

The Future of Natural Language Processing

Elizabeth D. Liddy

Center for Natural Language Processing
Syracuse University

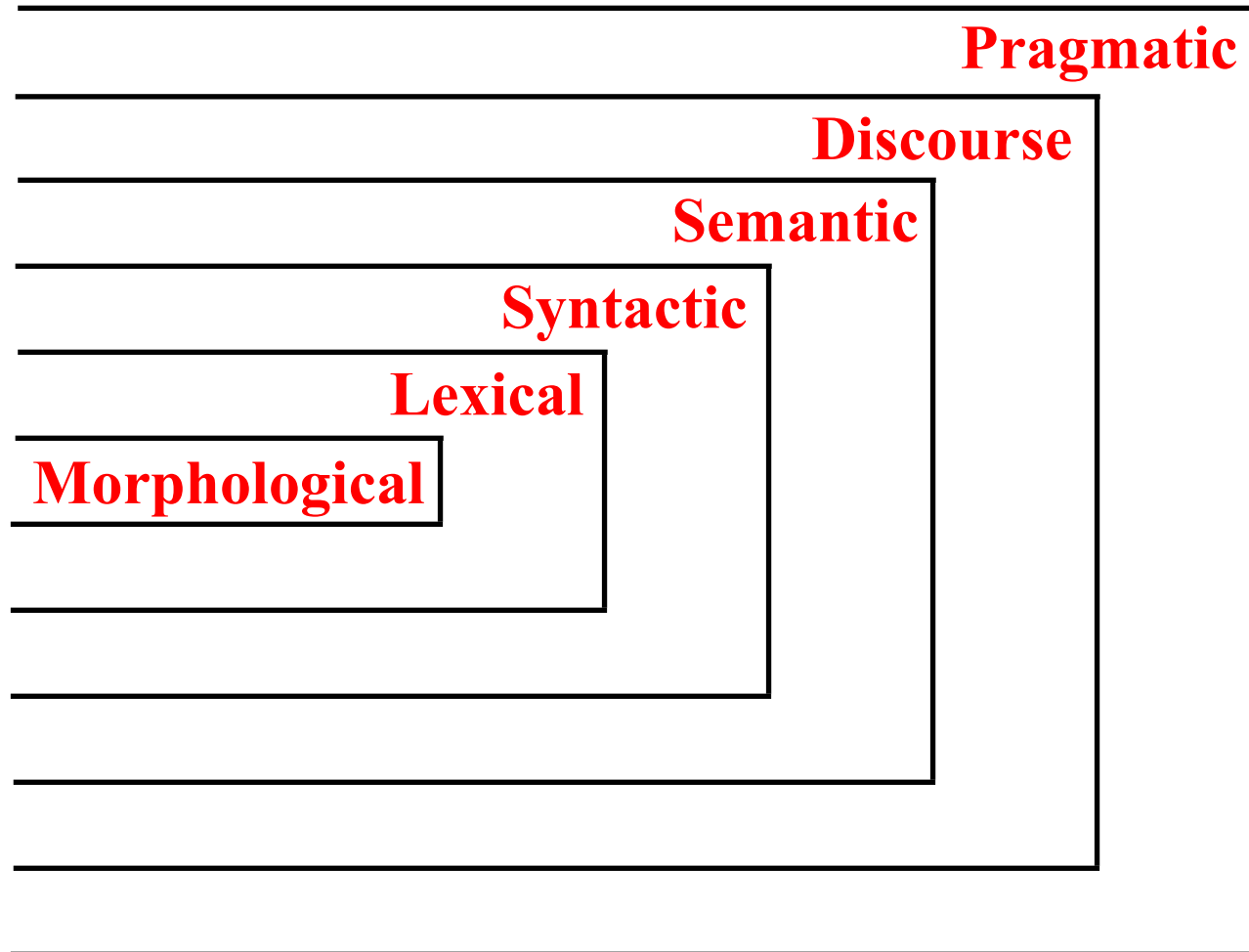
Overview

- Reflect up to a higher level of abstraction of the presentations across the past 2 days
 - Seeded by my pre-workshop thoughts
 - Little emphasis on specific technical details
 - Scant mention of our work at CNLP
- Capturing what we've heard in terms of:
 1. Needs from DHS, analysts, and their customers
 2. Goals from researchers that reflect current state of the art
- Results in a merged Needs / Goals inventory

Need / Goal I

- Accomplish higher levels of language analysis
- To-date, have achieved good performance on lower levels
 - Morphology, Lexical, Syntax, and some Semantics
- Now grapple with
 - Semantics, Discourse, Pragmatics (more than just knowledge bases)

Levels of Language Understanding



Differences in lower & higher levels

- Lower
 - Smaller units of analysis
 - Less variation
 - Rule-oriented
- Higher
 - Larger units of analysis
 - Great variation possible
 - Regularity-oriented
 - Take ever-widening context into account

Need / Goal II

- Ability to understand & utilize the implicit 'information' a text conveys
 - Not TOPIC or SUBJECT
 - Emotive / affective / opinion / evaluative aspects
 - Negative or positive attitude of reporter / participant
 - Right or left political leaning of speaker
 - Certainty or uncertainty about what's reported
- Capture *Intentionality*
 - Communicative goals of speaker / writer are key
 - Discourse / Pragmatic levels of linguistics provide theories
 - Gricean Maxims
 - Speech Acts

