LAW IX

The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015

Proceedings of the Workshop

June 5, 2015
Denver, Colorado, USA

Order print-on-demand copies from:

Curran Associates
57 Morehouse Lane
Red Hook, New York 12571
USA
Tel: +1-845-758-0400
Fax: +1-845-758-2633
curran@proceedings.com

ISBN 978-1-941643-47-1
Proceedings of the 9th Linguistic Annotation Workshop (LAW-IX)
Adam Meyers, Ines Rehbein and Heike Zinsmeister (eds.)

Introduction to the Workshop

The Linguistic Annotation Workshop (The LAW) is organised annually by the Association for Computational Linguistics' Special Interest Group for Annotation (ACL SIGANN). It provides a forum to facilitate the exchange and propagation of research results concerned with the annotation, manipulation, and exploitation of corpora; work towards harmonisation and interoperability from the perspective of the increasingly large number of tools and frameworks for annotated language resources; and work towards a consensus on all issues crucial to the advancement of the field of corpus annotation. The series is now in its ninth year, with these proceedings including papers that were presented at LAW IX, held in conjunction with the NAACL conference in Denver, Colorado, on June 5 2015.

This year's LAW IX has received 35 submissions, out of which 18 have been accepted to be presented at the workshop, either as a talk or as a poster. In addition to the papers, LAW IX also features a panel dedicated to this year's special theme, the Syntactic Annotation of Non-canonical Language. For the panel, we invited researchers who have vast experience with the manual annotation of language resources that can be described as "non-canonical", such as web data, learner language or non-standard language varieties, and who are – at the same time – aware of the problems arising when using the annotated data as training data for NLP tools or when trying to automatically predict syntactic analyses for non-canonical, noisy data. Before the workshop, we presented the panellists with a number of discussion points and asked them to write a short opinion piece addressing these issues. The resulting contributions are part of these proceedings.

Our thanks go to SIGANN, our organising committee, for its continuing organisation of the LAW workshops, and to the NAACL 2015 workshop chairs for their support. Also, we thank the NAACL 2015 publication chairs for their help with these proceedings. Most of all, we would like to thank all the authors for submitting their papers to the workshop, and our program committee members for their dedication and their thoughtful reviews.

Special Theme: Syntactic Annotation of Non-canonical Language

This year's LAW especially invited contributions addressing the special theme Syntactic Annotation of Non-canonical Language, and also features a panel dedicated to this topic. But what exactly does "non-canonical" mean? In the literature, we find different definitions that vary depending on the background and research interests of the respective research groups. Hirschmann et al. (2007), who focus on the analysis of learner data, use "non-canonical" as follows.

> "'Non-canonical' [...] refers to structures that cannot be described or generated by a given linguistic framework – canonicity can only be defined with respect to that framework. A structure may be non-canonical because it is ungrammatical, or it may be non-canonical because the given framework is not able to analyse it." (Hirschmann et al., 2007:1)

Dipper et al. (2013) use the term "(non)-standard" instead of "(non-)canonical" but emphasise that they do not intend a normative, prescriptive reading but simply refer to "de facto standard language found in newspaper texts". This definition, however, raises the follow-up question of what exactly is meant by standard, as the criterium "appears in newswire" seems to be a bit vague. A possible interpretation of

standard could be transcribed as everything that observes the (grammar) rules of the standard language variety. This approach allows us to better operationalise which structures are included and which are considered to be non-canonical. However, this definition is also not without problems. A minor problem is caused by new constructions that are used by many speakers of the language community but still considered to be ungrammatical by a large number of native speakers. But, and this is the more severe problem, how do we handle languages where we have more than one standard (e.g. British/North American/Australian/... Standard English) or where no standard exists at all? The latter is especially relevant for oral languages that have no script. Furthermore, the notion of a standard is often used to refer to the most prestigious variety of a language. As a result, language attitude also comes into play, making it even harder to arrive at an objective definition of what standard means.

While these topics have mostly been discussed in the theoretical linguistics literature in the areas of socio-linguistics or dialectology, other terms used to refer to non-canonical language in NLP include low-resourced languages and noisy data. The term "low-resourced languages" simply refers to any language for which no or only small-scale language resources exist. This means that, with respect to the definition given above, low-resourced languages can, but do not necessarily have to be non-canonical. We are not aware of any clear-cut definition for the term "noisy data". Usually, noisy data refers to text that contains spelling errors, abbreviations, non-standard words, missing punctuations, missing case information, and phenomena typical for spoken language, such as disfluencies or fillers. The term "noisy" does not distinguish between 'real' noise which was inserted unintentionally in the data (such as spelling errors, OCR errors etc.) and language features that fulfil a certain function and thus are part of the language system (but are, admittedly, challenging for NLP systems). An example are fillers which are often used as strategic devices for turn-taking and also fulfill pragmatic functions in the discourse. We would prefer to think of the latter as non-canonical (with regard to the rules of the standard variety) instead of noise.

