Language Generation 4: Discourse

CSCI 544
1/24/2012

Dirk Hovy
Words You’ll Hear Today

• pragmatics
• coherence
• coreference
• discourse structure
Absurdity like that I can deal with. I just want someone to fix my sh*t if I'm in a car accident. It's not always going to make sense or depict a situation that a normal person could relate to. Television is television, be it a sitcom or an ad for Popeye's Chicken. It's a given that nobody watches television commercials expecting to receive a heaping dose of reality. I really don't care if the Geico Caveman gets to suit up with the Washington Redskins or not.

www.cracked.com
It's a given that nobody watches television commercials expecting to receive a heaping dose of reality. Television is television, be it a sitcom or an ad for Popeye's Chicken. It's not always going to make sense or depict a situation that a normal person could relate to. I really don't care if the Geico Caveman gets to suit up with the Washington Redskins or not. I just want someone to fix my sh*t if I'm in a car accident. Absurdity like that I can deal with.
What’s All This Then?

• sentence ≠ random word order → syntax
• text ≠ random sentence order → pragmatics
• criterion: coherence
• coherent set of sentences = discourse
The Problem

- how to make texts coherent: labels/vocabulary?

My hovercraft is full of eels

You want a toe?
I can get you a toe, believe me.
There are ways, Dude.
You don't wanna know about it, believe me...

syntax

pragmatics
Coherence

• meaningful combination of sentences

• means to “link” sentences:
  • coreference
  • discourse relations
Coreference

- referring to one thing with multiple expressions
- crucial for all NLP tasks!
- intensive psycholinguistic research
Chomsky doesn’t believe in probabilities. He’d better! But the linguist-turned-political-activist ignores them.

Coreference chain: {Chomsky, He, linguist-turned-political-activist} {probabilities, them}
Coreference Problems

Peter gave Bob a piece of his mind.

He seemed to enjoy it.

It rained all day.
Coreference Algorithm 1: tree search for pronouns

Proposed antecedents: John
Coreference Algorithm 2: supervised ML

- identify all NPs
  \{Chomsky, probabilities, He, linguist-turned-political-activist, them\}

- pairwise match NPs and extract features
  features: gender, number, entity type, distance, …
  pairs: \(\text{(Chomsky, probabilities)}\)
  \(\text{(Chomsky, He)}\)
  \(\text{…}\)

- check for coreference w/ classifier

- sort into coreference chains
Coreference Overview

• performance: precision & recall of coref chains
• best models: supervised ML, ~70–90 F-score
• depends on data set, evaluation, etc.
Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Lionel Logue, a renowned speech therapist, was summoned to help the King overcome his speech impediment.

| Illinois Coreference Package | {Lionel Logue, a renowned speech therapist, his}  
|                            | {Queen Elizabeth}  
|                            | {the King, King}  
|                            | {King George VI, transforming her husband, her}  
|                            | {a viable monarch}  
| CherryPicker                | {Queen, her}  
|                            | {Elizabeth}  
|                            | {Husband}  
|                            | {King, George VI, King, his}  
|                            | {Lionel Logue}  
|                            | {therapist}  
| Reconcile                   | {Queen Elizabeth, her}  
|                            | {her husband, King George VI}  
|                            | {a Viable monarch}  
|                            | {Lionel Logue, a renowned speech therapist, the king, his}  
| ArkRef                      | {Queen Elizabeth, her}  
|                            | {her husband , King George VI}  
|                            | {Lionel Logue , a renowned speech therapist}  
|                            | {the King, his}  

false entity
wrong chain
Coherence

• meaningful combination of sentences
• means to “link” sentences:
  • coreference ✔
  • discourse relations
Discourse Relations

I will not buy this record because it is scratched

• what is their relation?
• how to name it?
• which sentence is independent?
• which order makes more sense?
• what if there are no markers?
Endless Possibilities

