Computational Linguistics

Coreference Resolution

Gosse Bouma

Information Science
University of Groningen

LOT 2009

Overview

1. Coreference
2. Applications
3. Corefence Resolution
   - Pronouns
   - Definite NPs
   - Follow-up Questions

Coreference

- Noun Phrases A and B coref if they refer to the same entity
- If A and B coref, and the interpretation of B depends on A, then A is the antecedent and B the anaphor.
- Anaphors can be pronouns, definite NPs, and Proper Names

Xavier Malisse heeft zich geplaatst voor de halve finales. Hij versloeg de Spanjaard Ramirez. In de halve finale treft de Belg een onbekende tegenstander.
Xavier Malisse goes to the semi finals. He beat Spaniard Ramirez.In the semi-finals the Belgian meets an unknown opponent.

Steve Stevaert dreigt met een regeringscrisis. Stevaert ergert zich aan de manier waarop de verschillende ministeries het dossier naar elkaar toeschuiven.
Steve Stevaert threathens with a crisis. Stevaert is annoyed by the way ministries pass the issue to each other

Relation Extraction

- Extract instances of a given relation from the corpus
- X is a symptom of Y

Een van de symptomen van een Vitamine A deficiëntie is een slecht reuk en/of smaakvermogen
One of the symptoms of Vitamin A deficiency is poor ability to smell

Blauwtong is een virusziekte die voornamelijk voorkomt bij schapen. Een van de symptomen van de ziekte is de blauwe tong die besmette dieren kunnen krijgen.
Blue tongue is a virus disease that occurs with sheep. One of the symptoms ofthe disease is the blue tongue infected animals can get.

Relation Extraction

- Extract instances of a given relation from the corpus
- X is a symptom of Y

Een van de symptomen van een Vitamine A deficiëntie is een slecht reuk en/of smaakvermogen
One of the symptoms of Vitamin A deficiency is poor ability to smell

Blauwtong is een virusziekte die voornamelijk voorkomt bij schapen. Een van de symptomen van de ziekte is de blauwe tong die besmette dieren kunnen krijgen.
Blue tongue is a virus disease that occurs with sheep. One of the symptoms ofthe disease is the blue tongue infected animals can get.

Question Answering

Vraag Wie sneuvelde bij Heiligerlee?

Antw Dat leidde tot de slag bij Heiligerlee, waarbij zijn broer Adolf sneuvelde [AD 2003]

Antw Adolf trok vervolgens in de troepenmacht van zijn andere broer Lodewijk mee naar het noorden, waar hij sneuvelde bij Heiligerlee. [wikipedia]

Que Who died at Heiligerlee?

Ans This lead to the battle of Heiligerlee, where his brother Adolf died

Ans Adolf moved north with the army of his brother Louis , where he died at Heiligerlee

Follow-Up questions

Who is the murderer of John Lennon?

10 may - 1955 – Mark David Chapman, murderer of John Lennon

How often was he hit?

he → John Lennon

Lennon was hit four times and died at 11.15 pm.

Where was he murdered?

he → John Lennon

John Lennon On All Music Guide

Follow-Up questions

Who is the murderer of John Lennon?

10 may - 1955 – Mark David Chapman, murderer of John Lennon

How often was he hit?

he → John Lennon

Lennon was hit four times and died at 11.15 pm.

Where was he murdered?

he → John Lennon

John Lennon On All Music Guide

Follow-Up questions

Who is the murderer of John Lennon?

10 may - 1955 – Mark David Chapman, murderer of John Lennon

How often was he hit?

he → John Lennon

Lennon was hit four times and died at 11.15 pm.

Where was he murdered?

he → John Lennon

John Lennon On All Music Guide

Follow Up Questions and Anaphora

personal and possessive pronouns

- When was Napoleon born?
- Which title was introduced by *him*?
- Who were *his* parents?

impersonal pronouns

- What is the KNMI?
- When was *it* founded?

deictic pronouns

- What is an ecological footprint?
- When was *this* introduced?

Follow Up Questions and Anaphora

personal and possessive pronouns

- When was Napoleon born?
- Which title was introduced by *him*?
- Who were *his* parents?

impersonal pronouns

- What is the KNMI?
- When was *it* founded?

deictic pronouns

- What is an ecological footprint?
- When was *this* introduced?

