CSCI 5832
Natural Language Processing

Lecture 23
Jim Martin

Today: 4/17

- **Finish Lexical Semantics**
- **Wrap up Information Extraction**

Inside Words

- Thematic roles: more on the stuff that goes on inside verbs.

Inside Verbs

- Semantic generalizations over the specific roles that occur with specific verbs.
- I.e. Takers, givers, eaters, makers, doers, killers, all have something in common
  - -er
  - They're all the agents of the actions
- We can generalize (or try to) across other roles as well

Thematic Roles

Thematic Role	De nition
AGENT	The volitional causer of an event
EXPERIENCER	The experiencer of an event
FORCE	The non-volitional causer of the event
THEME	The participant most directly affected by an event
RESULT	The end product of an event
CONTENT	The proposition or content of a propositional event
INSTRUMENT	An instrument used in an event
BENEFICIARY	The bene ciary of an event
SOURCE	The origin of the object of a transfer event
GOAL	The destination of an object of a transfer event

Thematic Role Examples

Thematic Role	Example
AGENT	*The waiter* spilled the soup.
EXPERIENCER	*John* has a headache.
FORCE	*The wind* blows debris from the mall into our yards.
THEME	Only after Benjamin Franklin broke *the ice...*
RESULT	The French government has built a *regulation-size baseball diamond...*
CONTENT	Mona asked *"You met Mary Ann at a supermarket"*?
INSTRUMENT	He turned to poaching cat sh, stunning them *with a shocking device...*
BENEFICIARY	Whenever Ann Callahan makes hotel reservations *for her boss...*
SOURCE	I ew in *from Boston.*
GOAL	I drove *to Portland.*

Why Thematic Roles?

- It's not the case that every verb is unique and has to introduce unique labels for all of its roles; thematic roles let us specify a fixed set of roles.
- More importantly it permits us to distinguish surface level shallow semantics from deeper semantics

Example

- From the WSJ…
  - *He melted her reserve with a husky-voiced paean to her eyes.*
  - If we label the constituents He and reserve as the Melter and Melted, then those labels lose any meaning they might have had literally.
  - If we make them Agent and Theme then we don't have the same problems

Tasks

- **Shallow semantic analysis is defined as**
  - Assigning the right labels to the arguments of verb in a sentence. Aka
    - Case role assignment
    - Thematic role assignment

Thematic Role	De nition
AGENT	The volitional causer of an event
EXPERIENCER	The experiencer of an event
FORCE	The non-volitional causer of the event
THEME	The participant most directly affected by an event
RESULT	The end product of an event
CONTENT	The proposition or content of a propositional event
INSTRUMENT	An instrument used in an event
BENEFICIARY	The bene ciary of an event
SOURCE	The origin of the object of a transfer event
GOAL	The destination of an object of a transfer event

Example

- **Newswire text**
  - [*British forces* $_{agent}$] [*believe* $_{target}$] that [*Ali was killed in a recent air raid* $_{theme}$]
  - British forces believe that [*Ali* $_{theme}$] was [*killed* $_{target}$] [*in a recent air raid* $_{temporal}$]

Resources

- PropBank
  - Annotate every verb in the Penn Treebank with its semantic arguments.
  - Use a fixed (25 or so) set of role labels (Arg0, Arg1…)
  - Every verb has a set of frames associated with it that indicate what its roles are.
    - So for *Give* we're told that Arg0 -> Giver

Resources

- Propbank
  - Since it's built on the treebank we have the trees and the parts of speech for all the words in each sentence.
  - Since it's a corpus we have the statistical coverage information we need for training machine learning systems.

Resources

- **Propbank**
  - Since it's the WSJ it contains some fairly odd (domain specific) word uses that don't match our intuitions of the normal use of the words
  - Similarly, the word distribution is skewed by the genre from "normal" English (whatever that means).
  - There's no unifying semantic theory behind the various frame files (*buy* and *sell* are essentially unrelated).

Resources

- **FrameNet**
  - Instead of annotating a corpus, annotate domains of human knowledge a domain at a time (called frames)
    - Then within a domain annotate lexical items from within that domain.
    - Develop a set of semantic roles (called frame elements) that are based on the domain and shared across the lexical items in the frame.

Cause_Harm Frame

Cause_harm

Definition:

The words in this frame describe situations in which an Agent or a Cause injures a Victim. The Body_part of the Victim which is most directly affected may also be mentioned in the place of the Victim. In such cases, the Victim is often indicated as a genitive modifier of the Body_part, in which case the Victim FE is indicated on a second FE layer.

