Natural Language Processing for Knowledge Representation and Reasoning

Michaël Thomazo

May 18th, 2014
Dresden

Recap

Previous courses:

- motivation for mixing NLP and KR
- Markov Logic
- inference algorithm

Today: presentation of the coreference resolution problem and of some algorithms.

Mentions

It is the process in which one identifies the mentions that are referring to the same real-world entity.

- name mention: Joe Smith
- nominal mention: the guy wearing a blue shirt
- pronoun mentions: he, him

Finding Mentions

The historian who taught herself Cobol

Three mentions:

- the whole phrase
- who
- herself

Mention head

Let us consider the following mention:

the new glass-clad skyscraper

Mention head

Let us consider the following mention:

the new glass-clad skyscraper

The head is *skyscraper*

Finding mentions and mention heads

This is already a non-trivial task. A tool often used for that purpose is described in:
Dan Klein, Christopher D. Manning: Accurate Unlexicalized Parsing. ACL 2003: 423-430

Entities (1)

Source: ACE (Automatic Content Extraction) English Annotation Guidelines for Entities

Entities are limited to the following seven types:

- Person (PER). Limited to humans. A person may be a single individual or a group.
- Organization (ORG). Organization entities are limited to corporations, agencies, or other groups of people defined by an established organizational structure.
- Geo-political Entity (GPE). GPE entities are geographical regions defined by political and/or social groups. A GPE entity subsumes and does not distinguish between a nation, its region, its government, or its people.

Entities (2)

Source: ACE (Automatic Content Extraction) English Annotation Guidelines for Entities

- Location (LOC) - Location entities are limited to geographical entities such as geographical areas and landmasses, bodies of water, and geological formations.
- Facility (FAC) - Facility entities are limited to buildings and other permanent man-made structures and real estates improvements
- Vehicle (VEH)
- Weapon (WEA)

First approach

Co-reference resolution seen as a classification problem: given a pair of mention (m_1, m_2), decide if m_1 and m_2 are co-referent. For clustering mentions in a text, the following is done for each mention m_i:

- look at mention $m_{i-1}, \ldots, m_1$ until some m_j is considered co-referent with m_i;
- create a link between m_i and m_j.

An example of output

See the blackboard.

Training set creation

Creation of both positive and negative examples.
If (m_j, m_i) are such that m_i is the first mention coreferent to m_j on its left, then:

- create a positive example for (m_j, m_i)
- create a negative example for each (m_j, m_k) with k between i and j.

Evaluation Metrics

- pairwise F-score
- MUC (based on links)
- B^3 (based on mentions)
- CEAF (Constraint Entity-Alignment F-Measure)

Features (1)

Soon et al. 2001

- SOON_STR: C if after discarding determines, both strings match, I otherwise
- PRONOUN_1: Y if first is a pronoun, else N
- PRONOUN_2: Y if second is a pronoun, else N
- DEFINITE_2: Y if second starts with "the"; else N
- DEMONSTRATIVE_2: Y if second start with "this","that"...

Features (2)

Soon et al. 2001

- NUMBER: C if both agree in number, I if they disagree, NA if number information is not available for at least one
- BOTH_PROPER_NOUNS: C if both NPs are proper names; NA if exactly one; I otherwise
- APPOSITIVE: C if the NPs are in an appositive relationship; else I
- WNCLASS: C if the NPs have the same WordNet semantic class; I if they do not; NA if not available
- SENTNUM: distance between the NPs in terms of the number of sentences;

Comments on this approach

Markov Logic approach

Base MLN: Use of the following predicates:

- InClust(m, c)
- head(m, h)
- Type(x, e)
- Number(x, n)
- Gender(x, g)
- IsPrn

Basic Knowledge Base

- $\text{InClust}(+m, +c)$;
- $\text{InClust}(m, +c) \wedge \text{Head}(m, +t)$;
- $\neg\, \text{IsPrn}(m) \wedge \text{InClust}(m, +c) \wedge \text{Head}(m, +t)$;
- $\text{Type}(+x, +e)$;
- $\text{InClust}(m, c) \Rightarrow (\text{Type}(m, e) \Leftrightarrow \text{Type}(c, e))$;
- $\text{IsPrn}(m) \wedge \text{InClust}(m, c) \wedge \text{Head}(m, +t), \wedge \text{Type}(c, +e)$

Adding apposition

$$\text{Appo}(x, y) \Rightarrow (\ \text{InClust}(x, c) \Leftrightarrow\ \text{InClust}(y, c))$$

Rule-based MLN

The three agreement rules plus one clustering non-pronouns by their heads.

Multi-pass approach

Key idea: perform first easy deductions, and use them to perform harder deductions afterwards.

Reference paper:
Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nate Chambers, Mihai Surdeanu, Dan Jurafsky, Christopher D. Manning: A Multi-Pass Sieve for Coreference Resolution. EMNLP 2010: 492-501

Pass 1

Exact match: two mentions are linked only if they contain exactly the same extent text.

Pass 2

Precise construct: if any of the conditions below are satisfied:

- two nominals mentions are in appositive construction.
- subject-object relation
- role appositive [[actress] Rebecca Schaeffer]
- relative pronoun: the mention is a relative pronoun that modifies the head of the antecedent NP
- acronym
- demomy

Pass 3

Strict head matching. All the following features must be matched:

- the mention head matches any head word in the antecedent cluster
- all the non-stop words in the mention cluster are included in the set of non-stop words in the cluster of the antecedent candidate
- the mention modifiers are included in the modifiers of the antecedent candidate
- one mention cannot be a child in the other NP

- the mention head matches any head word in the antecedent cluster
- all the non-stop words in the mention cluster are included in the set of non-stop words in the cluster of the antecedent candidate
- the mention modifiers are included in the modifiers of the antecedent candidate
- one mention cannot be a child in the other NP

Pass 6

Relaxed Head Matching
Relaxes cluster head match heuristic by allowing the mention head to match any word in the cluster of the candidate antecedent.
Both mention and antecedent are required to be named entities with coinciding types.

Pass 7: Pronouns

Based on agreement constraints (number, gender,...).