

Overview of NLP/CL Fields

Frank Ferraro – ferraro@umbc.edu

CMSC 473/673

Today's Learning Goals

- Learn about where to find NLP resources/how the community works
- Learn about the many depths of language (the “iceberg”)

The NLP Research Community

- Papers

- ACL Anthology (<http://aclweb.org/anthology>) has nearly everything, free! As of 2016:
 - Over 36,000 papers!
 - Free-text searchable
 - Great way to learn about current research on a topic
 - New search interfaces currently available in beta
 - » Find recent or highly cited work; follow citations
 - Used as a dataset by various projects
 - Analyzing the text of the papers (e.g., parsing it)
 - Extracting a graph of papers, authors, and institutions (Who wrote what? Who works where? What cites what?)

The NLP Research Community

- **Conferences**

- Most work in NLP is published as 8-page conference papers with 3 double-blind reviewers.
- Main annual conferences: ACL, EMNLP, NAACL
 - Also EACL, IJCNLP, COLING
 - + journals (TACL, *Computational Linguistics* [CL])
 - + various specialized conferences and workshops
- Big events, and growing fast! [ACL 2015](#):
 - About 1500 attendees
 - 692 full-length papers [submitted](#) (173 accepted)
 - 648 short papers submitted (145 accepted)
 - 14 [workshops](#) on various topics

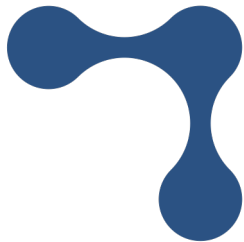
The NLP Research Community

- Standard tasks

- If you want people to work on your problem, make it easy for them to get started and to measure their progress. Provide:

- Test data, for evaluating the final systems
 - Development data, for measuring whether a change to the system helps, and for tuning parameters
 - An evaluation metric (formula for measuring how well a system does on the dev or test data)
 - A program for computing the evaluation metric
 - Labeled training data and other data resources
 - A prize? – with clear rules on what data can be used

The NLP Research Community: Standard Tasks (some of them)



GLUE

<https://gluebenchmark.com/>

GLUE Tasks

Name	Download
The Corpus of Linguistic Acceptability	
The Stanford Sentiment Treebank	
Microsoft Research Paraphrase Corpus	
Semantic Textual Similarity Benchmark	
Quora Question Pairs	
MultiNLI Matched	
MultiNLI Mismatched	
Question NLI	
Recognizing Textual Entailment	
Winograd NLI	
Diagnostics Main	

SuperGLUE Tasks

Name	Identifier
Broadcoverage Diagnostics	AX-b
CommitmentBank	CB
Choice of Plausible Alternatives	COPA
Multi-Sentence Reading Comprehension	MultiRC
Recognizing Textual Entailment	RTE
Words in Context	WiC
The Winograd Schema Challenge	WSC
BoolQ	BoolQ
Reading Comprehension with Commonsense Reasoning	ReCoRD
Winogender Schema Diagnostics	AX-g



<https://super.gluebenchmark.com/>

The NLP Research Community: Standard Tasks (some of them)

<https://semeval.github.io/SemEval2021/tasks>

(past years too!)

Lexical semantics

- **Task 1: Lexical Complexity Prediction** ([email organizers] [mailing list])
Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, Marcos Zampieri
- **Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation** ([email organizers])
Federico Martelli, Michael Bianchi, Gabriele Tola, Roberto Navigli
- **Task 3: Span- and Dependency-based Multilingual and Cross-lingual Semantic Role Labeling**
- **Task 4: Reading Comprehension of Abstract Meaning** ([email organizers] [mailing list])**
Xiaodan Zhu, Boyuan Zheng, Yu-Ping Ruan, Zhen-Hua Ling, Quan Liu, Si Wei

Social factors & opinion

- **Task 5: Toxic Spans Detection** ([email organizers] [mailing list])
John Pavlopoulos, Ion Androutsopoulos, Jeffrey Sorensen, Léo Laugier
- **Task 6: Detection of Persuasive Techniques in Texts and Images - website coming soon! (Sept 2)** ([email organizers] [mailing list])
Giovanni Da San Martino, Hamed Firooz, Preslav Nakov, Fabrizio Silvestri
- **Task 7: Hahackathon: Incorporating Demographic Factors into Shared Humor Tasks** ([email organizers])
J. A. Meaney, Steven Wilson, Walid Magdy, Luis Chiruzzo

