Preface

TLSX: The Proceedings

NICK GAYLORD, ALEXIS PALMER, ELIAS PONVERT

Computational linguistics for less-studied languages

The 10th annual Texas Linguistics Society (TLS) conference was held in November 2006, hosted by the department of Linguistics at the University of Texas at Austin. The conference topic was computational linguistics for less-studied languages.

The past decade has seen great developments in technologies for language documentation, particularly in the focus areas of speech and video recording and transcription, best practices for data collection and archiving, and ontology development. One aim of TLSX was to expand to other focus areas, particularly at the intersection of computational linguistics and language documentation, highlighting the application of techniques from computational linguistics to the management and analysis of language data for less-studied languages or less-studied varieties of well-studied languages.

TLSX brought together researchers from two largely unconnected disciplines, computational linguistics and documentary and descriptive linguistics. Many computational linguists are interested in theoretical issues such as the application of data-driven natural language processing (NLP) techniques to languages for which there exists relatively little digitally-available data. At the same time, many documentary and descriptive linguists are interested in improving technologies for language documentation and analysis. Through this venue, both sides became better aware of the challenges and needs of the other, as well as the real potential gains, both practical and intellectual, from interaction between the two.

The University of Texas is a uniquely appropriate venue for a workshop of this nature. Over the last decade, Texas has built a strong inter-disciplinary

Texas Lingusitics Society 10.
.

program in documentary and descriptive linguistics. Computational linguistics is also well-represented at UT, involving faculty and students with a variety of interests in the departments of Linguistics, Computer Sciences, and others.

Conference program

The conference program contained six keynote presentations and ten peer-reviewed talks chosen from submitted papers. The program also included two panel discussions, with one panel comprised of documentary and descriptive linguists and the other of computational linguists.

Keynote addresses

Stephen Bird (University of Melbourne, University of Pennsylvania) presented the first keynote address, entitled 'Linguistic Data Management with the Natural Language Toolkit'. The talk combined demonstrations of rapid, facile, and flexible data management using the Natural Language Toolkit and the Python programming language with discussion of existing data management issues and potential directions for development of technologies to address those issues. In addition, Bird encouraged researchers in both fields to broaden their approaches by taking some initial steps into the other field. For example, documentary/descriptive linguists might learn and apply some basic programming techniques, and computational linguists might get some field experience or develop more comprehensive views of technology development, consolidation, and reuse.

The keynote by Jason Baldridge (UT Austin Linguistics) was entitled 'Cutting Corpus Costs: Machine Learning and Annotation.' Baldridge gave an overview of different scenarios which may be encountered in the process of creating linguistically informative annotations for primary language data, and discussed two classes of approaches from computational linguistics which have the potential to reduce the cost of creating annotated corpora: semi-automated annotation and active learning. Semi-automated annotation is a two step process. First the data is labeled automatically by a system trained to do such annotation. The labels provided by the system, certain to contain many errors, are then corrected by a human expert. Active learning, on the other hand, identifies in the output of the automated labeler the most difficult examples, those which are most difficult for the machine to label given the currently-available training data. These examples are then manually labeled by the human expert and used to train a new, better-performing system.

The next keynote address came from Emily Bender (University of Washington), who spoke on 'The Grammar Matrix: A Crosslinguistic Resource to Promote Grammar Engineering for Linguistic Hypothesis Testing.' The

Grammar Matrix, a toolkit developed by Bender and her colleagues, facilitates rapid startup for grammar development, reducing the time required to produce a working grammar fragment for a language. This in turn supports the goal of establishing a database of grammatical patterns and constructions. As such, this technology provides a serious attempt to apply large-scale cross-linguistic engineering to the statement and evaluation of linguistic hypotheses. Bender's paper is included in these proceedings.

Katrin Erk's (UT Austin Linguistics) keynote address was titled 'Detecting Outliers: Useful for word sense assignment – and for aiding manual annotation?'. Many occurrences of a word will be readily describable by one of the existing senses listed in a dictionary, but some will not; by automatically identifying such cases, the researcher can focus his or her efforts on those points in the data that need particular attention while streamlining the rest of the process. Erk argued that the known usefulness of outlier detection for particular computational linguistic tasks could be extended to aid in the manual annotation of other types of data, including language documentation, by allowing researchers to focus their efforts on less-well-understood issues in their data.

The fifth keynote, by Raymond Mooney (UT Austin Computer Science), 'Maximizing the Utility of Small Training Sets in Machine Learning', addressed a crucial issue for researchers wishing to use standard natural language processing techniques on less-studied languages: many of those techniques are extremely data-intensive. Mooney surveyed five machine learning strategies suitable for use in contexts with relatively small amounts of available training data. For each of the five strategies – ensemble methods, active learning, transfer learning, unsupervised learning, and semisupervised learning – Mooney discussed the general learning strategy as well as specific examples of its application to small-data situations.

Mark Liberman (University of Pennsylvania) gave the final keynote, 'The Problems of Scale in Language Documentation.' The talk opened by asking how much and what kind of language data is needed to document a language, and in particular an endangered language. The proposed figure was 100-1000 hours of spoken language data, varying according to the nature and goals of the documentation effort. Liberman presented several ideas for recording this amount of data rapidly and in a maximally-informative manner and then opened the discussion up to the audience. From this discussion emerged a suggested mindset for tackling the problems of scale in language documentation: the linguistic community should view the collection of adequate amounts of data as a problem that must be solved. Accordingly, we should focus attention on budgetary realities and constraints, identifying where we are spending the most time and labor and then finding ways to reduce those expenses.

