyce Wiebe is Professor of Computer Science and Professor and Co-Director of the Intelligent Systems at the University of Pittsburgh. She has worked on issues related to private states for some time, originally in the context of tracking point of view in narrative (Wiebe, 1994), and later in the context of recognizing sentiment in other genres such as news articles (Wilson et al., 2005). She has approached the area from the perspective of corpus annotation (Wiebe et al., 2005; Deng et al., 2013), lexical semantics (Wiebe and Mihalcea, 2006), and discourse (Somasundaran et al., 2009). In addition to continuing these lines of research, she has recently begun investigating implicatures in opinion analysis (Deng and Wiebe, 2014).

She has received funding for her research from NSF, NIH, DARPA, ONR, NSA, ARDA, and Homeland Security. She was Program Chair of NAACL 2000 and Program Co-Chair of ACL-IJCNLP 2009. She has been on the editorial board of *Computational Linguistics* and is currently an action editor for *Transactions of the ACL*. `http://people.cs.pitt.edu/~wiebe/`

References

Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic N-grams. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 24–32, Athens, Greece, March. Association for Computational Linguistics.

Michael E. Bratman. 1999 [1987]. *Intention, Plans, and Practical Reason*. CSLI Publications.

Yejin Choi and Claire Cardie. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 793–801. Association for Computational Linguistics.

Laurence Danlos and Owen Rambow. 2011. Discourse relations and propositional attitudes. In *Proceedings of CID 2011 - Fourth International Workshop on Constraints in Discourse*.

Lingjia Deng and Janyce Wiebe. 2014. Sentiment propagation via implicature constraints. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 377–385, Gothenburg, Sweden, April. Association for Computational Linguistics.

Lingjia Deng and Janyce Wiebe. 2015. Entity/event-level sentiment annotation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (short paper)*. Association for Computational Linguistics, May.

Lingjia Deng, Yoonjung Choi, and Janyce Wiebe. 2013. Benefactive/malefactive event and writer attitude annotation. In *ACL 2013 (short paper)*. Association for Computational Linguistics.

Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 68–73, Suntec, Singapore, August. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Lauri Karttunen. 1971. Some observations on factivity. *Research on Language & Social Interaction*, 4(1):55–69.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical Structure Theory: A theory of text organization. Technical Report ISI/RS-87-190, ISI.

Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In *Annual Meeting-Association For Computational Linguistics*, page 432. Citeseer.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2010. Automatic committed belief tagging. In *Coling 2010: Posters*, pages 1014–1022, Beijing, China, August. Coling 2010 Organizing Committee.

Vinodkumar Prabhakaran, Michael Bloodgood, Mona Diab, Bonnie Dorr, Lori Levin, Christine D. Piatko, Owen Rambow, and Benjamin Van Durme. 2012. Statistical modality tagging from rule-based annotations and crowdsourcing. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 57–64, Jeju, Republic of Korea, July. Association for Computational Linguistics.

Owen Rambow. 1993. Rhetoric as knowledge. In *Proceedings of the ACL Workshop on Intentionality and Structure in Discourse Relations*, Columbus, OH.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1375–1384. Association for Computational Linguistics.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268. 10.1007/s10579-009-9089-9.

Swapna Somasundaran, Galileo Namata, Janyce Wiebe, and Lise Getoor. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 170–179, Singapore, August. Association for Computational Linguistics.

Veselin Stoyanov, Claire Cardie, and Janyce Wiebe. 2005. Multi-Perspective Question Answering using the OpQA corpus. In *Proceedings of the Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, pages 923–930, Vancouver, Canada.

Ivan Titov and Ryan T McDonald. 2008. A joint model

of text and aspect ratings for sentiment summarization. In *ACL*, volume 8, pages 308–316. Citeseer.

Marilyn Walker and Owen Rambow. 1994. The role of cognitive modeling in achieving communicative intentions. In *Proceedings of the Seventh International Workshop on Natural Language Generation*, Kennebunkport, ME.

Janyce Wiebe and Rada Mihalcea. 2006. Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1065–1072, Sydney, Australia, July. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language ann. *Language Resources and Evaluation*, 39(2/3):164–210.

Janyce M. Wiebe. 1990. Identifying subjective characters in narrative. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)*.

Janyce Wiebe. 1994. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, pages 347–354, Vancouver, Canada.

Theresa Wilson. 2007. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of private states*. Ph.D. thesis, Intelligent Systems Program, University of Pittsburgh.

Corpus Pattern for Semantic Processing

Patrick Hanks
University of Wolverhampton, UK
patrick.w.hanks@gmail.com

Elisabetta Jezek
University of Pavia, IT
jezek@unipv.it

Daisuke Kawahara
Kyoto University, JP
dk@i.kyoto-u.ac.jp

Octavian Popescu
IBM Research, US
o.popescu@us.ibm.com

1 Introduction

This tutorial presents a corpus-driven, pattern-based empirical approach to meaning representation and computation. Patterns in text are everywhere, but techniques for identifying and processing them are still rudimentary. Patterns are not merely syntactic but syntagmatic: each pattern identifies a lexico-semantic clause structure consisting of a predicator (verb or predicative adjective) together with open-ended lexical sets of collocates in different clause roles (subject, object, prepositional argument, etc.). If NLP is to make progress in identifying and processing text meaning, pattern recognition and collocational analysis will play an essential role, because:

> Many, if not most meanings, require the presence of more than one word for their normal realization. ... Patterns of co-selection among words, which are much stronger than any description has yet allowed for, have a direct connection with meaning. (J. M. Sinclair, 1998).

