12 – Anaphora resolution

IA161 Advanced Techniques of Natural Language Processing

M. Medved'

NLP Centre, FI MU, Brno

December 6, 2015

1. Linguistic fundamentals
   - Notation and terminology
   - Coreference
   - Varieties of anaphora according to the form of the anaphora
   - Types of anaphora according to the locations of the anaphora and the antecedent
   - Anaphora and ambiguity

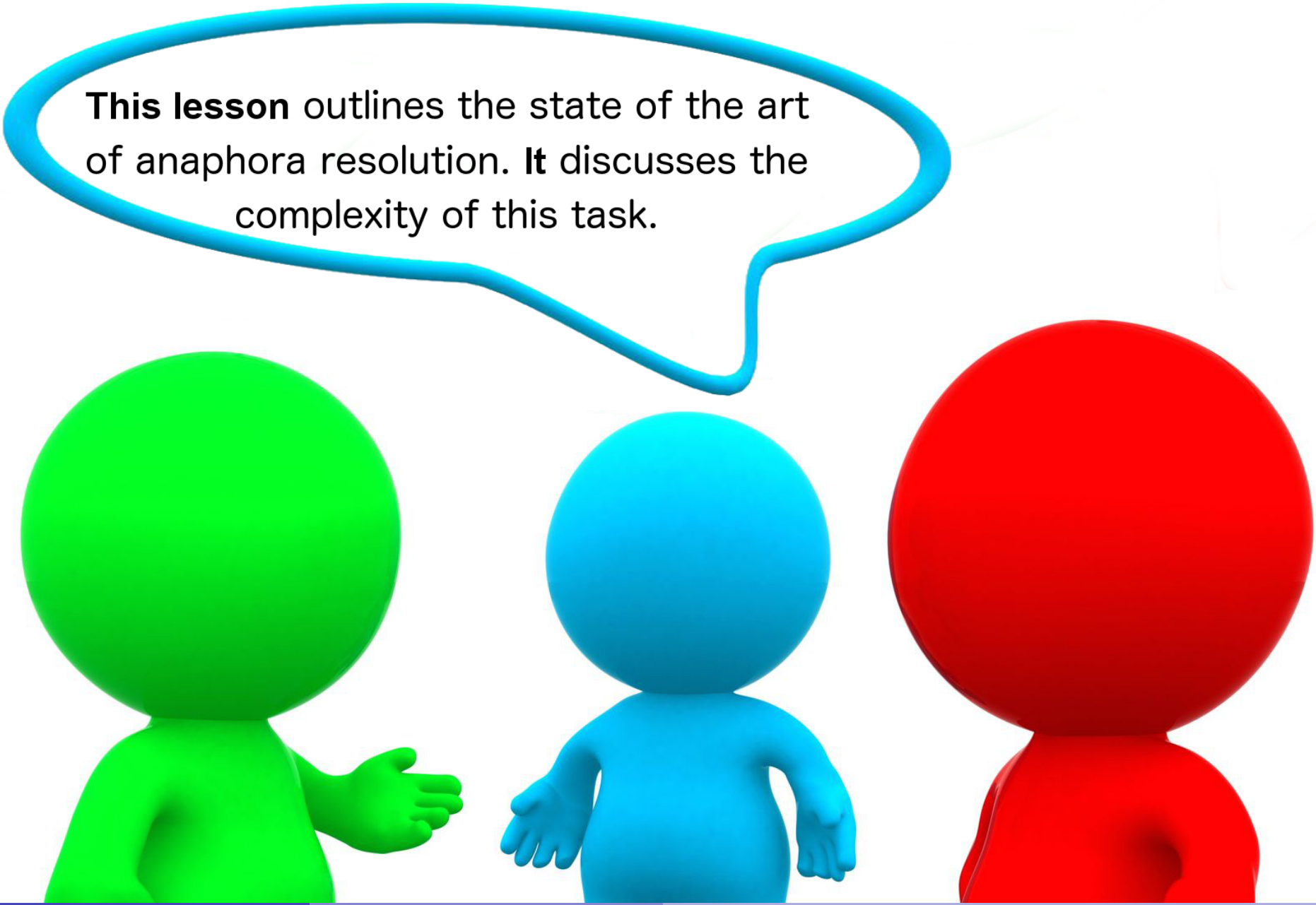
2. The process of automatic anaphora resolution
   - Anaphora resolution input
   - Anaphora resolution in practice
   - The resolution algorithm
     - The resolution algorithm - constraints
     - The resolution algorithm - preferences

3. Theories and formalisms used in anaphora resolution
   - Centering
   - Binding theory

4. Early anaphora resolution approaches
   - Hobbs algorithm

5. Recent anaphora resolution approaches
   - Main trends in recent anaphora resolution research
   - RAP algorithm
   - SAARA system - Czech
   - Results

Discourse


This lesson outlines the state of the art of anaphora resolution. **It** discusses the complexity of this task.

