Automatic Reference Resolution

A quick overview

Pascal Denis

Alpage
INRIA Paris-Rocquencourt
pascal.denis@inria.fr

Atelier JSM 2010
Nancy, 24 March 2010

Slides at: http://alpage.inria.fr/pdenis/jsm10.pdf

What this course is about

- Give you a short break from logic and λ-calculus (and donkeys!)
- Give you a feel for computational treatment of reference
  - How to build systems that perform reference resolution?
  - What are the main approaches?
  - How to evaluate them?
  - What corpora are out there?

Reference Resolution

An important part of natural language understanding

Clinton told National Public Radio that his answers to questions about Lewinsky were constrained by Starr's investigation. NPR reporter Mara Liasson asked Clinton "whether you had any conversations with her about her testimony, had any conversations at all."

Reference Resolution

An important part of natural language understanding

[Clinton] told [National Public Radio] that his answers to questions about [Lewinsky] were constrained by [Starr]'s investigation. [[NPR] reporter Mara Liasson] asked [Clinton] "whether [you] had any conversations with [her] about [her] testimony, had any conversations at all."

Reference Resolution

An important part of natural language understanding

[Clinton]$_{PER}$ told [National Public Radio]$_{ORG}$ that his answers to questions about [Lewinsky]$_{PER}$ were constrained by [Starr]$_{PER}$'s investigation. [[NPR]$_{ORG}$ reporter Mara Liasson]$_{PER}$ asked [Clinton]$_{PER}$ "whether [you]$_{PER}$ had any conversations with [her]$_{PER}$ about [her]$_{PER}$ testimony, had any conversations at all."

Reference Resolution

An important part of natural language understanding

Clinton told that his answers to questions about Lewinsky were constrained by 's investigation. NPR reporter Mara Liasson asked Clinton "whether you had any conversations with her about her testimony, had any conversations at all."

Reference Resolution

An important part of natural language understanding

Clinton told National Public Radio that his answers to questions about Lewinsky were constrained by Starr's investigation. NPR reporter Mara Liasson asked Clinton "whether you had any conversations with her about her testimony, had any conversations at all."

Reference Resolution
Two main tasks

Anaphora Resolution

Task of finding an antecedent for each anaphor (typically, pronouns)

Coreference Resolution

Task of partitionning the set of mentions into equivalence classes (or chains) that correspond to discourse entities

Motivations
Why is it useful?

A lot of applications depend on reference resolution:

- Information Extraction
  - Merge information about the same entity
- Question Answering / Information Retrieval
  - Facilitate question/answer and query/document matching
- Machine Translation
  - Literal translations of pronouns can produce funny translations (e.g., *elle* $\rightarrow$ {*she*,*it*})
- Text Summarization
  - Identify salient coreference chains

Motivations
Why is it hard?

- Ambiguity: most referential expressions refer to different things in different contexts
- Knowledge-rich: numerous information sources at play: linguistic and non-linguistic (AI-complete problem!)
- Defeasability: few of the sources are completely reliable
- Noisy Input: reference resolution is at the end of the NLP pipeline (error propagation)
- Different subproblems: different referring expressions need different resolution strategies (cf. Ariel 1988, Gundel et al. 1994)

Outline

Outline

Types of referring expressions

Natural languages provide many ways to refer to things:

- Proper names: full PNs (*George Walker Bush*) and short PNs (*Bush*, *W*)
- Indefinite NPs: *some dude from Texas*, *a donkey*
- Definite NPs: *the former president*, *the village idiot*
- Demonstrative NPs: *that dude*
- Pronouns: *he*, *his*, *that*, ∅

Information status / Salience

The use of a particular expression depends on the level of activation (salience, accessibility, givenness) of its referent in discourse model (Prince, 1981; Gundel et al., 1993; Ariel 2001)

Accessibility scale (Ariel, 2001)

Full PNs > Long Def. Desc. > Short Def. Desc. > last name > first name > distal dem. > proximal dem. > stressed pron. > unstressed pron.