The tutorial presents methods for building patterns on the basis of corpus evidence, using machine learning methods. It discusses some possible applications of pattern inventories and invites discussion of others. It is intended for an audience with heterogeneous competences but with a common interest in corpus linguistics and computational models for meaning-related tasks in NLP. We report on the methodologies for building resources for semantic processing and their contribution to NLP tasks. The goal is to provide the audience with an operative understanding of the methodology used to acquire corpus patterns and of their utility in NLP applications.

2 Overview

Natural language sentences make use of lexical, syntactic, semantic and pragmatic information in order to fulfill their role of conveying meaning. Previous research on computing the meaning of linguistic expressions - from approaches which consider overt distributional information on words to deep semantic ones, based on first order and lambda calculus representations - has highlighted two major issues: (1) the appropriate level of formalization for meaning representation cannot be founded only on premises derived from prior experience, (2) the lack of large-scale annotated corpora which combine different levels of semantic annotation hinders the development of machine-learning applications. In particular, in the framework of big data analytics for semantically processing large corpora, these two issues must be addressed.

The regular structure of normal clauses can be used as a basis in order to learn the rules that lie behind recurrent meaningful constructs in natural language. It has been shown (Hanks&Pustejovsky 2004, Pustejovsky&Jezek 2008, Popescu&Magnini 2007, Popescu 2013, Kawahara et al. 2014)

Proceedings of the Tutorials of the 53rd Annual Meeting of the ACL and the 7th IJCNLP, pages 12–15,
Beijing, China, July 26-31, 2015.

that it is possible to identify and to learn corpus patterns that encode the information that accounts for the senses of the verb and its arguments in the context. These patterns link the syntactic structure of clauses and the semantic types of argument fillers via the role that each of these play in the disambiguation of the clause as a whole. With regard to irregularities, there are quite a few clauses in a corpus where these patterns do not seem to match the text, because of the apparent incompatibility between the actual and the expected semantic types of the arguments (Jezek&Hanks 2010, Hanks 2012). However, it is possible to build statistical models that simultaneously generate both the regular and the innovative representation of a clause. Available solutions developed up to now range from supervised to totally unsupervised approaches. The patterns obtained encode the necessary information for handling the meaning of each word individually as well as that of the clause as a whole. As such they are instrumental in building better language models (Dligach&Palmer 2011). In the contexts matched by such patterns, any word is unequivocally disambiguated. The semantic types used in pattern representation play a discriminative role, therefore the patterns are sense discriminative and as such they can be used in word sense disambiguation and other meaning-related tasks (see among others Pustejovsky et al. 2004, Cumby&Roth 2003, Popescu&Magnini 2007, Pustejovsky et al. 2010, Popescu et al. 2014). Also, the meaning of a pattern as a whole is expressed as a set of basic implicatures. The implicatures are instrumental in textual entailment, semantic similarity and paraphrasing generation (Popescu et al. 2011, Nicolae&Popescu 2013, Vo et. al 2014). Depending on the proposed application, the implicatures associated with a pattern may be expressed in any of a wide variety of other ways, e.g. as a translation into another language or as a synonym set. The automatic aligning of the set of patterns of two languages via their shared semantic types is used in meaning-preserving translation tasks (Popescu&Jezek 2013).

The relatively recent research on corpus data has shown that intermediate text representations (ITRs), built in a bottom-up manner from corpus examples towards a complex representation of clauses, play an important role in dealing with the meaning disambiguation problem. ITRs offer an important degree of freedom in finding the right cut between various levels of semantic information. Large-scale corpus-driven lexical analysis leads to two apparently contradictory conclusions. On the one hand, the regularities of word use (valencies, collocations) are more regular than what most pre-corpus linguists would have predicted. On the other hand, the irregularities are more irregular. In particular, verb usage in language displays a continuous blend between regular constructs with clearly distinct senses and new and innovative usages. The Theory of Norms and Exploitations (Hanks 2013) maintains that language exhibits mainly a rule-governed behavior, but argues that there is not just one monolithic system of rules. Instead, there are two interactive sets of rules: 1) Norms: a set of rules for using words normally and idiomatically: these are the rules of grammar; they account for 70%-90% of all utterances - depending on the type of the verb, the topic, and the domain. However, they do not account for linguistic creativity, nor for changes in word meaning; 2) Exploitation rules, which account for creativity and innovative usage (about 10%-30% of corpus examples). Exploitation rules also account for phenomena such as meaning shift. Pattern Dictionaries are resources based on Corpus Pattern Analysis (CPA). They contains examples for each category for a large number of English and Italian verbs and are available at http://pdev.org.uk/ (Hanks 2004), and at http://tpas.fbk.eu/resource (Jezek et al. 2014).

The corpus-pattern methodology is designed to offer a viable solution to meaning representation. The techniques we present are widely applicable in NLP and they deal efficiently with data sparseness and open domain expression of semantic relationships.

The tutorial is divided into three main parts, which are strongly interconnected: (A) Building Corpus Patterns via the Theory of Norms and Exploitations, (B) Inducing Semantic Types and Semantic Task Oriented Ontologies, and (C) Machine Learning and Applications of Corpus Patterns.