1.
Linguistic fundamentals

Linguistic fundamentals

- cohesion:
  - collection of discourse sentences, phrases or words that are related
- anaphora:
  - **Informal**: cohesion which 'points back' to some previous discourse item
  - **Formal**: an expression the interpretation of which depends upon another expression in context before
- antecedent:
  - discourse entity to which the anaphora refers or for which it stands

Notaion and terminology

- anaphora resolution:
  - process of determining the antecedent of the anaphora
- coreference:
  - anaphora and its antecedent are coreferential when both have the same referent in real world

Example

The Queen is not here yet but **she** is expected to arrive in the next half an hour.

Coreference

- coreferential chain:
  - if more than one preceding noun phrases are coreferential

Example

This book is about anaphora resolution. **The book** is designed to help beginners in the field and **its** author hopes that **it** will be useful.

- definite NPs in copular relation are considered as coreferential:

Example

David Beckham is **the Manchester United midfielder**.

- anaphoric relation does not imply coreference relation between discourse entities:

Example

Every man has **his** own destiny.

Varieties of anaphora according to the form of the anaphora

- nominal anaphora:
  - antecedent has non-pronominal noun phrase
- pronominal anaphora:
  - personal pronouns, possessive pronouns, reflexive pronouns, demonstrative pronouns, relative pronouns
- non-anaphoric usage of pronouns:
  - non-anaphoric uses of **it** (peonastic **it**), generic usage of pronouns, deictic usage of pronouns (pointing to specific person in given situation)
- lexical noun phrase anaphora (definite descriptions and proper names):
- noun anaphora
- verb anaphora
- adverb anaphora
- zero anaphora (elipsis)

Types of anaphora according to the locations of the anaphora and the antecedent

- intrasentential:
  - anaphor and its antecedent are located in the same sentence
- intersentential:
  - antecedent is in a different sentence from the anaphor
- indirect anaphora:
  - reference becomes part of the hearer's or reader's knowledge indirectly rather than by direct mention
- identity-of-sense anaphora:
  - anaphora and the antecedent have the same referent in the real world and are therefore coreferential
- identity-of-reference:
  - does not denote the same entity as its antecedent, but one of a similar description

Anaphora and ambiguity

- problem when identifying antecedent

Example

Jane told Mary she was in love.

2.

The process of automatic anaphora resolution

Anaphora resolution input

- morphological and lexical knowledge
- syntactic knowledge
- semantic knowledge
- discourse knowledge
  - center or focus
- real-world (common-sense) knowledge

Anaphora resolution in practice

- identification of anaphors:
  - tools: morphological analyser, part-of-speech tagger, program for identifying non-anaphoric definite descriptions, parser, annotated corpora, ontology
- location of the candidates for antecedents:
  - tools: full parser (sentence splitter, tokeniser,part-of-speech(POS) taggers), semantic analyser, proper name recogniser
- the resolution algorithm:

The resolution algorithm

- constraints:
  - gender and number agreement
  - c-command constraints
  - selectional restrictions
- preferences:
  - the most recent NP
  - candidates in the main clause
  - NPs which are positioned higher in the parse tree
  - candidates in non-adjunct phrases
  - syntactic parallelism
  - center preference
  - subject preference

The resolution algorithm - constraints

- gender and number agreement

Example

Because **Klein** tried 'dirty tricks', they refused to support **him**.

- c-command constraints

Definition

A node **A** c-commands a node **B** if and only if (i) **A** does not dominate **B**, (ii) **B** does not dominate **A**, (iii) the first branching node dominating **A** also dominates **B**.

Example

She almost wanted Hera to know about the affair.

The resolution algorithm - constraints

- selectional restrictions
  - semantic restrictions that apply to the anaphor should apply to the antecedent as well

Example

George removed the disk from **the computer** and then shut down **it**.

The resolution algorithm - preferences

- syntactic parallelism:
  - noun phrases that have the same syntactic function as the anaphor

Example

The programmer successfully combined **Prolog** with C, but he had combined **it** with Pascal last time.

The resolution algorithm - preferences

- center preference:

Definition

Center is most prominent entity in utterance.

  - sentence that is likely to be pronominalised in a subsequent clause or sentence

Example

Tilly's mother had agreed to make her **a new dress** for the party. She worked hard on **the dress** for weeks and finally **it** was ready for Tilly to try on. Impatient to see what **it** would look like, Tilly tried on **the dress** over her skirt and ripped **it**.

The resolution algorithm - preferences

- subject preference:

Example

The customer lost patience and called the waiter. **He** ordered two 12-inch pizzas.

3.

