Entity Coreference Resolution

CMSC 473/673
UMBC
December 6th, 2017
Course Announcement 1: Assignment 4

Due Monday December 11th (~5 days)

Remaining late days can be used until 12/20, 11:59 AM

Any questions?
Course Announcement 2: Project

Due Wednesday 12/20, 11:59 AM

Late days cannot be used

Any questions?
Course Announcement 3: Final Exam

No mandatory final exam

December 20th, 1pm-3pm: optional second midterm/final

Averaged into first midterm score

No practice questions

Register by Monday 12/11:

https://goo.gl/forms/aXflKkP0BlRxhOS83
Course Announcement 4: Evaluations

Please fill them out! (We do pay attention to them)

Links from StudentCourseEvaluations@umbc.edu
Recap from last time...
Entailment: Underlying a Number of Applications

Question
Who bought Overture?

>>

Expected answer form
X bought Overture

Overture’s acquisition by Yahoo

entails

Yahoo bought Overture

Information extraction: X acquire Y
Information retrieval: Overture was bought for ...
Summarization: identify redundant information
MT evaluation

Adapted from Dagan, Roth and Zanzotto (2007; tutorial)
Applied Textual Entailment

A directional relation between two text fragments

\[ t \text{ (text) entails } h \text{ (hypothesis) } (t \rightarrow h) \text{ if} \]

humans reading \( t \) will infer that \( h \) is most likely true

\[ t \text{ probabilistically entails } h \text{ if:} \]

\[ P(h \text{ is true} | t) > P(h \text{ is true}) \]

t increases the likelihood of \( h \) being true

Positive PMI – \( t \) provides information on \( h \)’s truth

the value is the entailment confidence

Adapted from Dagan, Roth and Zanzotto (2007; tutorial)
<table>
<thead>
<tr>
<th>TEXT</th>
<th>HYPOTHESIS</th>
<th>TASK</th>
<th>ENTAILMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reagan attended a ceremony in Washington to commemorate the landings in Normandy.</td>
<td>Washington is located in Normandy.</td>
<td>IE</td>
<td>False</td>
</tr>
<tr>
<td>Google files for its long awaited IPO.</td>
<td>Google goes public.</td>
<td>IR</td>
<td>True</td>
</tr>
<tr>
<td>...: a shootout at the Guadalajara airport in May, 1993, that killed Cardinal Juan Jesus Posadas Ocampo and six others.</td>
<td>Cardinal Juan Jesus Posadas Ocampo died in 1993.</td>
<td>QA</td>
<td>True</td>
</tr>
<tr>
<td>The SPD got just 21.5% of the vote in the European Parliament elections, while the conservative opposition parties polled 44.5%.</td>
<td>The SPD is defeated by the opposition parties.</td>
<td>IE</td>
<td>True</td>
</tr>
</tbody>
</table>

Adapted from Dagan, Roth and Zanzotto (2007; tutorial)
Basic Representations

Meaning Representation

Inference

Logical Forms

Semantic Representation

Syntactic Parse

Local Lexical

Raw Text

Textual Entailment

Adapted from Dagan, Roth and Zanzotto (2007; tutorial)
Common and Successful Approaches (Features)

Measure similarity match between $t$ and $h$

Lexical overlap (unigram, N-gram, subsequence)
Lexical substitution (WordNet, statistical)
Syntactic matching/transformations
Lexical-syntactic variations (“paraphrases”)
Semantic role labeling and matching
Global similarity parameters (e.g. negation, modality)

Cross-pair similarity
Detect mismatch (for non-entailment)
Interpretation to logic representation
+ logic inference

Lexical baselines are hard to beat!

Lack of knowledge (syntactic transformation rules, paraphrases, lexical relations, etc.)

Lack of training data

Adapted from Dagan, Roth and Zanzotto (2007; tutorial)
Knowledge Acquisition

Direct Algorithms
- Concepts from text via clustering (Lin and Pantel, 2001)
- Inference rules – aka DIRT (Lin and Pantel, 2001)
...

Indirect Algorithms
- Hearst’s ISA patterns (Hearst, 1992)
- Question Answering patterns (Ravichandran and Hovy, 2002)
...

Iterative Algorithms
- Entailment rules from Web (Szepktor et al., 2004)
- Espresso (Pantel and Pennacchiotti, 2006)
...