3 Outline

3.1 Corpus, Language Usage and Computable Semantic Properties of Verb Phrases section

Basic Computational Semantic Concepts

Theory of Norm and Exploitation of Language Usage

Corpus Pattern Analysis in Sketch Engine

Sense Discriminative Patterns

3.2 Semantic Types and Ontologies

Argument Structures

Frames and Semantic Types

Inducing Semantic Types

Discriminative Patterns

3.3 Statistical Models for Corpus Pattern Recognition and Extraction. NLP Applications

Finite State Markov Chains

Naive Bayesian and Gaussian Random Fields for Conditional Probabilities over Semantic Types

Latent Dirichlet Analysis for Unsupervised Pattern Extraction

Probably Approximately Correct and Statistical Query Model

Joint Source Channel Model for Recognition of Norm and Exploitation

Textual Entailment, Paraphrase Generation and Textual Similarity with Corpus Patterns

4 Tutors

Patrick Hanks is Professor in Lexicography at the Research Institute of Information and Language Processing at the University of Wolverhampton. He is also a visiting professor at the Bristol Centre for Linguistics (University of the West of England). He studied English Language and Literature at Oxford and was awarded a PhD in Informatics at the Masaryk University in Brno, Czech Republic. In the 1980s he was the managing editor of Cobuild, an innovative corpus-based dictionary compiled at the University of Birmingham. In 1989-90 he co-authored with Ken Church and others a series of papers on statistical approaches to lexical analysis. For ten years (1990–2000) he was chief editor of Current English Dictionaries at Oxford University Press. He is the author of *Lexical Analysis: Norms and Exploitations* (MIT Press, 2013), which presents a new theory of word meaning and language in use. He is a consultant on lexicographical methodology and definition to several institutions throughout Europe, including Oxford University Press, and is a frequent invited plenary speaker at international conferences on lexicography, corpus linguistics, figurative language, onomastics, and phraseology.

Elisabetta Jezek has been teaching Syntax and Semantics and Applied Linguistics at the University of Pavia since 2001. Her research interests and areas of expertise are lexical semantics, verb classification, theory of Argument Structure, event structure in syntax and semantics, corpus annotation, computational Lexicography.

Daisuke Kawahara is an Associate Professor at Kyoto University. He is an expert in the areas of parsing, knowledge acquisition and information analysis. He teaches gradu-

ate classes in natural language processing. His current work is focused on automatic induction of semantic frames and semantic parsing, verb polysemic classes, verb sense disambiguation, and automatic induction of semantic frames.

Octavian Popescu is a researcher at IBM T. J. Watson Research Center, working on computational semantics with focus on corpus patterns for question answering, textual entailment and paraphrasing. He taught various NLP graduate courses in computational semantics at Trento University (IT), Colorado University at Boulder (US) and University of Bucharest (RO).

References

C. Cumby and D. Roth "On Kernel Methods for Relational Learning", in Proceedings of ICML 2003, Washington 2003

D. Dligach and M. Palmer: "Good Seed Makes a Good Crop: Accelerating Active Learning Using Language Modeling", in Proceedings of ACL, Oregon, 2011

P. Hanks, "Corpus Pattern Analysis". In Williams G. and S. Vessier (eds) Proceedings of the XI Euralex International Congress, Lorient, Université de Bretagne-Sud, 2004

P. Hanks and J. Pustejovsky. "Common Sense About Word Meaning:Sense in Context", in Proceedings of the TSD, Volume 3206, 2004.

P. Hanks "How People use words to make Meanings. Semantic Types meet Valencies". In A. Bulton and J. Thomas (eds.) Input, Process and Product: Developments in Teaching and Language Corpora. Masaryk University Press, 2012

P. Hanks "Lexical Analysis: Norms and Exploitations.". MITPress 2013

E. Jezek and P. Hanks, "What lexical sets tell us about conceptual categories", In Lexis, E-Journal in English Lexicology, 4, 7-22, 2010.

E. Jezek, B. Magnini, A. Feltracco, A. Bianchini, O. Popescu "T-PAS; A resource of Typed Predicate Argument Structures for linguistic analysis and semantic processing", in Proceedings of LREC, Reykjavik 2014

D. Kawahara, D. Pederson, O. Popescu, M. Palmer 2014. "Inducing Example-based Semantic Frames from a Massive Amount of Verb Uses", in Proceedings of the EACL, Gothenburg, 2014

V. Niculae and O. Popescu, "Determining is-a relationships for Textual Entailment", in Prceedings of JSSP , Trento, 2013

O. Popescu, B. Magnini "Sense Discriminative Patterns for Word Sense Disambiguation", in Proceedings of Semantic Content Acquisition and Representation, NODALIDA, Tartu, 2007.

O. Popescu, E. Cabrio, B. Magnini Journal Proceedings of the IJCAI Workshop Learning by Reasoning and its Applications in Intelligent Question-Answering, Barcelona 2011

O. Popescu , E. Jezek. "Pattern Based Translation", in Proceedings of Tralogy-II, Paris 2013

O. Popescu. "Learning Corpus Pattern with Finite State Automata", in Proceedings of IWSC, Berlin, 2013.

O. Popescu, P. Hanks, M. Palmer, "Mapping CPA onto Ontonotes Senses", in Proceedings of LREC, Reykjavik, 2014

J. Pustejovsky, P. Hanks, and A. Rumshisky. "Sense in Context", in Proceedings of COLING 2004, Geneva, 2004

J. Pustejovsky, E. Jezek "Semantic Coercion in Language: Beyond Distributional Analysis", Italian Journal of Linguistics 20, 1, 181-214, 2008.