Theories and formalisms used in anaphora resolution

Centering

Definition

Center is most prominent entity in utterance.

- utterances which continue the topic of preceding utterances

Example

Discourse A
(3.1) John works at Barclays Bank.
(3.2) He works with Lisa.
(3.3) John is going to marry Lisa.
(3.4a) He is looking forward to the wedding.
(3.4b) She is looking forward to the wedding.

Centering

- consists of:
  - forward-looking centre: correspond to the discourse entities evoked by the utterance
  - backward-looking centre: entity connects the current utterance to the previous discourse

Binding theory

- anaphors refer to antecedents that are in a so-called local domain

Example

(3.10) Victoria believed George had seen herself.
(3.11) Victoria believed George had seen him.

4.

Early anaphora resolution approaches

Hobbs algorithm

- one of the most influential works in the field
- algorithm traverses the surface parse tree in a particular order looking for a noun phrase of the correct gender and number

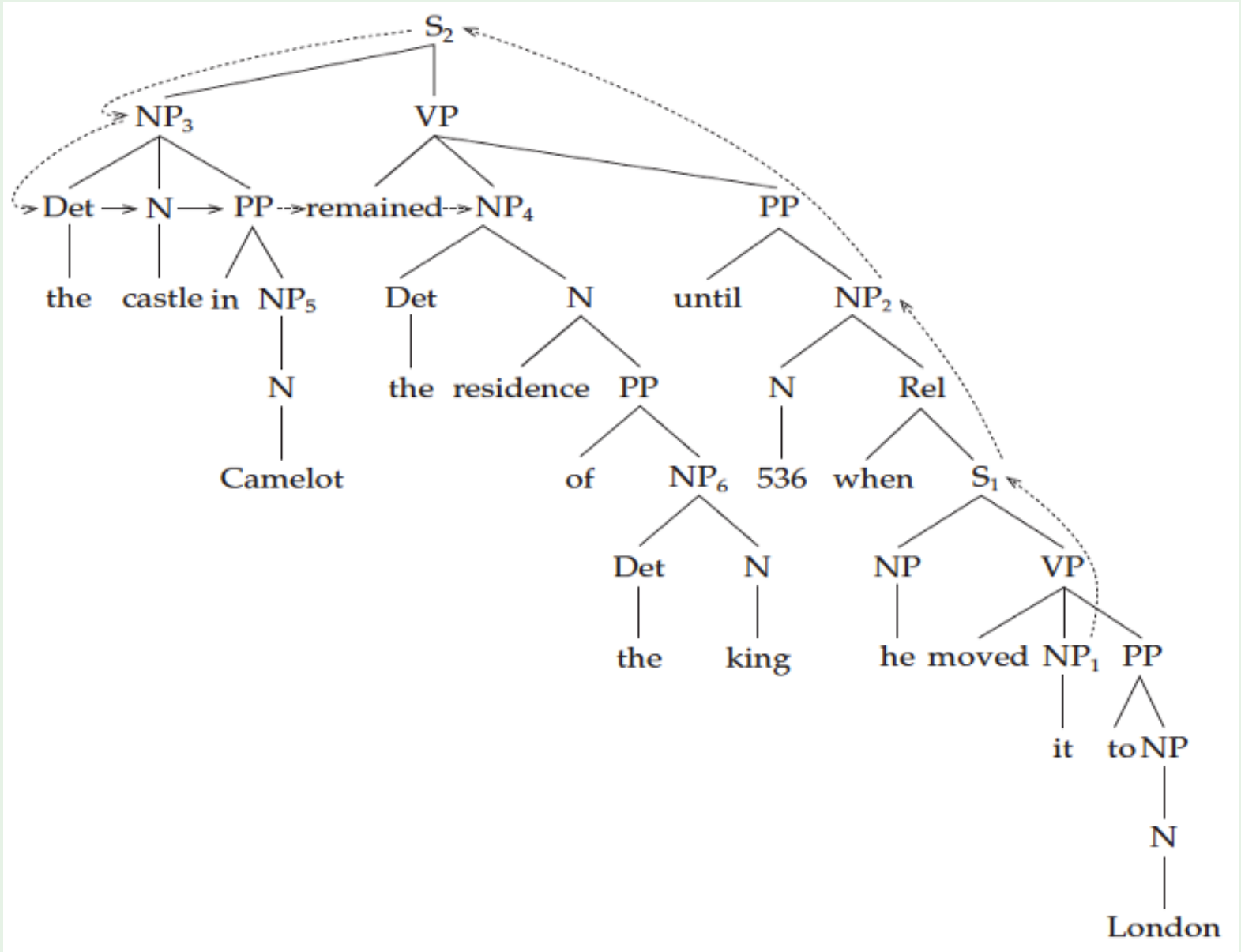
Description of Hobbs algorithm

- Begin at the NP node immediately dominating the pronoun in the parse tree of the sentence S.
- Go up the tree to the first NP or S node encountered. Call this node X, and call the path used to reach it p.
- Traverse all branches below node X to the left of path p in a left-to-right, breadth-first fashion.4 Propose as the antecedent any NP node encountered that has an NP or S node between it and X.
- If the node X is the highest S node in the sentence, traverse the surface parse trees of previous sentences in the text in order of recency, the most recent first; each tree is traversed in a left-to-right, breadth-first manner, and when an NP node is encountered, it is proposed as antecedent. If X is not the highest node in the sentence, proceed to step 5.