Adapted from Dagan, Roth and Zanzotto (2007; tutorial)
Choice of Plausible Alternatives (COPA; Roemmele et al., 2011)

Goal: test causal implication, not (likely) entailment

1000 questions
Premise, prompt, and 2 plausible alternatives
Forced choice, 50% random baseline
Forward and backward causality
Cohen’s Kappa = 0.95 (only 30 disagreements)

Forward causal reasoning:
The chef hit the egg on the side of the bowl. What happened as a RESULT?
A. The egg cracked.
B. The egg rotted.

Backward causal reasoning:
The man broke his toe. What was the CAUSE of this?
A. He got a hole in his sock.
B. He dropped a hammer on his foot.

http://ict.usc.edu/~gordon/copa.html

Adapted from Roemmele et al. (2011)
## COPA Test Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMI (window of 5)</td>
<td>58.8</td>
</tr>
<tr>
<td>PMI (window of 25)</td>
<td>58.6</td>
</tr>
<tr>
<td>PMI (window of 50)</td>
<td>55.6</td>
</tr>
<tr>
<td>Goodwin et al.: bigram PMI</td>
<td>61.8</td>
</tr>
<tr>
<td>Goodwin et al.: SVM</td>
<td>63.4</td>
</tr>
</tbody>
</table>

Performance of purely associative statistical NLP techniques?

Statements that are causally related often occur close together in text. Connected by causal expressions (“because”, “as a result”, “so”)

Approach: choose the alternative with a stronger correlation to the premise

PMI a la Church and Hanks, 1989

Adapted from Roemmele et al. (2011)
SNLI (Bowman et al., 2015)

<table>
<thead>
<tr>
<th>Bowman et al. (2015)</th>
<th>SNLI Test Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexicalized</td>
<td>78.2</td>
</tr>
<tr>
<td>Unigrams Only</td>
<td>71.6</td>
</tr>
<tr>
<td>Unlexicalized</td>
<td>50.4</td>
</tr>
<tr>
<td>Neural: sum of word vectors</td>
<td>75.3</td>
</tr>
<tr>
<td>Neural: LSTM</td>
<td>77.6</td>
</tr>
</tbody>
</table>

BLEU score between hypothesis and premise

# words in hypothesis - # words in premise

word overlap

unigram and bigrams in the hypothesis

Cross-unigrams: for every pair of words across the premise and hypothesis which share a POS tag, an indicator feature over the two words.

Cross-bigrams: for every pair of bigrams across the premise and hypothesis which share a POS tag on the second word, an indicator feature over the two bigrams

![Diagram of neural network model]
Pat and Chandler agreed on a plan.

He said Pat would try the same tactic again.
Pat and Chandler agreed on a plan.

He said Pat would try the same tactic again.
Pat and Chandler agreed on a plan.

He said Pat would try the same tactic again.

is “he” the same person as “Chandler?”
Coref Applications

- Question answering
- Information extraction
- Machine translation
- Text summarization
- Information retrieval
Winograd Schemas

I spread the cloth on the table to protect it.

I spread the cloth on the table to display it.
Rule-Based Attempts

Popular in the 1970s and 1980s
   Charniak (1972): Children’s story comprehension
       “In order to do pronoun resolution, one had to be able to do everything else.”

Focus on sophisticated knowledge & inference mechanisms

Syntax-based approaches (Hobbs, 1976)

Discourse-based approaches / Centering algorithms
   Kantor (1977), Grosz (1977), Webber (1978), Sidner (1979)
Basic System
Basic System

Input Text → Preprocessing
Basic System

Input Text → Preprocessing → Mention Detection
Basic System

Input Text → Preprocessing → Mention Detection → Coref Model
Basic System

Input Text → Preprocessing → Mention Detection → Coref Model → Output
Basic System

Input Text → Preprocessing → Mention Detection → Coref Model → Output
Preprocessing

POS tagging
Stemming
Predicate argument representation
  verb predicates and nominalization
Entity Annotation
  Stand alone NERs with a variable number of classes
Dates, times and numeric value normalization
Identification of semantic relations
  complex nominals, genitives, adjectival phrases, and adjectival clauses
Event identification
Shallow Parsing (chunking)
Preprocessing

POS tagging
Stemming
Predicate argument representation
  verb predicates and nominalization

**Entity Annotation**
  Stand alone NERs with a variable number of classes
Dates, times and numeric value normalization
Identification of semantic relations
  complex nominals, genitives, adjectival phrases, and adjectival clauses
Event identification

**Shallow Parsing (chunking)**
Basic System

Input Text → Preprocessing → Mention Detection → Coref Model → Output
What are Named Entities?