Anaphora vs. coreference

- Coreference: referent(α) = referent(β)
  - Equivalence relation
  - Not discourse-bound
- Anaphora: intepretation of α depends on intepretation of β
  - Irreflexive, non-symmetrical
  - Discourse-bound

Anaphoric relations beyond strict coreference

- Discontinuous sets

  [Jerry] owns [a Saab] and [George] owns [a LeBaron]. [They] drive [them] all the time.

- Bridging anaphora

  George almost bought [the Volvo], but [the color] wasn't at his taste.

- Bound anaphora

  [No Frenchman] believes that World Cup referees treated [his] team fairly.

- Generics

  Elaine ate [5 juicyfruits] yesterday. This tells you how much she likes [them].

Beyond individual anaphora

- Verbal anaphora

  Jerry [went to the dentist], and so did _ Kramer.
  Jerry [went to the dentist] and Kramer _ to the masseuse.
  Jerry [was mad at Elaine] but he didn't remember why _.

- Temporal anaphora

  Kramer entered the apartment. Jerry was talking to George.

- ...

Not all "anaphoric" expressions are anaphora

- Expletives

  It is half past two.
- First mention definites

  The US president uncovered his new healthcare plan.
- Exophora

  Pick that up and put it over there.
- Discourse deixis

  We'll talk about this in the next chapter.

Outline

(Morpho-)syntactic constraints

Gender, number, person agreement

- George bought [a LeBaron convertible]$_i$.
  - # She$_i$ /It$_i$ used to belong to Jon Voight.
  - He likes # them$_i$ /it$_i$.
  - He likes # you$_i$ /it$_i$.

Binding theory

- George$_i$ bought himself$_{i/*j}$ a LeBaron convertible.
- George$_i$ bought him$_{*i/j}$ a LeBaron convertible.

Semantic constraints

Selectional constraints

- Jerry bought coffee$_i$ from the store$_j$. George drank it$_{i/\#j}$.

Discourse accessibility

- George didn't buy a Volvo. #It was blue.

Preferences in pronoun interpretation

Recency

Jerry owns a Saab. George owns a LeBaron. Elaine likes to drive it. [it=a LeBaron]

Repetition

George needed a new car. His previous car got totaled. Jerry went with him to the car dealers. He bought a LeBaron. [He=George]

Grammatical role

Jerry went to the car dealers with George. He needed to talk to the mechanic. [He=Jerry]

Preferences in pronoun interpretation (cont'd)

Parallelism

George went with Jerry to the dentist. Elaine went with him to dermatologist. [him=Jerry]

Verb semantics and "implicit cause"

Jerry telephoned George. He lost his keys. [He=Jerry]
Jerry critized George. He lost his keys. [He=George]

Outline

A Generic Algorithm

1. **Identification of referential mentions**
   - Identify noun phrases that have referential content (e.g., filter pleonastic pronouns)
2. **Characterization of mentions**
   - Compute a set of values for $\{k_{i_1}, k_{i_2}, \ldots, k_{i_n}\}$ from n knowledge sources
3. **Anaphoricity determination** (optional)
   - Eliminate non-anaphoric expressions to cut search space
4. **Generation of antecedent candidates**
   - Compute for each anaphoric NP_j a list of antecedent candidates C_j

A Generic Algorithm (cont'd)

1. Filtering (optional)
   - Remove from C_j all candidates that violate hard constraints
2. Scoring/Ranking
   - Order members of C_j according to resolution preferences
3. Search/Clustering
   - Pick the best antecedent and/or cluster NPs that have the same antecedent

Trends in Reference Resolution

- Rule-based vs. corpus-based
- Knowledge-rich vs. knowledge-lean
- Semi-automatic vs. fully automatic preprocessing
- Small-scale vs. large-scale evaluation
- Pronominal anaphora vs. full coreference