J. M.Sinclair "The Lexical Item", in E. WEigand (ed.) Contrastive Lexical Semantics. Benjamins, 1998

N. Vo, O. Popescu, T. Caselli, "FBK-TR: SVM for Semantic Relatedness and Corpus Patterns for RTE", in Proceedings SemEval, Dublin, 2014

Matrix and Tensor Factorization Methods for Natural Language Processing

Guillaume Bouchard[*] **Jason Naradowsky**[#] **Sebastian Riedel**[#]
Tim Rocktäschel[#] and **Andreas Vlachos**[#]
[*] Xerox Research Centre Europe
guillaume.bouchard@xerox.com
[#] Computer Science Department
University College London
{j.narad, s.riedel, t.rocktaschel, a.vlachos}@cs.ucl.ac.uk

1 Tutorial Objectives

Tensor and matrix factorization methods have attracted a lot of attention recently thanks to their successful applications to information extraction, knowledge base population, lexical semantics and dependency parsing. In the first part, we will first cover the basics of matrix and tensor factorization theory and optimization, and then proceed to more advanced topics involving convex surrogates and alternative losses. In the second part we will discuss recent NLP applications of these methods and show the connections with other popular methods such as transductive learning, topic models and neural networks. The aim of this tutorial is to present in detail applied factorization methods, as well as to introduce more recently proposed methods that are likely to be useful to NLP applications.

2 Tutorial Overview

2.1 Matrix/Tensor Factorization Basics

In this part, we first remind essential results on bilinear forms, spectral representations of matrices and low-rank approximation theorems, which are often omitted in undergraduate linear algebra courses. This includes the link between eigenvalue decomposition and singular value decomposition and the trace-norm (a.k.a. nuclear norm) as a convex surrogate of the low-rank constraint on optimization problems. Then, an overview of the most efficient algorithms to solve low-rank constrained problems is made, from the power iteration method, the Lanczos algorithm and the implicitly restarted Arnoldi method that is implemented in the LAPACK library (Anderson et al., 1999). We show how to interpret low-rank models as probabilistic models (Bishop, 1999) and how we can extend SVD algorithms that can factorize non-standard matrices (i.e. with non-Gaussian noise and missing data) using gradient descent, reweighted SVD or Frank-Wolfe algorithms. We then show that combining different convex objectives can be a powerful tool, and we illustrate it by deriving the robust PCA algorithm by adding an L_1 penalty term in the objective function (Candès and Recht, 2009). Furthermore, we introduce Bayesian Personalized Ranking (BPR) for matrix and tensor factorization which deals with implicit feedback in ranking tasks (Rendle et al., 2009). Finally, will introduce the collective matrix factorization model (Singh and Gordon, 2008) and tensor extensions (Nickel et al., 2011) for relational learning.

2.2 Applications in NLP

In this part we will discuss recent work applying matrix/tensor factorization methods in the context of NLP. We will review the Universal Schema paradigm for knowledge base construction (Riedel et al., 2013) which relies on matrix factorization and BPR, as well as recent extensions of the RESCAL tensor factorization (Nickel et al., 2011) approach and methods of injecting logic into the embeddings learned (Rocktäschel et al., 2015). These applications will motivate the connections between matrix factorization and transductive learning (Goldberg et al., 2010), as well as tensor factorization and multi-task learning (Romera-Paredes et al., 2013). Furthermore, we will review work on applying matrix and tensor factorization to sparsity reduction in syntactic dependency parsing (Lei et al., 2014) and word representation learning (Pennington et al., 2014). In addition, we will discuss the connections between matrix factorization, latent semantic analysis and topic modeling (Stevens et al., 2012).

Proceedings of the Tutorials of the 53rd Annual Meeting of the ACL and the 7th IJCNLP, pages 16–18,
Beijing, China, July 26-31, 2015.

3 Structure

Part I: Matrix/Tensor Factorization Basics (90 minutes)

- Matrix factorization basics (40 min): bilinear forms, spectral representations, low rank approximations theorems, optimization with stochastic gradient descent, losses
- Tensor factorization basics (20 minutes): representations,notation decompositions (Tucker etc.)
- Advanced topics (30 minutes): convex surrogates, L1 regularization, alternative losses (ranking loss, logistic loss)

Break (15 minutes)

Part II: Applications in NLP (75 minutes)

- Information extraction, knowledge base population with connections to transductive learning and multitask learning (35 minutes)
- Lexical semantics with connections to neural networks, latent semantic analysis and topic models (30 minutes)
- Structured prediction (10 minutes)

4 About the Speakers

Guillaume Bouchard is a senior researcher in statistics and machine learning at Xerox, focusing on statistical learning using low-rank model for large relational databases. His research includes text understanding, user modeling, and social media analytics. The theoretical part of his work is related to the efficient algorithms to compute high dimensional integrals, essential to deal with uncertainty (missing and noisy data, latent variable models, Bayesian inference). The main application areas of his work includes the design of virtual conversational agents, link prediction (predictive algorithms for relational data), social media monitoring and transportation analytics. His web page is available at `www.xrce.xerox.com/people/bouchard`.