Named entity recognition (NER)

Identify proper names in texts, and classification into a set of predefined categories of interest

- Person names
- Organizations (companies, government organisations, committees, etc)
- Locations (cities, countries, rivers, etc)
- Date and time expressions

Cunningham and Bontcheva (2003, RANLP Tutorial)
What are Named Entities?

Named entity recognition (NER)

Identify proper names in texts, and classification into a set of predefined categories of interest

- Person names
- Organizations (companies, government organisations, committees, etc)
- Locations (cities, countries, rivers, etc)
- Date and time expressions
- Measures (percent, money, weight etc), email addresses, Web addresses, street addresses, etc.
- Domain-specific: names of drugs, medical conditions, names of ships, bibliographic references etc.

Cunningham and Bontcheva (2003, RANLP Tutorial)
Basic Problems in NE

Variation of NEs: John Smith, Mr Smith, John.

Ambiguity of NE types: John Smith (company vs. person)
- May (person vs. month)
- Washington (person vs. location)
- 1945 (date vs. time)

Ambiguity with common words, e.g. "may"

Cunningham and Bontcheva (2003, RANLP Tutorial)
More complex problems in NE

Issues of style, structure, domain, genre etc.

Punctuation, spelling, spacing, formatting

Dept. of Computing and Maths
Manchester Metropolitan University
Manchester
United Kingdom
Two kinds of NE approaches

Knowledge Engineering

- rule based
- developed by experienced language engineers
- make use of human intuition
- requires only small amount of training data
- development could be very time consuming
- some changes may be hard to accommodate

Learning Systems

- requires some (large?) amounts of annotated training data
- some changes may require re-annotation of the entire training corpus
- annotators can be cheap
Baseline: list lookup approach

System that recognises only entities stored in its lists (gazetteers).

Advantages - Simple, fast, language independent, easy to retarget (just create lists)

Disadvantages – impossible to enumerate all names, collection and maintenance of lists, cannot deal with name variants, cannot resolve ambiguity
Creating Gazetteer Lists

Online phone directories and yellow pages for person and organisation names; SSA database

https://www.ssa.gov/oact/babynames/

Locations lists

US GEOnet Names Server (GNS) data – 3.9 million locations with 5.37 million names (e.g., [Manov03])
UN site: http://unstats.un.org/unsd/citydata
Global Discovery database from Europa technologies Ltd, UK (e.g., [Ignat03])

Automatic collection from annotated training data

Cunningham and Bontcheva (2003, RANLP Tutorial)
Shallow Parsing Approach
(internal structure)

Internal evidence – names often have internal structure. These components can be either stored or guessed, e.g. location:

Cap. Word + \{City, Forest, Center, River\}
Sherwood Forest

Cap. Word + \{Street, Boulevard, Avenue, Crescent, Road\}
Portobello Street

Cunningham and Bontcheva (2003, RANLP Tutorial)
Problems with the shallow parsing approach

Ambiguously capitalized words (first word in sentence)
[All American Bank] vs. All [State Police]

Semantic ambiguity
"John F. Kennedy" = airport (location)
"Philip Morris" = organisation

Structural ambiguity
[Cable and Wireless] vs. [Microsoft] and [Dell];
[Center for Computational Linguistics] vs. message from [City Hospital] for [John Smith]

Cunningham and Bontcheva (2003, RANLP Tutorial)
Gazetteer lists for rule-based NE

Needed to store the indicator strings for the internal structure and context rules

Internal location indicators: river, mountain, street, road

Internal organisation indicators: Ltd., Inc.

Cunningham and Bontcheva (2003, RANLP Tutorial)
Named Entity Grammars

Phases run sequentially and constitute a cascade of FSTs over the pre-processing results

Hand-coded rules applied to annotations to identify NEs

Annotations from format analysis, tokeniser, sentence splitter, POS tagger, and gazetteer modules

Use of contextual information

Finds person names, locations, organisations, dates, addresses.

Cunningham and Bontcheva (2003, RANLP Tutorial)
NER and Shallow Parsing: Machine Learning
Sequence models (HMM, CRF) often effective
BIO encoding
NER and Shallow Parsing: Machine Learning

Sequence models (HMM, CRF) often effective

BIO encoding

Pat and Chandler Smith agreed on a plan.
NER and Shallow Parsing: Machine Learning
Sequence models (HMM, CRF) often effective

BIO encoding

Pat and Chandler Smith agreed on a plan.
NER and Shallow Parsing: Machine Learning

Sequence models (HMM, CRF) often effective

BIO encoding

Pat and Chandler Smith agreed on a plan.

B-NP  O  B-NP  I-NP  B-VP  O  B-NP  I-NP
NER and Shallow Parsing: Machine Learning

Sequence models (HMM, CRF) often effective

BIO encoding

Pat and Chandler Smith agreed on a plan.