Follow Up Questions and Anaphora

personal and possessive pronouns

- When was Napoleon born?
- Which title was introduced by *him*?
- Who were *his* parents?

impersonal pronouns

- What is the KNMI?
- When was *it* founded?

deictic pronouns

- What is an ecological footprint?
- When was *this* introduced?

Follow Up Questions and Anaphora

definite NPs

- Since when is Cuba ruled by Fidel Castro?
- When was the flag of *the country* designed?

deictic NPs

- Who lead the Russian Empire during the Russian-Turkish War of 1787-1792?
- Who won *this war*?

Follow Up Questions and Anaphora

definite NPs

- Since when is Cuba ruled by Fidel Castro?
- When was the flag of *the country* designed?

deictic NPs

- Who lead the Russian Empire during the Russian-Turkish War of 1787-1792?
- Who won *this war*?

Resolving Coreference

- Position
    - number of sentences or NPS between antecedent and anaphor
- Lexical, morphology
    - pronoun, proper noun, definiteness, number
- Syntax
    - SUBJ, OBJ, PREDC, APP, ...
- String-matching
- Semantic
    - gender, category of proper names, synonyms, hypernyms, ...

Resolving Pronouns (Bouma and Bouma)

Experimental Framework

Mur (2008) for PN and full NPs:

- Rule-based, string matching resolution of PNs
- Manually weighted linear model (cf Shallom & Lappin, 1997) for definite full NPs
- determination of anaphoricity of all NPs rule-based.

Runs on top of the Alpino parser

Experimental framework - pronoun module

Maximum Entropy ranking model for pronoun resolution

$$P(ant \mid pron) = \frac{1}{Z} \times \exp(\sum_{f \in Feats} w_f \, f(pron, ant))$$

Maxent models trained with TADM (Malouf, 2003), on the last mention of each compatible referent in the last 10 sentences. Several gaussian priors. During application we pick the most likely candidate, provided it is better than a set threshold.

Experimental framework - pronoun module

14 features, capturing:

- GF of candidate
- NP form of candidate
- Ontological status of candidate
- Distance between pronoun and candidate
- Frequency of mention candidate referent

Evaluation syntactic features

10-fold cross-validation on KNACK training corpus (Hoste & De Pauw, 2006)

Model	Opts	Prc Pron	MUC F	MUC P	MUC R
nearest		31.0	42.7	49.4	37.6
sub & sent	$-.5$	61.1	50.3	59.8	43.4
syntax	$1, -.5$	61.3	51.8	60.1	45.5

Cf Hendrickx et al (2008): 51.3 MUC F on this corpus.

Plausibility

“Gosse$_{=g}$ sneed een stuk parfait$_{=t}$ af en legde het mes$_{=m}$ weg. Het$_{=?}$ smaakte hem$_{=g}$ heerlijk!”
Frequency based predicate-argument association:
parfait–su– smaak vs *mes –su– smaak*

- Dagan et al. 1995: small improvement
- Kehler et al. 2004: no improvement
- Yang et al. 2005: some improvement esp with web as corpus

Predicate-argument frequencies

Predicate-argument cooccurrence info from automatically parsed TwNC & Wikipedia ($>$525mln words). 37mln sub-verb pairs, 18mln obj-verb pairs. Association between pred-arg Pointwise MI:

$$PMI(pred, arg) = \log \frac{P(pred, arg)}{P(pred) \times P(arg)}$$

After frequency filtering, about 2mln types (sub&obj1)
The value of the *association feature* for a candidate antecedent is the largest PMI of the coreferents.

Evaluation MI

Model	Opts	Prc Pron	MUC F	MUC P	MUC R
nearest		31.0	42.7	49.4	37.6
sub & sent	$-.5$	61.1	50.3	59.8	43.4
syntax	$1, -.5$	61.3	51.8	60.1	45.5
MI	$1, -.5$	61.2	51.8	60.1	45.6

Overcoming sparseness with similar words?

Kehler et al. (2004) & Yang et al. (2005) identify data sparseness as a problem.