Jason Naradowsky is a postdoc at the Machine Reading group at UCL. Having previously obtained a PhD at UMass Amherst under the supervision of David Smith and Mark Johnson, his current research aims to improve natural language understanding by performing task-specific training of word representations and parsing models. He is also interested in semi-supervised learning, joint inference, and semantic parsing. His web page is available at `http://narad.github.io/`.

Sebastian Riedel is a senior lecturer at University College London and an Allen Distinguished Investigator, leading the Machine Reading Lab. Before, he was a postdoc and research scientist with Andrew McCallum at UMass Amherst, a researcher at Tokyo University and DBCLS with Tsujii Junichi, and a PhD student with Ewan Klein at the University of Edinburgh. He is interested in teaching machines how to read and works at the intersection of Natural Language Processing (NLP) and Machine Learning, investigating various stages of the NLP pipeline, in particular those that require structured prediction, as well as fully probabilistic architectures of end-to-end reading and reasoning systems. Recently he became interested in new ways to represent textual knowledge using low-rank embeddings and how to reason with such representations. His web page is available at `http://www.riedelcastro.org/`.

Tim Rocktäschel is a PhD student in Sebastian Riedel's Machine Reading group at University College London. Before that he worked as research assistant in the Knowledge Management in Bioinformatics group at Humboldt-Universität zu Berlin, where he also obtained his Diploma in Computer Science. He is broadly interested in representation learning (e.g. matrix/tensor factorization, deep learning) for NLP and automated knowledge base completion, and how these methods can take advantage of symbolic background knowledge. His webpage is available at `http://rockt.github.io/`.

Andreas Vlachos is postdoc at the Machine Reading group at UCL working with Sebastian Riedel on automated fact-checking using low-rank factorization methods. Before that he was a postdoc at the Natural Language and Information Processing group at the University of Cambridge and at the University of Wisconsin-Madison. He is broadly interested in natural language understanding (e.g. information extraction, semantic parsing) and in machine learning approaches that would help us towards this goal. He has also worked on active learning, clustering and biomedical text mining. His web page is available at `http://sites.google.com/site/andreasvlachos/`.

References

[Anderson et al.1999] Edward Anderson, Zhaojun Bai, Christian Bischof, Susan Blackford, James Demmel, Jack Dongarra, Jeremy Du Croz, Anne Greenbaum, S Hammerling, Alan McKenney, et al. 1999. *LAPACK Users' guide*, volume 9. SIAM.

[Bishop1999] Christopher M Bishop. 1999. Bayesian PCA. In *Advances in Neural Information Processing Systems*, pages 382–388.

[Candès and Recht2009] Emmanuel J Candès and Benjamin Recht. 2009. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772.

[Goldberg et al.2010] Andrew Goldberg, Ben Recht, Junming Xu, Robert Nowak, and Xiaojin Zhu. 2010. Transduction with matrix completion: Three birds with one stone. In *Advances in Neural Information Processing Systems 23*, pages 757–765.

[Lei et al.2014] Tao Lei, Yu Xin, Yuan Zhang, Regina Barzilay, and Tommi Jaakkola. 2014. Low-rank tensors for scoring dependency structures. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1381–1391.

[Nickel et al.2011] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 809–816.

[Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

[Rendle et al.2009] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 452–461.

[Riedel et al.2013] Sebastian Riedel, Limin Yao, Benjamin M. Marlin, and Andrew McCallum. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, GA.

[Rocktäschel et al.2015] Tim Rocktäschel, Sameer Singh, and Sebastian Riedel. 2015. Injecting Logical Background Knowledge into Embeddings for Relation Extraction . In *Proceedings of the 2015 Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*.

[Romera-Paredes et al.2013] Bernardino Romera-Paredes, Hane Aung, Nadia Bianchi-Berthouze, and Massimiliano Pontil. 2013. Multilinear multitask learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1444–1452.

[Singh and Gordon2008] Ajit P. Singh and Geoffrey J. Gordon. 2008. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 650–658.

[Stevens et al.2012] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961.

Scalable Large-Margin Structured Learning: Theory and Algorithms

Liang Huang **Kai Zhao**
The City University of New York
{liang.huang.sh, kzhao.hf, lemaoliu}@gmail.com

1 Motivations

Much of NLP tries to map structured input (sentences) to some form of structured output (tag sequences, parse trees, semantic graphs, or translated/paraphrased/compressed sentences). Thus structured prediction and its learning algorithm are of central importance to us NLP researchers. However, when applying machine learning to structured domains, we often face scalability issues for two reasons:

1. Even the fastest exact search algorithms for most NLP problems (such as parsing and translation) is too slow for repeated use on the training data, but approximate search (such as beam search) unfortunately breaks down the nice theoretical properties (such as convergence) of existing machine learning algorithms.

2. Even with inexact search, the scale of the training data in NLP still makes pure online learning (such as perceptron and MIRA) too slow on a single CPU.

This tutorial reviews recent advances that address these two challenges. In particular, we will cover principled machine learning methods that are designed to work under vastly inexact search, and parallelization algorithms that speed up learning on multiple CPUs. We will also extend structured learning to the latent variable setting, where in many NLP applications such as translation and semantic parsing the gold-standard derivation is hidden.

