Computational Linguistics 1

CMSC/LING 723, LBSC 744

Kristy Hollingshead Seitz
Institute for Advanced Computer Studies
University of Maryland

Lecture 26: 8 December 2011

Agenda

- HW7 due next Tuesday
  - Please provide a brief writeup of what you changed, and why you thought it would help. Include some examples where your changes improved the results, where it didn't, and report on the overall accuracy change as appropriate.
  - Also submit your code
- Review for the final next Tuesday
- Questions, comments, concerns?
- Summarization
- Information Extraction (IE)
  - Co-reference Resolution

Summarization

- Some key dimensions within which summarization systems differ
  - Extracting versus abstracting
  - Single document versus multi-document
  - Query driven versus general summarization
- Summarization systems these days are most typically
  - Extractive: abstracting is very difficult
    - Rarely, might see systems that 'fuse' sentences
  - Multi-document: to exploit redundancy
    - Introduces problems of co-reference, conflicting info
  - Query driven: summarization doesn't occur in a vacuum

Extractive Summarization

- Summary composed of extracted parts of documents
  - Common granularity at sentence level, but not required
  - In fact, get close to abstracting if unit is smaller
- Must first segment the document into extractable units
  - Naive sentence segmentation: when you see a period, segment
- Then rank the extractable units
  - Naive sentence ranking: position in document (earlier better)
- Then extract some number of the ranked extractable units
  - Naive sentence selection: pick top n units

Sentence Selection

- Sentences can be similar across multiple documents
- The top of the list may include very similar sentences
- If we have limited space (say 250 words), repetition is not good
- Want to respect the ranking, yet penalize redundancy somehow
- Integrating rank with penalty can be tricky

Comparison to IR

- Both involve ranking according to some score; and selection
- Ranking criterion will differ . . .
  - For IR, terms are weighted based on how they distinguish document(s) from others in the collection
  - For summarization, want sentences that are "central" to the document(s)
- . . . but related: still care about term frequency, inverse document frequency and stop lists
- Pretty different use of vector space, however

Sentence-Ranking Possibilities

- Suppose we use TF*IDF term weighting
- Create an n-dimensional normalized vector for document: d
- Create an n-dimensional normalized vector for each sentence: s
- Rank sentences by:

$$\cos(d, s) = \sum_{i=1}^{n} d[i]s[i]$$

- Maybe other useful features?:
  - Position of sentence in the document
  - Distribution of terms across documents in the set
  - Query terms (maybe by influencing term weighting?)

Query Expansion

- Relevant to both IR and summarization
- Given a query, expand to include related terms
- Can solve some issues from text normalization
  - query term: 'ncx1'; expand with: 'NCX1'
  - query term: 'regulation'; expand with 'regulating', 'regulate'
- Can also help by including semantically related words
  - query term: 'cow'; expand with 'bovine'
- Possible problem: swamping original terms with expanded terms

Summarization Evaluation

- This is a big, difficult topic: what makes a good summary
- Here we will focus on automatic evaluation given references
- Basic intuition, comparing two summaries, S_1 and S_2
  - To the extent that S_1 overlaps more with a reference summary R, it is "better" than S_2
- Key questions
  - How does one measure overlap?
  - What about summary length?
  - What if there are multiple reference summaries?

ROUGE

- Measure overlap by counting matching n-grams
  - n-gram is a word sequence of length n
  - e.g., unigram 'dog'; bigram 'dog food'; trigram 'dog food can'
- Let $c(x, s)$ be the count of n-gram x in summary s
- Let $c(x, r)$ be the count of n-gram x in reference r
- For a given n, the ROUGE score is

$$\text{ROUGE-}n = \frac{\sum_x \min\{c(x,r), c(x,s)\}}{\sum_x c(x,r)}$$

- Denominator reference count, thus a recall metric

Further topics of exploration

- Moving beyond TF*IDF term weighting
  - How can we find "central" sentences in new ways?
  - Graph based random walk methods
- Including multiple factors in sentence ranking
- Using the query to influence the term weighting
- Query expansion
- Is there a principled way to reduce redundancy?