B-NP  O  B-NP  I-NP  B-VP  O  B-NP  I-NP
Basic System

Input Text → Preprocessing → Mention Detection → Coref Model → Output
Pat and Chandler agreed on a plan.

He said Pat would try the same tactic again.
Pat and Chandler agreed on a plan.

He said Pat would try the same tactic again.
Pat and Chandler agreed on a plan. He said Pat would try the same tactic again.

Model Attempt 1: Binary Classification

naïve approach (take all non-positive pairs): highly imbalanced!

Soon et al. (2001): heuristic for more balanced selection
Pat and Chandler agreed on a plan.

He said Pat would try the same tactic again.

possible problem: not transitive
Model Attempt 1: Binary Classification

Pat and Chandler agreed on a plan.

He said Pat would try the same tactic again.

solution: go left-to-right

possible problem: not transitive

for a mention $m$, select the closest preceding coreferent mention

otherwise, no antecedent is found for $m$
Anaphora

does a mention have an antecedent?
Anaphora

does a mention have an antecedent?

Chris told Pat he aced the test.
Pat and Chandler agreed on a plan.

He said Pat would try the same tactic again.
Pat and Chandler agreed on a plan.

He said Pat would try the same tactic again.
Model 2: Entity-Mention Model

Pat and Chandler agreed on a plan.

He said Pat would try the same tactic again.

**advantage:** featurize based on all (or some or none) of the clustered mentions
Model 2: Entity-Mention Model

Pat and Chandler agreed on a plan.

He said Pat would try the same tactic again.

**advantage**: featurize based on *all* (or some or none) of the *clustered mentions*

**disadvantage**: clustering doesn’t address anaphora
Model 3: Cluster-Ranking Model 
(Rahman and Ng, 2009)

Pat and Chandler agreed on a plan.

He said Pat would try the same tactic again.

learn to rank the clusters and items in them
Stanford Coref (Lee et al., 2011)
Stanford Coref (Lee et al., 2011)

John is a musician. He played a new song. A girl was listening to the song. “It is my favorite,” John said to her.
[John] is [a musician]. [He] played [a new song]. [A girl] was listening to [the song]. “[It] is [[my] favorite],” [John] said to [her].
[John] is [a musician]. [He] played [a new song]. [A girl] was listening to [the song]. "[It] is [[my] favorite]," [John] said to [her].
John is a musician. He played a new song. A girl was listening to the song. "It is my favorite," John said to her.
[John] is [a musician]. [He] played [a new song]. [A girl] was listening to [the song]. “[It] is [[my] favorite],” [John] said to [her].
[John] is [a musician]. [He] played [a new song]. [A girl] was listening to [the song]. “[It] is [[my] favorite],” [John] said to [her].
[John] is [a musician]. [He] played [a new song]. [A girl] was listening to [the song]. “[It] is [[my] favorite],” [John] said to [her].
[John] is [a musician]. [He] played [a new song]. [A girl] was listening to [the song]. "[It] is [[[my] favorite]]," [John] said to [her].
[John] is [a musician]. [He] played [a new song]. [A girl] was listening to [the song]. “[It] is [[my] favorite],” [John] said to [her].

Stanford Coref (Lee et al., 2011)
[John] is [a musician]. [He] played [a new song]. [A girl] was listening to [the song]. “[It] is [[my] favorite],” [John] said to [her].
Other Sieve-Based Approaches

Ratinov & Roth (EMNLP 2012)

Each sieve is a machine-learned classifier

Later sieves can override earlier sieves’ decisions

Can recover from errors as additional evidence is available
Other Sieve-Based Approaches

Ratinov & Roth (EMNLP 2012)

Each sieve is a machine-learned classifier

Later sieves can override earlier sieves’ decisions

Can recover from errors as additional evidence is available

• Nested (e.g., {city of {Jerusalem}})
• Same Sentence both Named Entities (NEs)
• Adjacent (Mentions closest to each other in dependency tree)
• Same Sentence NE&Nominal (e.g., Barack Obama, president)
• Different Sentence two NEs
• Same Sentence No Pronouns
• Different Sentence Closest Mentions (no intervening mentions)
• Same Sentence All Pairs
• All Pairs

Ng (2006; IJCAI Tutorial)
Possible Classifiers

Perceptron → Structured Perceptron

Learn more in 678
Possible Classifiers

Perceptron → Structured Perceptron

Structured Maxent Classifiers

Neural Networks

Learn more in 678