“Gosse$_{=g}$ sneed een stuk parfait$_{=t}$ af en legde het mes$_{=m}$ weg. Het$_{=?}$ smaakte hem$_{=g}$ heerlijk!”

parfait –su– smaak

ijs –su– smaak
cake –su– smaak
gebak –su– smaak
taart –su– smaak

Overcoming sparseness with similar words?

Word similarity can be calculated with pred-arg frequencies, too (Bouma & vd Plas, 2005).
Our approach:

- for each noun, form a vector of PMIs with predicates
- use DICE to calculate the similarity between vectors
- pick the 15 most similar words to form a cluster
- *association feature* is now the maximum of maximum

Evaluation MI Similar Words

Model	Opts	Prc Pron	MUC F	MUC P	MUC R
nearest		31.0	42.7	49.4	37.6
sub & sent	$-.5$	61.1	50.3	59.8	43.4
syntax	$1, -.5$	61.3	51.8	60.1	45.5
MI	$1, -.5$	61.2	51.8	60.1	45.6
MI Sim	$1, -.5$	59.7	51.4	59.6	45.2

Model inspection

subject	0.1036
direct object	-0.0066
indirect object	0.0062
oblique complement	-0.0058
head noun	0.0862
sentence	-0.4694
human	0.0002
pronoun	0.0353
definite	-0.0693
same paragraph	0.1083
preceding sentence	0.0111
mentions	0.0762
mi(su)	0.0575
mi(obj1)	0.0174

Discussion

- Small improvement over baseline with syntactic information.
- No noticeable effect of PMI as association information.
- More similar words?
- Another way of calculating a combined association score?
- Learn MI from other data (i.e. coref-corpus = Flemish, MI-corpus = Dutch)

Resolving Definite NP anaphors

- Definite NP often mentions the semantic class of the antecedent

Todd Martin was the opponent of the quiet Ivanisevic in December 1995. The American, who defeated the local hero Boris Becker a day earlier, was beaten by the 26-year old Croatian during the finals of the Grand Slam Cup in 1995

- Relevant knowledge can be found in apposition relations

Acquiring ISA relations

- Corpus is searched exhaustively for ⟨ Concept, app, Instance ⟩
  - museum Hermitage, Madame Tussaud, National Gallery, ...(1.945)
  - bondscoach Guus Hiddink, Jorge Valdano, Louis van Gaal, ... (14K)
  - Argentinian newspaper La Nación, biologist Lilian Ramos, supermarket Disco, ... (1.861)
- 3.2M appositions extracted for 660K Named Entities

Using Anaphora Resolution in IE

Increase in # of extracted facts

	original	anaphora
age	17.038	20.119
born_date	1.941	2.034
born_loc	753	891
died_age	847	885
died_date	892	1.061
died_how	1.470	1.886
died_loc	642	646

Using Anaphora Resolution in IE

Accuracy on 400 random coref. facts

	# facts
new facts(corr.)	168
new facts(incorr.)	128
increase in frequency(corr.)	91
increase in frequency(incorr.)	6

Anaphora Resolution in Questions

- Antecedent has to be a named entity
- From first question or the answer to the first question

What is the capital of Russia?

Moscow

How many inhabitants does it have

8 million

Anaphora Resolution Results

Questions	200	
Qs with Anaphor	56	100%
Correct Antecedent	29	52%
Wrong Antecedent	15	27%
Missed	12	21%

Problematic Cases

Antecedent is not a named entity

- Wat is mede?
- Hoe heet het in India?

Locative and Temporal Anaphora

- Hoe groot is Pitcairn?
- Welke talen worden er gesproken?

- Wanneer werd Contra-Aquincum gesticht?
- 294
- Welke keizer was destijds aan de macht?

Problematic Cases

Antecedent is not a named entity

- Wat is mede?
- Hoe heet het in India?

Locative and Temporal Anaphora

- Hoe groot is Pitcairn?
- Welke talen worden er gesproken?

- Wanneer werd Contra-Aquincum gesticht?
- 294
- Welke keizer was destijds aan de macht?

Problematic Cases

Bridging?

- In welke gemeente ligt Helvoirt?
- Hoe heet het jaarlijkse evenement rond Hemelvaartsdag?

- Wanneer werd de Efteling geopend ?
- Welke nieuwe attractie werd geopend in 1993 ?