Data Intensive Linguistics — Lecture 13
Semantics and discourse
Philipp Koehn
20 February 2006

Semantics

- What is meaning?
- What is the meaning of the word cat?
  - not a specific cat
  - not all cats
  - abstract notion of any cat
- Atomic semantic units: concepts
  - example: cat → CAT

WordNet: an ontology of concepts

ENTITY
ANIMAL
MAMMAL
CARNIVORE
FELINE CANINE BEAR
CAT DOG WOLF FOX

Semantic relationships

- Hypernym / hyponym
  - CAT is-a FELINE
  - basis of hierarchical relationships in WordNet
- Part / whole
  - CAT has-part PAW
  - PAW is-part of CAT
- Membership
  - FACULTY has-member PROFESSOR
  - PROFESSOR is-member-of FACULTY
- Antonym / opposite
  - LEADER is-opposite-of FOLLOWER

Thematic roles

- Words play semantic roles in a sentence

  see the woman with the telescope

- Specific verbs typically require arguments with specific thematic roles and allow adjuncts with specific thematic roles.

Semantic frames

- Complex concepts can be defined by semantic frames, whose slots are filled by concrete information

  SOCCER-GAME
  - HOME-TEAM: Heart of Midlothian
  - AWAY-TEAM: FC Motherwell
  - SCORE: 3-0
  - TIME-STARTED: 2006-02-18 1600 GMT
  - LOCATION: Tynecastle Stadium, Edinburgh

- Information extraction: can we fill semantic frames from text?

Source of semantic knowledge

- Semantic knowledge is not directly observable
- Building semantic knowledge bases
  - for instance WordNet, an ontology
  - labor intensive
  - may not contain all information we want, e.g.
    - pigeon is a typical bird
    - penguin is not a typical bird
- Can we automatically learn semantics?

Learning semantics

The meaning of a word is its use.
Ludwig Wittgenstein, Aphorism 43

- Represent context of a word in a vector
  - Similar words have similar context vectors
- Example: Google sets http://labs.google.com/sets
  - one meaning of cat
    - enter: cat, dog
    - return: cat, dog, horse, fish, bird, rabbit, cattle, ...
  - another meaning of cat
    - enter: cat, more
    - return: more, cat, is, rm, mv, cd, cp, ...
Learning prejudices

- Detecting national stereotypes with Google
  - Enter: Scots are known to be *
    - fragal, friendly, generous, thrifty, ...
  - Enter: Englishmen are known to be *
    - prudish, great sports-lovers, people with manners, courteous, cold, ...
  - Enter: Germans are known to be *
    - pathetic, hard-nosed, arrogant, very punctual, fanatical, hard-working, ...

Text segmentation

- Some text types have very pronounced topic shifts
  - news broadcasts cover different stories
- Also other long texts may cover multiple topics
  - lectures
  - speeches
  - essays
- Task text segmentation
  - given: text
  - wanted: segmentation into smaller units with different topics

Segmentation by vocabulary change

- At a topic boundary, use of vocabulary changes
  - By comparing vocabulary of neighboring text parts, boundaries can be detected
- Example: Stargazers text from Hearst [1994]
  - intro: the search for life in space
  - the moons chemical composition
  - how early proximity of the moon shaped it
  - how the moon helped life evolve on earth
  - improbability of the earth-moon system

Rhetorical relations

- Rhetorical Structure Theory (RST): relations between spans of EDUs
  - Example:

  ![Rhetorical Structure Theory Example](image)

  \[ \text{The bank also says} \]
  \[ \text{it will use its network to channel investments} \]

Types of rhetorical relations

- Mono-nuclear: Nucleus is more salient than satellite, which contains supporting information
- Multi-nuclear: joining spans have equal importance

- 78 types of relations in 16 classes
  - attribution, background, cause, comparison, condition, contrast, elaboration, enablement, evaluation, explanation, joint, manner-means, topic-comment, summary, temporal, topic-change

- More detail, see: Building a discourse-tagged corpus in the framework of rhetorical structure theory by Lynn Carlson, Daniel Marcu, and Mary Ellen Kurowski [SIGDIAL 2001]

Discourse parsing

- Human annotator agreement on rhetorical relations is not very high
  - 71.9% if 18 relation types are used
  - 77.0% if 130 relation types are used
- Probabilistic parsing model [Soricut and Marcu, NAACL 2003]
  - probabilistic chart parser
  - achieves similar performance
- Experiments done on the sentence level.
  - Discourse parsing should be useful for, e.g., summarization
Anaphora

Violent protests broke out again in Happyland. According to the country’s department of peace, flowers will be handed out tomorrow. A spokesman of the department announced that they will be blue and green. This will demonstrate the country’s commitment to alleviate the situation.

- A text contains often multiple references to the same objects:
  - flowers — they
  - Happyland — the country
  - department of peace — the department
  - violent protests — the situation
  - handing out flowers — this

- Anaphora resolution (matching the references) is a hard problem

Sentiment detection

- What is the overall sentiment of a text
- Example: movie review
  - is it a recommendation or a negative review?
  - can be framed as a text classification problem
  - see: “Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales” by Bo Pang and Lillian Lee [ACL 2005]

- Similar questions
  - is a text critical of a person?
  - does the text have a bias (political, etc.)?