Trends in Reference Resolution

- Rule-based vs. corpus-based
- Knowledge-rich vs. knowledge-lean
- Semi-automatic vs. fully automatic preprocessing
- Small-scale vs. large-scale evaluation
- Pronominal anaphora vs. full coreference

Hobbs "Naive" algorithm

Hobbs (1978)

- Simple syntax-based algorithm for third person anaphoric pronouns
- Hobbs algorithm relies on:
  - a syntactic parser
  - a morphological gender and number checker
- Algorithm searches syntactic trees for current and preceding sentences in a breadth-first, left-to-right manner. Stops as soon as it finds a candidate NP that matches in gender and number.
- Gender and number work as hard constraints and ranking of antecedent candidates corresponds to number of NPs skipped by algorithm

The Hobbs algorithm

Hobbs (1978)

1. Begin at NP node immediately dominating the pronoun.
2. Go up the tree to first NP or S node. Call this X, and the path *p*.
3. Traverse all branches below X to the left of *p* in a left-to-right fashion. Propose as antecedent any NP that has a NP or S between it and X.
4. If X is the highest S in the sentence, traverse the parse trees of the previous sentences in the order of recency. Traverse left-to-right, breadth first. When a NP is encountered, propose as antecedent. If not the highest node, go to Step 5.
5. From node X, go up the tree to the first NP or S. Call it X, and the path *p*.
6. If X is an NP and the path to X did not pass through the nominal that X dominates, propose X as antecedent
7. Traverse all branches below X to the *left* of the path, in a left-to-right, breadth first manner. Propose any NP encountered as the antecedent
8. If X is an S node, traverse all branches of X to the *right* of the path but do not go below any NP or S encountered. Propose any NP as the antecedent.
9. Go to Step 4.

The Hobbs algorithm (cont'd)

Simplified version

1. right-to-left search in the current sentence, starting with the first c-commanding NP to the left of the pronoun
2. while no antecedent is found, left-to-right search in preceding sentence
3. if still no antecedent found, search current sentence from left-to-right, starting with the first NP to the right of the pronoun (for cataphora)

Exercise: Find an example for which Hobbs algorithm predicts the wrong antecedent.

The Hobbs algorithm (cont'd)