Document Understanding Conference (DUC)

- Last few years focused on query-driven multi-doc summarization
- Has transitioned to "Text Analysis Conference" (TAC)
- New tasks of interest: update summaries; opinion summaries
- Focus on multiple evaluation metrics
  - Some manual, e.g., pyramid analysis
  - Some automatic, e.g., ROUGE, BE
- Lots of new ideas tried every year, mainly extractive techniques
- Data created through bakeoffs can be used for training systems

Agenda

- HW7 due next Tuesday
- Review for the final next Tuesday
- Questions, comments, concerns?
- Summarization
- Information Extraction (IE)
  - Co-reference Resolution

Information Extraction

- IR versus IE
  - IR retrieves relevant documents from collections
    - Looking for documents or larger passages
    - Using information theory, probabilistic theory, statistics
  - IE extracts relevant information from documents
    - Looking for structure
    - Using natural language processing

IE Tasks

- Most IE tasks involve:
  1. Document segmentation
  2. Labeling of segments
  3. Discovery of relations between labeled segments
- Examples:
  - Form filling from classified advertisements
  - Named entity recognition
  - Discovery of part/whole relations

Example IE Task

From the 7th Message Understanding Conference (MUC-7)

- Find the description of a launch event, and fill in:
  - Vehicle
  - Payload
  - Mission Date
  - Mission Site
  - Mission Type (Military, Civilian)
  - Mission Function (Test, Deploy, Retrieve)
  - Mission Status (Succeeded, Failed, In Progress, Scheduled)

Other MUC tasks:

- Latin American terrorism, Joint ventures, Microelectronics, Company management changes

IE as Tagging Tasks

- Most segments in typical IE tasks are non-hierarchical and non-overlapping
- Thus can be modeled using simple finite-state models
  - For each label X, a word can begin (B-X) or be inside (I-X)
  - Some words may be outside (O) any labeled segment
- Ad hoc IR treats document as unordered set. . .
- . . . but IE must take into account sequence information
  - Previous word's tag may influence that of the current word
  - Typically include information about word sequence

Named Entity Recognition (NER): an IE Task

Begin segment words underlined and in red. Inside segment words just in red. Everything else outside segment.

```
Exchange activity in cardiomyocytes is regulated
by several factors. It is activated by cytosolic
Ca2+ and MgATP (20) and inhibited by cytosolic
sodium (21) and ATP depletion (22). A high
affinity Ca2+-binding domain has been identified
in the large cytoplasmic loop (residues 371-508)
that is believed to be responsible for calcium
regulation (23). It is also inhibited by the
exchanger inhibitory peptide, XIP, that
corresponds to a 20-amino acid segment at the N
terminus of the large cytoplasmic loop (24).
```

Co-reference Resolution: an IE Task

- Introduced as a task at MUC-6
- Recognize referential relations among expressions
  - Whole-part relations
  - Set-subset
  - Type-token
- Recognize identify of reference among (similar) noun phrases

Co-reference Examples

- Names and aliases
  - International Business Machines, IBM, Big Blue
  - Mr. William H. Gates, Mr. Gates, Bill Gates
- Definite noun phrases
  - the big computer company, the Armonk-based giant
  - the head honcho at Microsoft, the world's richest man
- Pronouns
  - he, her, his, it, its, we, they, theirs, them, ours, your, ...

Another Co-reference Example

John went to Bill's car dealership to check out an Acura Integra. He looked at it for about an hour.

- Possible interpretations:
  - John looked at Bill's car dealership
  - Bill looked at John
  - John looked at an Acura Integra
  - John looked at Bill etc.
- Not just pronouns:

  John went to Bill's car dealership to check out an Acura Integra. The car was just what he wanted.