Description of Hobbs algorithm: continue

- From node X, go up the tree to the first NP or S node encountered. Call this node X and call the path traversed to reach it p.
- If X is an NP node and if the path p to X did not pass through the N-bar node that X immediately dominates, propose X as the antecedent.
- Traverse all branches below the node X to the left of path p in a left-to-right, breadth-first manner. Propose any NP node encountered as the antecedent.
- If X is S node, traverse all branches of node X to the right of path p in a left-to-right, breadth-first manner, but do not go below any NP or S node encountered. Propose any NP node encountered as the antecedent.
- Go to step 4.

Example (The castle in Camelot remained the residence of the king until 536 when he moved it to London.)



5.

Recent anaphora resolution approaches

Main trends in recent anaphora resolution research

- knowledge-poor approach: cheaper and more reliable corpus-based NLP tools
- corpora:
  - co-occurrence rules
  - training decision trees
  - identify anaphor–antecedent pairs

RAP algorithm

- relies on salience measures derived from the syntactic structure as well as on a simple dynamic model of attentional state to select the antecedent of a pronoun from a list of NP candidates
- does not employ semantic information or real-world knowledge in choosing from the candidates

RAP algorithm components

- An intrasentential syntactic filter for ruling out coreference between a pronoun and an NP on syntactic grounds.
- A morphological filter for ruling out coreference between a pronoun and an NP due to non-agreement of person, number, or gender features.
- A procedure for identifying pleonastic pronouns.
- An anaphor binding algorithm for identifying the possible antecedent of a reflexive or reciprocal pronoun within the same sentence.

RAP algorithm components: continue

- A procedure for assigning values to several salience parameters for an NP, including syntactic role, parallelism of syntactic roles, frequency of mention, proximity, and sentence recency. Higher salience weights are assigned to (i) subject over non-subject NPs, (ii) direct objects over other complements, (iii) arguments of a verb over adjuncts and objects of prepositional phrase adjuncts of the verb, and (iv) head nouns over complements of head nouns.
- A procedure for identifying anaphorically linked NPs as an equivalence class for which a global salience value is computed as the sum of the salience values of its elements.
- A decision procedure for selecting the preferred element from a list of antecedent candidates for a pronoun.

RAP: resolution algorithm

- First a list of all NPs in the current sentence is created and the NPs are classified according to their type (definite NP, pleonastic pronoun, other pronoun, indefinite NP).
- All NPs occurring in the current sentence are examined.
  - NPs that evoke new discourse referents are distinguished from NPs that are presumably coreferential with already listed discourse referents as well as from those used non-referentially (e.g. pleonastic pronouns).
  - Salience factors are applied to the discourse referents evoked in the previous steps as appropriate.
  - The syntactic filter and reflexive binding algorithm are applied.
    - ★ If the current sentence contains any personal or possessive pronouns, a list of pronoun–NP pairs from the sentence is generated. The pairs for which coreference is ruled out on syntactic grounds are identified.
    - ★ If the current sentence contains any reciprocal or reflexive pronouns, a list of pronoun-NP pairs is generated so that each pronoun is paired with all its possible antecedent binders.
  - If any non-pleonastic pronouns are present in the current sentence, their resolution is attempted in the linear order of pronoun occurrence in the sentence.

SAARA system - Czech

- containing re-implementations and variants of selected salience-based algorithms
- contains tree layers:
  - technical layer
  - markable layer
  - supervisor layer
- anaphora resolution algorithm:
  - BFP algorithm
  - RAP algorithm

State-of-the-art results

- **BART** system reports **68%** precision, **64%** recall, **66%** F-measure
- **ARKref** system reports **66%** precision, **55%** recall, **60%** F-measure
- **Reconcile** system reports **66%** precision, **67%** recall, **66%** F-measure

References I

Mitkov, R. (2002).
Anaphora Resolution.
Studies in Language and Linguistics. Longman.

Němčík, V. (2012).
Saara: Anaphora resolution on free text in czech.
In Aleš Horák, P. R., editor, *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2012*, pages 3–8, Brno. Tribun EU.