An example where it gets it wrong

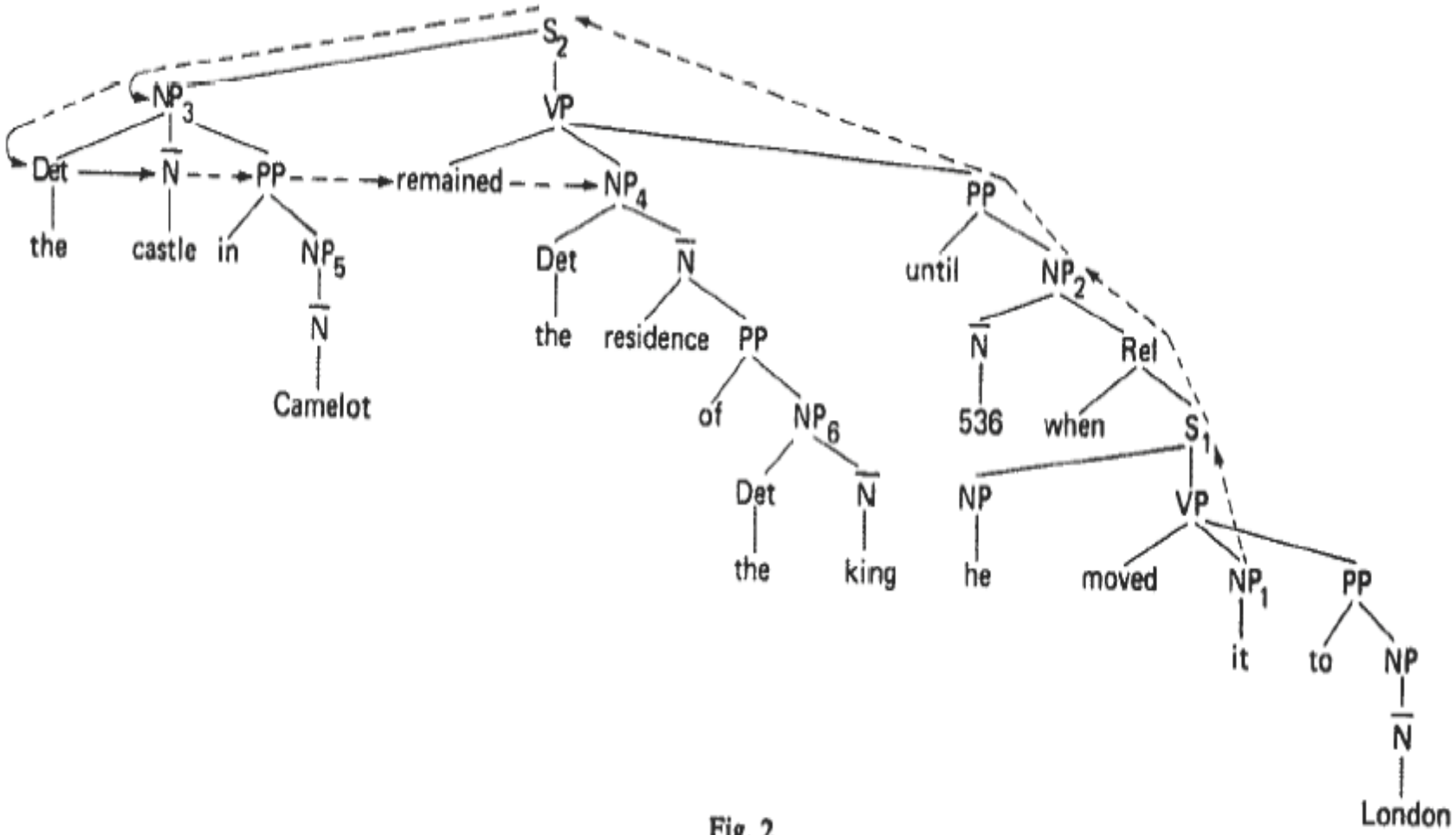

Fig. 2.

The Hobbs algorithm (cont'd)
Summary

- Hobbs (1978) reports an accuracy of 88% based on manual evaluation and perfect syntactic and morphological info (with 90+% of intra-sententials)
- Requires full parser
  - not available for some languages
  - highly sensitive to parsing errors (parsing accuracy still low on entire sentences)
- Does not capture all constraints and preferences on pronoun resolution: approximate binding theory, recency, and grammatical role
- Still useful baseline for evaluating pronominal anaphora resolution algorithms
  - More complex systems sometimes incorporate "Hobbs distance" (e.g., Kehler et al. 2004)

Lappin & Leass (1994)
Resolution of Anaphora Procedure (RAP)

- salience-based resolution system for third person pronoun
- Combines syntactically-based preferences with the effects of the recency in a simple discourse model
- RAP in a nutshell:
  - Add every new discourse entity to discourse model
  - Update its value based on salience factors (recency, grammatical roles)
  - Values cut in half each time a new sentence (recency enforcement)
- Scoring/ranking of antecedent candidates based on additive weights

Lappin & Leass (1994) (cont'd)

Initial weighting scheme

Salience factor	weight
Sentence Recency	100
Subject Emphasis	80
Existential Emphasis	70
Accusative	50
Indirect Object	40
Non-adverbial Emphasis	50
Head-noun Emphasis	80

Lappin & Leass (1994) (cont'd)
Summary

- Using the perfect output from a morphological analyzer and a full syntactic parser, L&L report 86% (on 360 pronoun occurrences)
- Also requires full parser, although there are knowledge-lean extensions of RAP: Kennedy & Boguraev (1996) and Mitkov (2001)
- Interesting attempt to combine different types resolution preferences, but weighting scheme for combining preferences is "fiddly": no guarantee that it is optimal, that it is "portable"