Co-reference Resolution

- Two referring expressions that refer to the same referent are said to **co-refer**
- A **referring expression** is a natural language expression referring to an entity called the **referent**, e.g. the word *Shaq* and that guy dunking over there
- A referring expression *licensing* the use of another is known as the **antecedent**
  - e.g. John went to ..., and *he* ...
- Pronouns can be **bound** by quantifiers
  - Every boy drinks milk with his lunch

Types of Referring Expressions

- *indefinite NPs*: used to introduce new entities to discourse
  - Waiter, there's *a fly* in my soup.
- *definite NPs*: used to refer to an identifiable entity
  - Yes, sir, *the fly* is doing *the backstroke*.
  - *The Trailblazers* are in rebuilding mode.
- *Pronouns* are a form of definite reference, usually to highly **salient** referents
  - *Reflexive* pronouns, e.g. *himself*, *herself*
  - *Demonstrative* pronouns, e.g. *this* or *that*
- *One* anaphora, has properties of both indefinite and definite reference
  - Everybody has cell phones these days, but John doesn't want *one*.

Other Kinds of Reference

- Inferable reference: referent not explicitly mentioned, but inferable
  - We climbed Mt. Shasta, but *the summit* was somewhat disappointing.
  - Part/whole relations, and the results of processes:
    After raking the yard, put *the leaves* in a bag and put it at the curb
- Discontinuous sets: *pairwise* reading
  - John gets letters from his aunt, and Jimmy gets letters from his cousin. They love getting them.
- Generic reference: general classes of previously mentioned entities
  - I had a pickle with lunch. It's my favorite vegetable.
- These make an already difficult problem harder

Co-reference Constraints

- Agreement: Number, Person, Case, Gender. Violations:
  - Number: Mary met Mel Gibson. They didn't like him.
  - Person: Mary met Mel Gibson. You didn't like him.
  - Case: Mary met Mel Gibson. She didn't like he.
  - Gender: Mary met Mel Gibson. She didn't like her.
- Syntactic constraints
  - John likes himself. (himself=John)
  - John likes him. (him ≠ John)
  - Bill said that John likes himself. (himself=John)
  - Bill said that he likes himself. (himself=he)

NP Co-Reference

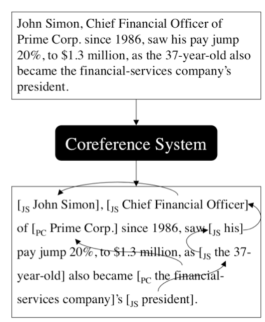
- Much literature on pronoun co-reference resolution
- Non-pronoun NP co-reference is also a fairly difficult problem
- Content in NPs that can provide clues to co-reference
  - Overlap in words between expressions, e.g. “Shaquille O'Neal” and “Shaq”
  - Inferable features, such as gender and animacy
  - Definiteness

The Role of Co-reference in IE

- Objects involved in relevant events and relationships are referred to in many different ways and often at widely separate locations in a text

```
Motor Vehicles International Corp. announced a major
management shakeup. MVI said its CEO had resigned.
The big automaker is attempting to regain market
share. It will announce significant losses for the
3rd quarter. A company spokesman said the company
will be moving their operations...MVI, the first
automaker to announce quarterly results, is the
biggest American auto exporter to Latin America.
```


Co-reference System [from Cardie and Wagstaff (1999)]



Syntactic Analysis in IE

- Generally directed toward shallow, simple parses of core constituents
- Semantic analysis involves only ground propositions
- Prepositional attachment only for arguments of domain-relevant verbs
- Locative and temporal adverbials processed, others ignored

IE Evaluation Metrics

- Precision (P)

$$Precision = \frac{\#\ \text{correct answers}}{\#\ \text{answers produced}}$$

- Recall (R)

$$Recall = \frac{\#\ \text{correct answers}}{\#\ \text{total possible corrects}}$$

- F-measure

$$F = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

Agenda

- HW7 due next Tuesday
- Review for the final next Tuesday
- Questions, comments, concerns?
- Summarization
- Information Extraction (IE)
  - Co-reference Resolution