Individual talks

The first session consisted of two talks focused on tools and other resources for language documentation. Dorothee Beermann and Atle Prange presented TypeCraft, an online tool for producing interlinearized glosses of natural language sentences and phrase-level tokens. In addition to facilitating peer review of annotation, TypeCraft aims to develop a repository of data from less-studied languages. H. Andrew Black and Gary F. Simons presented FieldWorks Language Explorer (FLEX) and its approach to morphological parsing. One aim of FLEX is to replace the simplistic pattern matching of Shoebox/Toolbox applications with a more sophisticated model for morpheme parsing and segmentation.

The talks in the second session discussed applications of techniques from computational linguistics to language-specific analysis problems. Vijay John presented his algorithm for enhancing search in Mandarin Chinese using transliteration. John also showed the algorithm's potential for improving search in a number of other languages. Frederick Hoyt applied Maximum Entropy part-of-speech tagging models to the problem of vocalization in Arabic Text. The talk addressed inflectional rather than derivational or lexical vocalization, and thus pertains to phenomena such as subject-verb agreement, case marking on nouns, nominal definiteness, and verbal mood. One key insight of Hoyt's work is to model inflectional vocalization in Arabic as a tagging task, following from the fact that vocalization is largely determined by syntactic context, the same sort of context that is important to part-of-speech taggers.

The next session's talks were again language-specific, one addressing grammar development in a Mayan language, the other two using computational techniques for grammar development and morphological analysis of American Sign Language (ASL). Elias Ponvert presented a fragment of the Mayan language Popti' in Combinatory Categorial Grammar (CCG), offering original analyses of two grammatical phenomena: constraints on relative clause and topic formation, and incorporated pronouns. Ponvert's analysis was implemented using OpenCCG. Tony Wright presented a grammar fragment of ASL, also offering an analysis using CCG and implemented in OpenCCG. Wright specifically addressed ASL multiple embeddings of topic-comment structures and the spatial-path morphology used to express thematic relations in ASL. Ponvert's and Wright's analyses both illustrate the effectiveness of grammar engineering toward establishing and testing linguistic hypotheses. Aaron Shield's presentation was also concerned with ASL; in joint work with Jason Baldridge, Shield presented a finite-state morphological analyzer for ASL verbs which relates surface forms to abstract formal representations, implemented in the Xerox Finite State Toolkit. Shield and Baldridge's work is a novel application of finite-state technology to the

phonology of signed languages, and involved a new and potentially useful representation of sign forms, amenable to phonological analysis.

The final session contained two talks on morphological analysis and a talk on parse projection. Alfonso Medina-Urrea presented a technique for affix discovery using entropy and economy measurements. The talk included results from evaluations of the method on two unrelated languages of the Americas, Ralámuli (Uto-Aztecan) and Chuj (Mayan). Dan Jinguji presented a joint paper by William Lewis, Fei Xia and himself on annotating and enriching data for lesser-studied languages via alignment and projection of structure from other sources: namely, from annotated and parsed data for resource-rich languages such as English. This work featured not only the novel application of recent machine learning methodology to languages with little pre-existing annotation, but along the way made use of the large body of Web-based language data, toward ultimately providing a kind of linguistic structure search. Finally, Robert Elwell presented an analysis of the verbal morphology of the Bantu language Ekegusii using finite state methods, with an implementation of the analysis done in XFST.

The panel discussions

Each hour-long panel discussion focused on one of the TLSX focus areas. For the sake of brevity, we present here only the names of the participants and the organizing concepts and/or questions for each panel. The first panel discussion took place Friday afternoon, and the second closed the conference on Sunday afternoon.

The first panel, comprised of linguists with extensive experience in linguistic fieldwork, discussed their own needs with respect to technologies for language documentation and description. Some key focus areas were gaps in the current tool set, problems of current technologies, and the variability of needs according to the characteristics of the language being studied. Panelists were Nora England, Pattie Epps, Liberty Lidz, B'alam Mateo-Toledo, and Christina Willis, and the panel was moderated by Tony Woodbury.

The second panel was made up of the six keynote speakers. The panelists spoke primarily from the point of view of the computational linguist, aiming to answer questions like the following: how can computationial linguistics address the needs of documentary and descriptive linguistics, and how will doing so further the state of research in the field of computational linguistics? What are fruitful directions for future research? What sorts of collaborations would be useful for both subfields? Where do we go from here?

Acknowledgments

Perhaps the most exciting outcome of TLSX was the engaging conversation that developed between documentary/descriptive linguists and computational linguists. For this we thank all of our speakers and conference attendees. Thanks too to panelists and session chairs: Pascal Denis, Nora England, Pattie Epps, Liberty Lidz, B'alam Mateo-Toledo, Carlota Smith, Steve Wechsler, Christina Willis, and Tony Woodbury.

We are very grateful for the support of our two faculty advisors: Jason Baldridge for assistance with reviewing and conference organization, and Steve Wechsler for help with the conference proceedings and as our liaison with CSLI. We also thank CSLI for hosting the online proceedings of the TLS conference series.

Thanks to the program committee, comprised of Steven Abney, Jason Baldridge, Emily Bender, Steven Bird, Cem Bozsahin, Inge de Bleecker, Katrin Erk, Jeff Good, Frederick Hoyt, Jonas Kuhn, Terry Langendoen, William Lewis, Mark Liberman, Liberty Lidz, Chris Manning, Raymond Mooney, Martha Palmer, Alexandré Sevigny, Gary Simons, Mark Steedman, and Tony Woodbury.

We received invaluable support of all sorts from Brian Price and Ben Rapstine in the Linguistics department, and Candace Pruett provided home-baked breakfast goodies! Thanks also to all of our student volunteers.

Finally, enormous thanks go to the sponsors of TLSX: UT Austin Department of Linguistics, 21st Century Technologies, The Artificial Intelligence Lab at UT Austin, the UT Austin Center for Middle Eastern Studies, and the UT Austin Department of Asian Studies.