2 Contents

1. Overview of Structured Learning
   (a) key challenge 1: search efficiency
   (b) key challenge 2: interactions between search and learning
2. Structured Perceptron
   (a) the basic algorithm
   (b) convergence proof – a purely geometric approach **(updated in 2015)**
   (c) voted and averaged perceptrons, and efficient implementation tricks
   (d) applications in tagging, parsing, etc.
   (e) inseparability and generalization bounds **(new in 2015)**
3. Structured Perceptron under Inexact Search
   (a) convergence theory breaks under inexact search
   (b) early update
   (c) violation-fixing perceptron
   (d) applications in tagging, parsing, etc.

   —coffee break—
4. Large-Margin Structured Learning with Latent Variables
   (a) examples: machine translation, semantic parsing, transliteration
   (b) separability condition and convergence proof **(updated in 2015)**
   (c) latent-variable perceptron under inexact search
   (d) applications in machine translation
5. Parallelizing Large-Margin Structured Learning
   (a) iterative parameter mixing (IPM)
   (b) minibatch perceptron and MIRA

Proceedings of the Tutorials of the 53rd Annual Meeting of the ACL and the 7th IJCNLP, pages 19–20,
Beijing, China, July 26-31, 2015.

3 Instructor Biographies

Liang Huang is an Assistant Professor at the City University of New York (CUNY). He received his Ph.D. in 2008 from Penn and has worked as a Research Scientist at Google and a Research Assistant Professor at USC/ISI. His work is mainly on the theoretical aspects (algorithms and formalisms) of computational linguistics, as well as theory and algorithms of structured learning. He has received a Best Paper Award at ACL 2008, several best paper nominations (ACL 2007, EMNLP 2008, and ACL 2010), two Google Faculty Research Awards (2010 and 2013), and a University Graduate Teaching Prize at Penn (2005). He has given three tutorials at COLING 2008, NAACL 2009 and ACL 2014.

Kai Zhao is a Ph.D. candidate at the City University of New York (CUNY), working with Liang Huang. He received his B.S. from the University of Science and Technology in China (USTC). He has published on structured prediction, online learning, machine translation, and parsing algorithms. He was a summer intern with IBM TJ Watson Research Center in 2013, Microsoft Research Redmond in 2014, and Google Research NYC in 2015.

Detecting Deceptive Opinion Spam using Linguistics, Behavioral and Statistical Modeling

Arjun Mukherjee
Department of Computer Science
University of Houston
501 PGH, 4800 Calhoun Rd. Houston, TX
arjun@cs.uh.edu

1 Introduction

With the advent of Web 2.0, consumer reviews have become an important resource for public opinion that influence our decisions over an extremely wide spectrum of daily and professional activities: e.g., where to eat, where to stay, which products to purchase, which doctors to see, which books to read, which universities to attend, and so on. Positive/negative reviews directly translate to financial gains/losses for companies. This unfortunately gives strong incentives for *opinion spamming* which refers to illegal human activities (e.g., writing fake reviews and giving false ratings) that try to mislead customers by promoting/demoting certain entities (e.g., products and businesses). The problem has been widely reported in the news. Despite the recent research efforts on detection, the problem is far from solved. What is worse is that opinion spamming is widespread. While credit card fraud is as rare as 0.2%, based on our research we estimated that up to 30% of the reviews on many Web sites could be fake. Thus, detecting fake reviews and opinions is a pressing and also profound issue as it is critical to ensure the trustworthiness of the information on the web. Without detecting them, the social media could become a place full of lies, fakes, and deceptions, and completely useless.

Major review hosting sites and e-commerce vendors have already made some progress in detecting fake reviews. However, the task is still extremely challenging because it is very difficult to obtain large-scale ground truth samples of deceptive opinions for algorithm development and for evaluation, or to conduct large-scale domain expert evaluations. Further, in contrast to other kinds of spamming (e.g., Web and link spam, social/blog spam, email spam, etc.) opinion spam has a very unique flavor as it involves fluid sentiments of users and their evaluations. Thus, they require a very different treatment. Since our first paper in 2007 (Jindal and Liu, 2007) on the topic, our group and many other researchers have proposed several algorithms and bridged algorithmic methodologies from various scientific disciplines including computational linguistics (Ott et al., 2011), social and behavioral sciences (Jindal and Liu, 2008; Mukherjee et al., 2013a, b), machine learning, data mining and Bayesian statistics (Mukherjee et al., 2012; Fei et al., 2013; Mukherjee et al., 2013c; Li et al., 2014b; Li et al., 2014a) to solve the problem. The field of deceptive opinion spam has gained a lot of interest in communications (Hancock et al., 2008), psycholinguistics communities (Gokhman et al., 2012), and economic analysis (Wang, 2010) apart from mainstream NLP and Web mining as attested by publications in top tier venues in their respective communities. The problem has far reaching implications in various allied NLP topics including Lie Detection, Forensic Linguistics, Opinion Trust and Veracity Verification and Plagiarism Detection. However, owing to the inherent nature of the problem, a unique blend of NLP, data mining, machine learning, social, behavioral, and statistical techniques are required which many NLP researchers may not be familiar with.