- no agreed-upon set of sentence-sentence relations:
  - McKeown: schemas (4)
  - Hobbs: inference-based (13)
  - Mann & Thomas: RST (23)
  - Survey by Hovy & Maier: 350+
RST

• Rhetorical Structure Theory (Mann & Thompson 1987)
• quasi-standard by now
• **nucleus**: sentence that can be interpreted on its own
• **satellite**: needs a nucleus to make sense

I will not buy this record  
It is scratched
• use RST to parse documents, similar to syntactic parsers
• Marcu 1997
• performance: P=47.0, R=78.4, F1=58.8
• difficulties: circular dependencies, graph instead of tree
With its distant orbit (-50 percent farther from the sun than Earth -), Mars experiences frigid weather conditions.

Surface temperatures typically average about \(-70\) degrees Fahrenheit at the equator and can dip to \(-123\) degrees C near the poles. Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion.

Marcu 2000
Parsing Discourse Algorithm

- determine cues in text
- find sections, paragraphs, sentences, clauses, etc.
- for all sections, paragraphs, sentences:
  - hypothesize rels b/w all units at that level based on markers and cohesion
  - use proof theory to infer trees below level
  - assign weights and find tree w/ max. weight
  - merge best trees at each level

Marcu 2000
Discourse Structure in Language Generation

- McKeown: use 4 schemas (Identification, Constituency, Attributive, Compare&Contrast)
- each schema: various sub-labels for each clause
- realize via ATNs (recursive FSAs)
Clause Functions

• = labels: {adversative, amplification, analogy, answer, attributive, cause-effect, comparison, constituency, contrast, evidence, explanation, generalization, identification, inference, particular-illustration, positing question, problem, renaming, representative, restriction}
Discourse Structure in Language Generation

- choose schema
- extract KB info
- walk ATN

Identification schema:
### Schema Example

<table>
<thead>
<tr>
<th>Schema</th>
<th>Functions</th>
<th>Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IDENTIFICATION</strong></td>
<td>identification</td>
<td>Eltville, (Germany): an important wine village of the Rheingau region.</td>
</tr>
<tr>
<td></td>
<td>evidence</td>
<td>The vineyards make wines that are emphatically of the Rheingau style,</td>
</tr>
<tr>
<td></td>
<td>amplification</td>
<td>with a considerable weight for a white wine.</td>
</tr>
<tr>
<td></td>
<td>particular</td>
<td>Taubenberg, Sonnenberg, and Langenstück are among vineyards of note.</td>
</tr>
<tr>
<td></td>
<td>illustration</td>
<td></td>
</tr>
</tbody>
</table>
Works Both Ways

- like FSA for grammar:
  - can recognize sentence as some schema S if it can apply functions and reach end state S
  - can generate sentences starting from S and following arcs in ATN (~Markov chain)
• A hobie is a brand of catamaran,
• which is a kind of sailboat.
• Large hobies have a jib, smaller ones don’t.
• Hobies have a cockpit connecting the pontoons.
• Catamarans have sail and mast like other sailboats,
• but they have two hulls instead of one.
• That’s a catamaran there!
Focus, Please!

- A hobie is a brand of catamaran, which is a kind of sailboat.
- Catamarans have sail and mast like other sailboats, but they have two hulls instead of one.
- That's a catamaran there!
- Hobies have a cockpit connecting the pontoons.
- Large hobies have a jib, smaller ones don't.
Pragmatics: In Theory

- using other sentences as context
- could help resolve ambiguities
- provides more information
- produces better text (summarization, MT)
Pragmatics: In (NLP) Practice

- across-sentence/document processing
- increases complexity
- introduces noise

⇒ still infrequently used by NLP systems
If You Remember Nothing Else…

- **pragmatics**: study of text structure
- **coherence**: sentence order sensibility measure
- **coreference**: refer to one entity with multiple expressions. Important but hard!
- **discourse relations**: rules & labels for sentence order. Parsing and generation approaches
Thanks for listening!