Machine learning reference resolution

The last decade of research in NLP has seen the emergence of machine learning approaches

- ML approaches are more robust, making them easier to integrate into larger systems
  - better at handling noisy data (e.g., errors pre-processing)
  - better at dealing with exceptions
- They are more portable to different domains and different languages (provided one has an annotated corpus)
- They also perform better than their hand-crafted correspondants

Machine learning, informally

- Previous approaches are deductive: they rely on hand-crafted rules/heuristics to perform resolution
- By contrast, machine learning approaches are inductive: they "learn" from examples seen in a "training" set (i.e., an annotated corpus)
- Typically, the function to learn is a classification function
- Examples are represented as feature vectors: e.g. [0,0,0,1,0,0,1,0]
- Learning consists in finding the weights associated with the various features using numerical methods
- Typical learning criteria:
  - maximize likelihood of training set (Naive Bayes, MaxEnt)
  - minimize error rate on training set (perceptron, SVM)

Reference resolution as binary classification

Soon et al, 2001, Ng & Cardie, 2002, Kehler et al. 2004 (inter alia)

- Reference resolution is reformulated as a binary classification task: pair of mentions is classified coreferential (1) or not (0)
- Training data consist pairs of mentions, represented as feature vectors, with a class label (0 or 1)
- A binary classifier is induced from the training data using a particular learning algorithm (Decision Trees, MaxEnt)
- Classifier is then applied to test instances: determines whether two mentions are being coreferential (based on score or probability)
- Antecedent is the NP that maximizes this score/probability
- Coreference chains can be obtained by applying greedy clustering on the coreferential pairs

Typical features

Soon et al. (2001) features

Feature type	Feature Name	Description
String matching	STR_MATCH	T if, after removing DET's, the strings of NP*i* and NP*j* match; else F
Recency	S_DIST	Number of sentences between NP*i* and NP*j*
NP type	I_PRO	T if NP*i* is a pronoun; else F
	J_PRO	T if NP*j* is a pronoun; else F
	J_DEF	T if NP*j* starts with word *the*; else F
	J_DEM	T if NP*j* starts with word *this*, *that*, *these*, *those*; else F
	BOTH_PN	T if NP*i* and NP*j* are both PNs; else F
Morphosyntactic Agreement	NUM_AGR	T if NP*i* and NP*j* agree in number; F if they disagree; UNK if either NP's number cannot be determined
	GEN_AGR	T if NP*i* and NP*j* agree in gender; F if they disagree; UNK if either NP's gender cannot be determined
Sortal Agreement	WN_CLASS_AGR	T if NP*i* and NP*j* have the same WN class; F if they don't; UNK if either NP's class cannot be determined
Alias	ALIAS	T if one NP is an alias of the other; else F
Appositive	APPOSITIVE	T if the NPs are in an appositive relation; else F

Outline

Most popular corpora

- Two main corpora for English:
  - Message Understanding Conferences (MUC): MUC-6 and MUC-7 corpora (60 and 50 documents, resp.)
  - Automatic Content Extraction (ACE) corpora (3 datasets with around 200 documents)
- Both available from Linguistic Data Consortium (LDC)
- Main motivation: evaluating of Information Extraction (IE) systems
- Unique domain: Newspaper texts

Except from MUC-6

```
<COREF ID="4" TYPE="IDENT" REF="0">American
Airlines</COREF> said <COREF ID="5" TYPE="IDENT"
REF="4">it</COREF> has called for <COREF ID="6"
TYPE="IDENT" REF="7" MIN="mediation">federal
mediation in <COREF ID="9" TYPE="IDENT" REF="10"
MIN="talks" STATUS="OPT"><COREF ID="8" TYPE="IDENT"
REF="5">its</COREF> contract talks with <COREF
ID="11" TYPE="IDENT"REF="12" MIN="unions">unions
representing <COREF ID="19" MIN="pilots"><COREF
ID="13" TYPE="IDENT" REF="8">its</COREF>
pilots</COREF> and <COREF ID="21"
MIN="attendants">flight
attendants</COREF></COREF></COREF></COREF>.
```