In this tutorial, we aim to cover the problem in its full depth and width, covering diverse algorithms that have been developed over the past 7 years. The most attractive quality of these techniques is that many of them can be adapted for cross-domain and unsupervised settings. Some of the methods are even in use by startups and established companies. Our focus is on insight and understanding, using illustrations and intuitive deductions. The goal of the tutorial is to make the inner workings of these techniques transparent, intuitive and their results interpretable.

2 Content Overview

The first part of the tutorial presents the problem in its various flavors, the NLP techniques, and the algorithms motivated from social and behavioral sciences. It also presents a detailed insight into commercial vs. crowdsourced deceptive opinions using information theory and linguistics. The second section includes detailed math and algorithms for training supervised, unsupervised, semi-supervised, and partially supervised machine learning and statistical models for deceptive opinion spam

Proceedings of the Tutorials of the 53rd Annual Meeting of the ACL and the 7th IJCNLP, pages 21–22,
Beijing, China, July 26-31, 2015.

detection. These algorithms allow us to work on unlabeled data which is a key aspect of the problem as generating high quality labels of fake reviews in large scale is hard if not impossible. We also discuss some new evaluation methods. Additionally, we draw connections to Authorship Attribution to discover fake reviewers with multiple accounts based on their writing styles, which is a new frontier in deceptive opinion spamming. The last part of the tutorial gives a general overview of the different applications of the methods in allied NLP problems and domains, data sources, and the limitations of the existing methods.

3 Tutorial Outline

I. Introduction
- a. The socio-economic value of opinions
- b. Deceptive Opinion Spam and Fraud
- c. Opinion Spam Types: Individual, Group, Singular, and Campaigns

II. Leveraging Linguistic Signals
- a. N-gram language models
- b. Psycholinguistics
- c. Stylometry

III. Leveraging Behavioral Signals
- a. Rating, Reviewing, & Collusion Behaviors
- b. Distributional and Time-Series Analysis
- c. Graph Based Methods
- d. Linguistic vs. Behavioral Features: A case study on Commercial vs. Crowdsourced Fake Reviews

IV. Machine Leaning & Statistical Modeling
- a. Supervised vs. Unsupervised Methods
- b. Positive and Unlabeled (PU) and Semi-Supervised Learning
- c. Latent Variable Models

V. The Next Frontier: Sockpuppets
- a. Authorship Attribution and Beyond
- b. Modeling Latent Spaces of Language
- c. Learning in Similarity Spaces

VI. Discussion and Resources
- a. Applications
- b. Data sources
- c. Evaluation
- d. Discussion

4 Instructor Biography

Arjun Mukherjee is an Assistant Professor in the Department of Computer Science at the University of Houston. He is an active researcher in the area of opinion spam, sentiment analysis and Web mining. He is the lead author behind several influential works on opinion spam research. These include group opinion spam, commercial fake review filters (e.g., Yelp), and various statistical models for detecting singular opinion spammers, burstiness patterns, and campaign. His work on opinion mining including deception detection have also received significant media attention (e.g., ACM Tech News, NYTimes, LATimes, Business Week, CNet, etc[1]). Mukherjee has also served as program committee members of WWW, ACL, EMNLP, and IJCNLP.

References

G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh. 2013. Exploiting Burstiness in Reviews for Review Spammer Detection. *ICWSM*.

S. Gokhman, J. Hancock, P. Prabhu, M. Ott, and C. Cardie. 2012. In search of a gold standard in studies of deception. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*.

J. T. Hancock, L. E. Curry, S. Goorha, and M. Woodworth. 2008. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*.

N. Jindal and B. Liu. 2007. Review spam detection. *WWW*.

N. Jindal and B. Liu. 2008 Opinion Spam and Analysis. *WSDM*.

H. Li, B. Liu, A. Mukherjee, and J. Shao. 2014a. Spotting Fake Reviews using Positive-Unlabeled Learning. *Computaci{ó}n y Sistemas*, 18(3).

H. Li, A. Mukherjee, B. Liu, R. Kornfield, and S. Emery. 2014b. Detecting Campaign Promoters on Twitter using Markov Random Field. *ICDM*.

A. Mukherjee, V. Venkataraman. 2014. Opinion Spam Detection: An Unsupervised Approach using Generative Models. *UH-CS-TR-2014-07*.

A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance. 2013a. What Yelp Fake Review Filter might be Doing? *AAAI ICWSM*.

A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance. 2013b. Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews. *UIC-CS-2013-03*

A. Mukherjee, A. Kumar, B. Liu, J. Wang, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013c. Spotting Opinion Spammers using Behavioral Footprints. *KDD*.

A. Mukherjee, B. Liu, and N. Glance. 2012. Spotting Fake Reviewer Groups in Consumer Reviews. *WWW*.

M. Ott, Y. Choi, C. Cardie, and J. T Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. *ACL*.

Z. Wang. 2010. Anonymity, Social Image, and the Competition for Volunteers: A Case Study of the Online Market for Reviews. *The B.E. Journal of Economic Analysis & Policy*, 10(1):1–34, January.

[1] http://www.cs.uic.edu /~liub/FBS/media-coverage.html

What You Need to Know about Chinese for Chinese Language Processing

Chu-Ren Huang
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
churen.huang@inet.polyu.edu.hk

1 Introduction

The synergy between language sciences and language technology has been an elusive one for the computational linguistics community, especially when dealing with a language other than English. The reasons are two-fold: the lack of an accessible comprehensive and robust account of a specific language so as to allow strategic linking between a processing task to linguistic devices, and the lack of successful computational studies taking advantage of such links. With a fast growing number of available online resources, as well as a rapidly increasing number of members of the CL community who are interested in and/or working on Chinese language processing, the time is ripe to take a serious look at how knowledge of Chinese can help Chinese language processing.