Need / Goal IIb

- Implicit information is of increasing importance as number and range of informal sources explodes
 - Blogs
 - Message boards
 - Discussion groups
 - Homegrown web sites
- Speech input uses additional prosodic features to convey implicit information
 - Surprise, pleasure, disdain, doubt, etc

Need / Goal III

- Contextualized NLP
 - temporal / localized production and analysis
- Text analysis:
 - for a particular person
 - for a specific purpose
 - at a given time
 - in a particular place
 - (user / task / time / history)
- Potential metadata elements for re-use

Need / Goal IV

- Information extraction of not just pre-defined generic set of extractions
 - Insufficiently individualized for needs of users
 - Specialization brings increased informativeness & value
 - Requires more modeling of the domain - portability
 - Specialized content-based metadata schemas to drive extraction
 - Trends towards learning algorithms
 - For quick specialization to domain
 - Utilizing just local, lower-level linguistic structure
 - Need to maintain means to incorporate knowledge of analysts
 - Temporal extraction and tagging is vital
 - Fuzzy time computed from non-specific text references
 - Timelines for understanding event sequences

Need / Goal V

- Opposing needs to expand and abbreviate text
 - Based on analyst's task
- Expand current mention with profile of person
 - Requires that all types of reference, both explicit and implicit, be resolved accurately
 - Requires full, rich, accurate IE
- Abbreviate across numerous reports
 - Query-specific
 - Summarizing salient aspects based on specific need

Need / Goal VI

- Primacy of users' needs / practices
- Specializable by end-users
 - Don't need to know NLP
 - Know their own domain and application and use this knowledge without knowing about NLP
- Tools to enable users to:
 - Understand nature of data they're analyzing
 - Adapt rules / models with simple highlighting of examples and desired output
- Move towards user guidance vs. full automation
 - SPUD tool
 - User involvement improves IR
 - Automatic content-based metadata generation
- Transparency of what the NLP system has done
 - More sharing of what the system is doing supports understanding & trust

Need / Goal VII

- Must solve same problem that plagues social scientists
 - Work alone or in small groups
 - Grapple with voluminous field notes, case studies, interviews
 - But can deal with only a few due to effort involved in learning & using current commercial content-analytic tools
 - Do much hand coding and analysis
 - Results in overly specific models based on a few instances
- Small science → big science
- Desired capabilities
 - Richer analysis of more data more efficiently to provide more conceptual insight into data
 - Room for easily incorporating researchers' insights
- Goal → more explanatory, predictive models

Need / Goal VIII

- There has been a pendulum swing
 - Early days of AI when applications were domain-dependent
 - Critiqued for brittleness & lack of transportability
- General, all-purpose applications
 - Factoid QA is doable but not as useful for real world tasks
 - What happens when you're not querying against huge collections and redundancy doesn't help?
 - Generic summarization is not as useful as task specific
 - 300 page public health reports need model-specific elements driving the summarization strategy
- Now swinging back to domain-dependent solutions
 - Require re-usable systems, methodologies & tools for quickly adapting to new domains

Need / Goal IX

- **Speech-based NLP Applications**
 - Vital input & access option in certain environments
 - Have greatest growth potential where it is most natural
 - Foreign Disclosure Officers in the field
 - ASR has the potential for significant improvement
 - Currently under-utilizes higher levels of language
 - Can use Discourse & Pragmatic levels of NLP for interpreting acoustic signal
 - E.g. Understanding communicative goal of a speaker enables better interpretation at phoneme level → leads to more accurate ASR
 - Promising results for information analysis for the blind

Need / Goal X

- Multi-linguality - All information is not in English
 - CLIR - queries in one language retrieve documents in other language(s)
 - Translation needed for matching and reading
 - Allow monolingual searchers to judge relevance based on machine translated results
 - CLIR performs below Monolingual IR
 - MT, IE, transliteration efforts required
 - Capabilities that assist monolingual NLP are even more powerful here
 - Keeping user involved

Current Enquiry: Iraq board of elections Language: English

[Include All](#) [Exclude All](#)

Include	Source Language	Target Language	Part-of-Speech	Sense
<input type="checkbox"/>	board	شامل	adjective	broad in scope or content
<input type="checkbox"/>	board	محيط	adjective	broad in scope or content
<input type="checkbox"/>	board	جامع	adjective	broad in scope or content
<input type="checkbox"/>	board	كلي	adjective	broad in scope or content
<input type="checkbox"/>	board	أضاف	verb	live and take one's meals at or in
<input type="checkbox"/>	board	ركب	verb	get on board of (trains, buses, ships, aircraft, etc.)
<input type="checkbox"/>	board	امتطى	verb	get on board of (trains, buses, ships, aircraft, etc.)
<input type="checkbox"/>	board	مجلس	noun	a committee having supervisory powers
<input type="checkbox"/>	board	لجنة	noun	a committee having supervisory powers
<input type="checkbox"/>	board	جلسة	noun	a committee having supervisory powers
<input type="checkbox"/>	board	خوران	noun	food or meals in general
<input type="checkbox"/>	election	انتخاب	noun	a vote to select the winner of a position or political office

Summary

- Are these your needs? Are these our goals?
- If not, what else should be added?