Linguistic problems with these corpora

- What counts as a markable?
  - Full NPs, NEs, modifiers, adjectives
  - Annotated named entities are very application-oriented (persons, locations, organizations, but also vehicles and weapons!)
- Very loose definition of coreference
  - Predication not distinguished from coreference:
  - Intensionality problems: *Henry Higgins might be the man you have talked to.*

Other corpora

- MATE/GNOME/ARRAU (Poesio et alii)
  - linguistically sounder: e.g., identity, different predication
  - different types of anaphoric relations: e.g., bridging
  - different domains: WSJ, dialogues, narratives, ...
- Ontonotes
  - WSJ portion of PTB with WSD, semantic roles, coreference (incl. event coreference)
  - high-quality annotations: 90% agreement
- French corpora
  - Ananas from CNRTL (Suzanne Alt-Salmon)
  - Dédé from CNRTL (Hélène Manuelian, Claire Gardent)
  - 1M word corpus from ELDA (Agnès Tutin)
- Anaphoric Bank: international initiative to collect and share anaphoric annotations in different languages

  `www.anaphoricbank.org/`

Evaluation metric for Anaphora Resolution

Accuracy

$$Accuracy = \frac{\#\ anaphors\ that\ are\ correctly\ resolved}{total\ \#\ of\ anaphors}$$

Precision/Recall

$$Recall = \frac{\#\ anaphors\ that\ are\ correctly\ resolved}{\#\ true\ anaphors}$$

$$Precision = \frac{\#\ anaphors\ that\ are\ correctly\ resolved}{\#\ predicted\ anaphors}$$

Evaluation metric for Coreference Resolution
The problem

- We are comparing partitions over sets of mentions
- Three types of errors are possible:
  - wrong resolutions
  - missing resolutions: an anaphor is not resolved
  - spurious resolutions: a non-anaphor is resolved
- Intuitively, we want to reward a system that produces a partition that comes as close as possible to the true partition
- Idea: use minimal link edition needed to perfectly align system chains with reference chains

Evaluation metric for Coreference Resolution

The MUC metric (Vilain et al., 1995)

Let T be the true partition and S the system partition:

$$Recall = \frac{\sum_{s \in S \cap t \in T \neq \emptyset} |s \cap t| - 1}{\sum_{t \in T} |t| - 1} \qquad Precision = \frac{\sum_{s \in S \cap t \in T \neq \emptyset} |s \cap t| - 1}{\sum_{s \in S} |s| - 1}$$

Example

$$T = \{\{A, B, C\}, \{D, E, F, G\}\} \qquad S = \{\{A, B\}, \{C, D\}, \{F, G\}\}$$

$$Recall = \frac{(2-1)+(1-1)+(1-1)+(2-1)}{(3-1)+(4-1)} = \frac{2}{5}$$

$$Precision = \frac{(2-1)+(1-1)+(1-1)+(2-1)}{(2-1)+(2-1)+(2-1)} = \frac{2}{3}$$

Problems with the MUC metric

- Does not always give ver y intuitive evalutions: e.g., $\{A, B, C, D, E, F, G\}$ gets R = 1.0 and P = 5/6 = .83
- First problem: metric intrinsically favors systems that produce fewer, big chains
- Second problem: metric ignores single-mention entities, since they contain no link (cf. MUC annotation)

Exercise: Compute recall and precision scores for the following partitions $S_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}, \{G\}\}$ and $S_2 = \{\{A, B, C, D\}, \{E, F, G\}\}$

Outline

- More sophisticated machine learning models: more global solutions
- Use of unlabeled data
- More semantic knowledge: using lexical databases, using the Web
- More challenging tasks: bridging anaphora, event coreference
- New, collaborative ways of creating annotated data: Phrase Detectives webgame (`www.phrasedetectives.org/`), Amazon Mechanical Turks
- Integration within larger NLP systems and extrinsic evaluation