Information in scientific & clinical text

- **Task 8: MeasEval: Counts and Measurements** ([email organizers] [mailing list])
Corey Harper, Jessica Cox, Ron Daniel, Paul Groth, Curt Kohler, Antony Scerri
- **Task 9: Statement Verification and Evidence Finding with Tables** ([email organizers] [mailing list])
Nancy Xin Ru Wang, Sara Rosenthal, Marina Danilevsky, Diwakar Mahajan
- **Task 10: Source-Free Domain Adaptation for Semantic Processing** ([email organizers] [mailing list])
Steven Bethard, Egoitz Laparra, Timothy Miller, Özlem Uzuner
- **Task 11: NLPContributionGraph** ([email organizers] [mailing list])
Jennifer D'Souza, Sören Auer, Ted Pederson

Other phenomena

- **Task 12: Learning with Disagreements** ([email organizers] [mailing list])
Alexandra Uma, Tommaso Fornaciari, Barbara Plank, Edwin Simpson, Massimo Poesio

The NLP Research Community

- Software

- Lots of people distribute code for these tasks
 - Or you can email a paper's authors to ask for their code
- Some [lists](#) of software, but no central site ☹️
- Some [end-to-end pipelines](#) for text analysis
 - “One-stop shopping”
 - Cleanup/tokenization + morphology + tagging + parsing + ...
 - [NLTK](#) is easy for beginners and has a [free book](#) (intersession?)
 - [GATE](#) has been around for a long time and has a bunch of modules

The NLP Research Community

- **Software**

- To find good or popular tools:
 - Search current papers, ask around, use the web
- Still, often hard to identify the **best** tool for your job:
 - Produces appropriate, sufficiently detailed output?
 - Accurate? (on the measure you care about)
 - Robust? (accurate on your data, not just theirs)
 - Fast?
 - Easy and flexible to use? Nice file formats, command line options, visualization?
 - Trainable for new data and languages? How slow is training?
 - Open-source and easy to extend?

The NLP Research Community

- **Datasets**

- Raw text or speech corpora
 - Or just their [n-gram counts](#), for super-big corpora
 - Various languages and genres
 - Usually there's some metadata (each document's date, author, etc.)
 - Sometimes $\exists$ licensing restrictions (proprietary or copyright data)
- Text or speech with manual or automatic annotations
 - What kind of annotations? That's the rest of this lecture ...
 - May include translations into other languages
- Words and their relationships
 - [Morphological](#), [semantic](#), translational, evolutionary
- [Grammars](#)
- [World Atlas of Linguistic Structures](#)
- Parameters of statistical models (e.g., grammar weights)

The NLP Research Community

- **Datasets**

- Read papers to find out what datasets others are using
 - [Linguistic Data Consortium](#) (searchable) hosts many large datasets
 - Many projects and competitions post data on their websites
 - But sometimes you have to email the author for a copy
- [CORPORA mailing list](#) is also good place to ask around
- [LREC Conference](#) publishes papers about new datasets & metrics
- [Amazon Mechanical Turk](#) – pay humans (very cheaply) to annotate your data or to correct automatic annotations
 - **Old task, new domain:** Annotate parses etc. on *your* kind of data
 - **New task:** Annotate something new that you want your system to find
 - **Auxiliary task:** Annotate something new that your system may benefit from finding (e.g., annotate subjunctive mood to improve translation)
- Can you make annotation so much [fun](#) or so [worthwhile](#) that they'll do it for free?

The NLP Research Community

- **Standard data formats**
 - Often just simple *ad hoc* text-file formats
 - Documented in a README; easily read with scripts
 - Some standards:
 - [Unicode](#) – strings in any language (see [ICU](#) toolkit)
 - PCM (.wav, .aiff) – uncompressed audio
 - BWF and AUP extend w/metadata; also many compressed formats
 - [XML](#) – documents with embedded annotations
 - [Text Encoding Initiative](#) – faithful digital representations of printed text
 - [Protocol Buffers](#), [JSON](#) – structured data
 - [UIMA](#) – “unstructured information management”; Watson uses it
 - Standoff markup: raw text in one file, annotations in other files (“ $\exists$ noun phrase from byte 378—392”)
 - Annotations can be independently contributed & distributed

The NLP Research Community

- **Survey articles**
 - May help you get oriented in a new area
 - Synthesis Lectures on Human Language Technologies
 - Handbook of Natural Language Processing
 - Oxford Handbook of Computational Linguistics
 - Foundations & Trends in Machine Learning
 - Survey articles in journals – JAIR, CL, JMLR
 - ACM Computing Surveys?
 - Online tutorial papers
 - Slides from tutorials at conferences
 - Textbooks

To Write A Typical Paper

- Need some of these ingredients:
 - A domain of inquiry Scientific or engineering question
 - A task Input & output representations, evaluation metric
 - Resources Corpora, annotations, dictionaries, ...
 - A method for training & testing Derived from a model?
 - An algorithm
 - Analysis of results Comparison to baselines & other systems, significance testing, learning curves, ablation analysis, error analysis
- There are other kinds of papers too: theoretical papers on formal grammars and their properties, new error metrics, new tasks or resources, etc.