Having shed some light on what we mean by non-canonical language, we now briefly discuss why we think this to be a relevant topic for a panel. When the first linguistically annotated corpora were built, research mostly focussed on written text from the newspaper domain. Meanwhile, also other corpora are available, including spoken language, learner data, or historical texts. The advent of Digital Humanities has further advanced this trend, and many projects exist that work with data from domains other than newspaper. Especially data from the social media has attracted lots of attention, and many new corpus projects are now under way. The new projects follow different approaches, some using existing annotation schemes as they are, others extend and adapt existing schemes to the particularities of their data, and others again invent their own scheme for annotation. Concerning the granularity of the annotations, we can also find a huge range of detail. Some use rather coarse-grained label sets while others aim at very fine-grained distinctions.

This again brings us to the question of the reliability of the annotations. There have been discussions of whether it is worthwhile employing expert annotators, given the time requirements and high costs, or whether one could achieve similar results with untrained annotators. Also, and this has been in the focus of last year's LAW: "The good, the bad, and the perfect: How good does annotation need to be?" The answer to this question is closely related to the next one: What type of annotators do we need? Is crowdsourcing reliable enough, and can it be employed efficiently for treebanking?

We think that future work on the linguistic aspects of non-canonical language as well as on processing it will benefit from a discussion on best practices for the syntactic annotation of non-standard language.

As panellists, we invited Ann Bies (LDC), Aoife Cahill (ETS), Barbara Plank (CST, University of Copenhagen) and Nathan Schneider (ILCC, University of Edinburgh), and presented them with the discussion points below.

- What are the factors that lead to the adoption of a totally new annotation scheme rather than using an existing annotation scheme?
- How do you decide on the granularity of the distinctions you choose to annotate? Give examples.
- For building new resources for NCLs, is it still worthwhile to invest a huge amount of time and human labour for manual annotation, considering that the annotators spend most of their time making arbitrary decisions, and that the aim of building 'high-quality resources' for NCLs might not be realistic?
- On a related note, what are the considerations when choosing the level of expertise of the annotators? When is crowd sourcing appropriate? When do we need linguistic experts?
- Can the concept of "gold annotations" be applied to non-canonical languages where the inherent ambiguity in the data makes it hard to decide on the "ground truth" of an utterance?

The resulting papers are part of the proceedings. We would like to thank the panellists for their insightful contributions and hope that this will foster future discussions on that matter.

Adam Meyers, Ines Rehbein and Heike Zinsmeister, program co-chairs

LAW Co-chairs:

Adam Meyers, New York University
Ines Rehbein, Potsdam University
Heike Zinsmeister, University of Hamburg

Organising Committee:

Stefanie Dipper, Ruhr University Bochum
Chu-Ren Huang, The Hong Kong Polytechnic University
Nancy Ide, Vassar College
Lori Levin, Carnegie-Mellon University
Antonio Pareja-Lora, SIC & ILSA, UCM / ATLAS, UNED
Massimo Poesio, University of Trento
Sameer Pradhan, Harvard University
Manfred Stede, University of Potsdam
Katrin Tomanek, VigLink Inc.
Fei Xia, University of Washington
Nianwen Xue, Brandeis University

Program Committee:

Collin Baker, UC Berkeley
Ann Bies, LDC
Archna Bhatia, Carnegie Mellon University
Marie Candito, Université Paris Diderot - INRIA
Özlem Çetinoğlu, University of Stuttgart
Christian Chiarcos, University of Frankfurt
Markus Dickinson, Indiana University
Stefanie Dipper, Ruhr University Bochum
Tomaž Erjavec, Josef Stefan Institute
Kilian Evang, University of Groningen
Pablo Faria, Universidade Estadual de Campinas
Jennifer Foster, Dublin City University
Andrew Gargett, University of Birmingham
Kim Gerdes, Sorbonne Nouvelle, Paris 3
Nizar Habash, New York University Abu Dhabi
Udo Hahn, University of Jena
Chu-Ren Huang, The Hong Kong Polytechnic University
Nancy Ide, Vassar College
Aravind Joshi, University of Pennsylvania
Varada Kolhatkar, University of Toronto

Valia Kordoni, Humboldt University Berlin
Sandra Kübler, Indiana University
John S. Y. Lee, City University of Hong Kong
Els Lefever, University College Ghent
Lori Levin, Carnegie-Mellon University
Amália Mendes, Universidade di Lisboa
Anna Nedoluzhko, Chareles University Prague
Kemal Oflazer, Carnegie-Mellon University, Qatar
Lilja Øvrelid, University of Oslo
Alexis Palmer, University of Stuttgart
Antonio Pareja-Lora, SIC & ILSA, UCM / ATLAS, UNED
Massimo Poesio, University of Trento
Sameer Pradhan, Harvard University
James Pustejovsky, Brandeis University
Arndt Riester, University of Stuttgart
Benoît Sagot, Inria, Université Paris 7
Nathan Schneider, Carnegie-Mellon University
Djamé Seddah, Université Paris Sorbonne & INRIA's Alpage Project
Kiril Simov, Bulgarian Academy of Sciences, Sofia
Anders Søgaard, University of Copenhagen
Caroline Sporleder, University of Trier
Manfred Stede, University of Potsdam
Joel Tetrault, Yahoo! Labs
Katrin Tomanek, VigLink Inc.
Reut Tsarfaty, Weizmann Institute of Science, Israel
Yulia Tsvetkov, Carnegie-Mellon University
Andreas Witt, IDS Mannheim
Fei Xia, University of Washington
Nianwen Xue, Brandeis University