The tutorial will be organized according to the structure of linguistic knowledge of Chinese, starting from the basic building block to the use of Chinese in context. The first part deals with characters as the basic linguistic unit of Chinese in terms of phonology, orthography, and basic concepts. An ontological view of how the Chinese writing system organizes meaningful content as well as how this onomasiological decision affects Chinese text processing will also be discussed. The second part deals with words and presents basic issues involving the definition and identification of words in Chinese, especially given the lack of conventional marks of word boundaries. The third part deals with parts of speech and focuses on definition of a few grammatical categories specific to Chinese, as well as distributional properties of Chinese PoS and tagging systems. The fourth part deals with sentence and structure, focusing on how to identify grammatical relations in Chinese as well as a few Chinese-specific constructions. The fifth part deals with how meanings are represented and expressed, especially how different linguistic devices (from lexical choice to information structure) are used to convey different information. Lastly, the sixth part deals with the ranges of different varieties of Chinese in the world and the computational approaches to detect and differentiate these varieties. In each topic, an empirical foundation of linguistics facts are clearly explicated with a robust generalization, and the linguistic generalization is then accounted for in terms of its function in the knowledge representation system. Lastly this knowledge representation role is then exploited in terms of the aims of specific language technology tasks. In terms of references, in addition to language resources and various relevant papers, the tutorial will make reference to Huang and Shi's (2016) reference grammar for a linguistic description of Chinese.

2 Resources

- Huang, Chu-Ren. 2009. Tagged Chinese Gigaword Version 2.0. Philadelphia: Lexical Data Consortium. University of Pennsylvania. ISBN 1-58563-516-2
- Sinica Corpus: Academia Sinica Balanced Corpus for Mandarin Chinese. http://www.sinica.edu.tw/SinicaCorpus
- Sinica BOW: Academia Sinica Bilingual Ontological Wordnet http://BOW.sinica.edu.tw
- Sinica TreeBank http://TreeBank.sinica.edu.tw/
- Chinese Wordnet 2005. http://cwn.ling.sinica.edu.tw
- Hantology 2006. http://hantology.ling.sinica.edu.tw

3 Outline

The tutorial will have six components according to the nature of linguistic knowledge of Chinese: 1)

Proceedings of the Tutorials of the 53rd Annual Meeting of the ACL and the 7th IJCNLP, pages 23–24,
Beijing, China, July 26-31, 2015.

characters, 2) words, 3) Parts of Speech, 4) Sentence and Structure, 5) Meaning: Representation and Expressive, and 6) Variations and Changes. Under each knowledge component, there will be 3 to 5 focus areas. In addition, relevant resources and language technology applications will be introduced together with the linguistic description or at the end of the lecture sections (for those language processing applications involving more than one linguistic issue.) Overall, two lecture sections of 80 minutes each will be given, each containing 5 topical groups (each topical group covers 2-3 focus areas described above). It is estimated that each topic group will take about 15 minutes to cover. Although the 15 minutes will not be enough for explication of finer details, participants will be able to access and acquire additional details from a comprehensive list references.

The three hour teaching plan is given below.

00:00-01:20 Characters, Words, and Parts-of-Speech

- -Component structure of Chinese characters: encoding and ontological issues
- -Writing system and processing of Chinese texts: myths and facts
- -Definition and identification of words in Chinese: with special foci on segmentation, and compounds
- -PoS and tagging in Chinese, with special foci on de, adjectives (or verbs), prepositions, and classifiers
- -Related issues and examples in Chinese Language processing

01:20-01:40: Coffee Break

01:40-03:00 Sentence, Meaning, and Variations

- -Aspectual and eventive systems of Chinese
- -Identification of grammatical relations: ba/bei, topic/argument, separable compounds and oblique arguments
- -Semantic relations and semantic selection
- -World Chineses: variations and changes and how to identify them
- -Related issues and examples in Chinese Language processing

4 Instructor

Chu-Ren Huang is currently a Chair Professor at the Hong Kong Polytechnic University. He is a Fellow of the Hong Kong Academy of the Humanities, a permanent member of the International Committee on Computational Linguistics, and President of the Asian Association of Lexicography. He currently serves as Chief Editor of the Journal Lingua Sinica, as well as Cambridge University Press? Studies in Natural Language Processing. He is an associate editor of both Journal of Chinese Linguistics, and Lexicography. He has served advisory and/or organizing roles for conferences including ALR, ASIALEX, CLSW, CogALex, COLING, IsCLL, LAW, OntoLex, PACLIC, ROCLING, and SIGHAN. Chinese language resources constructed under his direction include the CKIP lexicon and ICG, Sinica, Sinica Treebank, Sinica BOW, Chinese WordSketch, Tagged Chinese Gigaword Corpus, Hantology, Chinese WordNet, and Emotion Annotated Corpus. He is the co-author of a Chinese Reference Grammar (Huang and Shi 2016), and a book on Chinese Language Processing (Lu, Xue and Huang in preparation).

Author Index