Lexical Units

Lexical Units

bash.v, batter.v, bayonet.v, beat up.v, beat.v, belt.v, biff.v, bludgeon.v, boil.v, break.v, bruise.v, buffet.v, burn.v, butt.v, cane.v, chop.v, claw.v, clout.v, club.v, crack.v, crush.v, cudgel.v, cuff.v, cut.v, elbow.v, electrocute.v, electrocution.n, flagellate.v, flog.v, fracture.v, gash.v, hammer.v, hit.v, horsewhip.v, hurt.v, impale.v, injure.v, jab.v, kick.v, knee.v, knife.v, knock.v, lash.v, maim.v, maul.v, mutilate.v, pelt.v, poison.v, poisoning.n, pummel.v, punch.v, slap.v, slice.v, smack.v, smash.v, spear.v, squash.v, stab.v, stone.v, strike.v, thwack.v, torture.v, transfix.v, welt.v, whip.v, wound.v

Created by hcb on Wed May 23 15:26:58 PDT 2001

FrameNet

- Frames and frame elements are entities in a hierarchy.
  - Cause_Harm inherits from Transitive_Action
  - Corporal_Punishment inherits from Cause_Harm
  - The victim FE in Cause_Harm inherits from the patient FE of Transitive_Action
  - And the evaluee of the Corporal_Punishment frame inherits from the victim of the Cause_Harm frame.

FrameNet

- Framenet.icsi.berkeley.edu

Break

Thursday we'll turn to discourse (Chapter 20).

Next week Stat MT

Final quiz will be on May 1.

HLT Certificate

You may be on your way to the…
Human Language Technology Certificate

For typical CS students
5 courses
CS: NLP, UI design, AI
Ling: Syntax and Morphology, Phonetics

Information Extraction

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York

Information Extraction

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

Named Entity Recognition

- **Find the named entities and classify them by type.**
- **Typical approach**
  - **Acquire training data**
  - **Encode using IOB labeling**
  - **Train a sequential supervised classifier**
  - **Augment with pre- and post-processing using available list resources (census data, gazeteers, etc.)**

Information Extraction

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York

Temporal and Numerical Expressions

- **Temporals**
  - **Find all the temporal expressions**
  - **Normalize them based on some reference point**
- **Numerical Expressions**
  - **Find all the expressions**
  - **Classify by type**
  - **Normalize**

Information Extraction

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York

Event Detection

- **Find and classify all the events in a text.**

Information Extraction

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York

Relation Extraction

- Basic task: find all the classifiable relations among the named entities in a text (populate a database)…
  - Employs
    - { <American, Tim Wagner> }
  - Part-Of
    - { <United, UAL>, {American, AMR} >

Relation Extraction

- Typical approach:

  For all pairs of entities in a text
  - Extract features from the text span that just covers both of the entities
    - Use a binary classifier to decide if there is likely to be a relation
    - If yes: then apply each of the known classifiers to the pair to decide which one it is
- Use supervised ML to train the required classifiers from an annotated corpus

Information Extraction

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York

Template Analysis

- **Many news stories have a script-like flavor to them. They have fixed sets of expected events, entities, relations, etc.**
- **Template, schemas or script processing involves:**
  - **Recognizing that a story matches a known script**
  - **Extracting the parts of that script**

Template Analysis

- So airlines often try to raise fares. Sometimes it sticks, sometimes it doesn't; it depends on how the other airlines react to the increase.
  - Airline that starts it off: United
  - Effective date of the increase: Thursday
  - Amount of the increase: $6
  - Followers: American
  - Routes: …

Template Processing

- Builds on earlier steps; obviously helps to know the entity types of the things that can fill the slots in the script.
- One approach…
  - Use supervised ML (with IOB labeling) to label all the candidate segments with their roles.
  - Collect all the candidate slots and resolve
    - If there's only one candidate take it
    - If not then vote or take the candidate with highest confidence score

Information Extraction

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York

Information Extraction Summary

- **Named entity recognition and classification**
- **Coreference analysis**
- **Temporal and numerical expression analysis**
- **Event detection and classification**
- **Relation extraction**
- **Template analysis**

Information Extraction

- Ordinary newswire text is often used in typical examples.
  - And there's an argument that there are useful applications there
- The real interest/money is in specialized domains
  - Bioinformatics
  - Patent analysis
  - Specific market segments for stock analysis
  - Intelligence analysis
  - Etc.