Panellists:

Ann Bies, LDC
Aoife Cahill, ETS
Barbara Plank, University of Copenhagen
Nathan Schneider, University of Edinburgh

Table of Contents

Workshop Program

Friday, June 5, 2015

8:45–9:00 ***Opening Remarks***

9:00–10:30 ***Session 1***

Oral Presentations

9:00–9:30 *Scaling Semantic Frame Annotation*
Nancy Chang, Praveen Paritosh, David Huynh and Collin Baker

9:30–10:00 *An Analytic and Empirical Evaluation of Return-on-Investment-Based Active Learning*
Robbie Haertel, Eric Ringger, Kevin Seppi and Paul Felt

10:00–10:30 *Annotating genericity: a survey, a scheme, and a corpus*
Annemarie Friedrich, Alexis Palmer, Melissa Peate Sørensen and Manfred Pinkal

10:30–11:00 ***Coffee break***

11:00–12:30 ***Session 2***

Poster presentations

Design and Annotation of the First Italian Corpus for Text Simplification
Dominique Brunato, Felice Dell'Orletta, Giulia Venturi and Simonetta Montemagni

On the Discoursive Structure of Computer Graphics Research Papers
Beatriz Fisas, Horacio Saggion and Francesco Ronzano

Semantic Annotation of Japanese Functional Expressions and its Impact on Factuality Analysis
Yudai Kamioka, Kazuya Narita, Junta Mizuno, Miwa Kanno and Kentaro Inui

A Qualitative Analysis of a Corpus of Opinion Summaries based on Aspects
Roque Lopez, Thiago Pardo, Lucas Avanço, Pedro Filho, Alessandro Bokan, Paula Cardoso, Márcio Dias, Fernando Nóbrega, Marco Cabezudo, Jackson Souza, Andressa Zacarias, Eloize Seno and Ariani Di Felippo

Friday, June 5, 2015 (continued)

Developing Language-tagged Corpora for Code-switching Tweets
Suraj Maharjan, Elizabeth Blair, Steven Bethard and Thamar Solorio

Annotating Geographical Entities on Microblog Text
Koji Matsuda, Akira Sasaki, Naoaki Okazaki and Kentaro Inui

The Annotation Process of the ITU Web Treebank
Tuğba Pamay, Umut Sulubacak, Dilara Torunoğlu-Selamet and Gülşen Eryiğit

Part of Speech Annotation of Intermediate Versions in the Keystroke Logged Translation Corpus
Tatiana Serbina, Paula Niemietz, Matthias Fricke, Philipp Meisen and Stella Neumann

A Hierarchy with, of, and for Preposition Supersenses
Nathan Schneider, Vivek Srikumar, Jena D. Hwang and Martha Palmer

Bilingual English-Czech Valency Lexicon Linked to a Parallel Corpus
Zdenka Uresova, Ondřej Dušek, Eva Fucikova, Jan Hajic and Jana Sindlerova

Correction Annotation for Non-Native Arabic Texts: Guidelines and Corpus
Wajdi Zaghouani, Nizar Habash, Houda Bouamor, Alla Rozovskaya, Behrang Mohit, Abeer Heider and Kemal Oflazer

12:30–14:00 ***Lunch break***

14:00–15:30 ***Session 3***

Friday, June 5, 2015 (continued)

Panel

Balancing the Existing and the New in the Context of Annotating Non-Canonical Language
Ann Bies

Parsing Learner Text: to Shoehorn or not to Shoehorn
Aoife Cahill

Non-canonical language is not harder to annotate than canonical language
Barbara Plank, Héctor Martínez Alonso and Anders Søgaard

What I've learned about annotating informal text (and why you shouldn't take my word for it)
Nathan Schneider

16:00–18:00 ***Session 4***

Oral presentations

16:00–16:30 *On Grammaticality in the Syntactic Annotation of Learner Language*
Markus Dickinson and Marwa Ragheb

16:30–17:00 *Across Languages and Genres: Creating a Universal Annotation Scheme for Textual Relations*
Ekaterina Lapshinova-Koltunski, Anna Nedoluzhko and Kerstin Anna Kunz

17:00–17:30 *Annotating the Implicit Content of Sluices*
Pranav Anand and Jim McCloskey

17:30–18:00 *Annotating Causal Language Using Corpus Lexicography of Constructions*
Jesse Dunietz, Lori Levin and Jaime Carbonell

18:00–18:10 ***Closing***