

EGK Seminar 13.06.2002 Olga Ourioupina

## Overview

- Main Subtasks
- Existing Algorithms

pronominal anaphor definite descriptions – coreference definite descriptions – bridging

My Proposal, Plans, ...

aquiring knowledge using knowledge



### Main Subtasks

- Identifying discourse-new entities
- Finding possible antecedent(s)
- Identifying the Relation Type

## Pronominal anaphor

#### Baseline algorithms:

- take the previous NP
- take the previous subject NP

Accuracy - 60-70%

## 1

### Pronominal anaphor

#### Traditional approaches

- RAP (Lappin & Leass, 1994) syntax-based
- Centering (Grosz, Sidner, 1986) focus tracking
- Mitkov 1994, 1996 syntax-based, incl. semantic/domain modules

#### Accuracy

RAP - 86%

LRC - 80%

Mitkov – 87%

## Pronominal anaphor

#### Alternative approaches

<ul><li>Dagan &amp; Itai, 1990 – corpus-based</li></ul>	87%
---	-----

- Kennedy & Boguraev, 1996 RAP-based, no parsing 75%
- Mitkov, 1996 no parsing
   90%
- Baldwin, 1997 (COGNIAC) no parsing R=64%, P=92%



## Definite NPs (Coreference)

#### Baseline algorithms

- All NPs are coreferential
- All NPs with at least one common word are coreferential
- All NPs with the same head noun are coreferential

#### Accuracy on MUC-6 data (Soon et al., 2001)

	R	Р	F
ALL	89.9	31.8	47.0
ONE_WRD	55.4	36.6	44.1
HD_WRD	56.4	50.4	53.2



## Definite NPs (coreference)

## Models based on Commonsense Reasoning

- Extensive use of hand-coded commonsense knowledge
- Evaluation impossible

Sidner, 1972

Carter, 1987

Alshawi, 1992 (Core Language Engine)

Gardent & Konrad, 1999



## Definite NPs (Coreference)

## Real applications MUC-6 (1995)

```
R P F
best 59% 72% 65%
worst 36% 44% 40%
```

#### MUC-7 (1998)

	R	Р	F
best	56.6%	84.3%	67.7%
worst	52.5%	21.4%	30.4%



## Definite NPs (Coreference)

### Vieira, Poesio, Teufel,.. – knowledgebased approach

- WordNet
- Various heuristics
- (Corpora)

#### Accuracy (F):

Identifying first-mentioned entities 70%

Same-head NPs71-77%

Bridging (incl. Synonyms)33%

## -

## Definite NPs (Coreference)

### Other approaches

Cardie & Wagstaff, 1999 – Coreference as Clustering

Distance Metric based on Feature Vectors

Features: Distance, Animacy,..., Semantic Class (WordNet)

Accuracy on MUC-6 data: R=53%, P=55%, F=54%

 Hartrumpf, 2001 – Combining Syntactico-Semantic rules and Corpus Statistics (German)

**ENTITY and SORT features from MultiNet** 

Accuracy: R=55%, P=82%, F=66%



## Definite NPs (bridging)

Asher & Lascarides

Theoretical analysis

Vieira, Poesio, Teufel

Implemented system, however, the performance is low.



### Gardent&Konrad – Using Model Generation for Definite NPs Resolution

- Huge hand-coded KB required
- Semantic representation of the whole sentence required
- Salience and precedence information not included
- Extremely slow for more than 4-5 entities



### What kind of knowledge do we need?

- (Almost) all the coreference resolution systems make use of WordNet, GermaNet,...
- Soon et al., 2001: 75% mistakes due to the lack of semantic knowledge (63.3% – not enough features, 11.7% – errors in class determination).
- Not too sophisticated knowledge (sortal information, for example).



# Classification of Definite Descriptions based on the information required for their processing (Poesio & Vieira)

- Discourse new
- Anaphoric (same head)
- Syn/Hyp/Mer
- Names
- Compounds
- Events
- Discourse Topic
- Inference

The National Assembly, for the past year,..



# Classification of Definite Descriptions based on the information required for their processing (Poesio & Vieira)

- Discourse new
- Anaphoric (same head)
- Syn/Hyp/Mer
- Names
- Compounds
- Events
- Discourse Topic
- Inference

President Roh Tae Woo's administration

The administration



# Classification of Definite Descriptions based on the information required for their processing (Poesio & Vieira)

- Discourse new
- Anaphoric (same head)
- Syn/Hyp/Mer
- Names
- Compounds
- Events
- Discourse Topic
- Inference

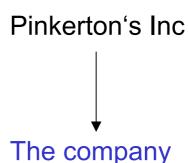
President Roh Tae Woo's administration

The government



# Classification of Definite Descriptions based on the information required for their processing (Poesio & Vieira)

- Discourse new
- Anaphoric (same head)
- Syn/Hyp/Mer
- Names
- Compounds
- Events
- Discourse Topic
- Inference





# Classification of Definite Descriptions based on the information required for their processing (Poesio & Vieira)

- Discourse new
- Anaphoric (same head)
- Syn/Hyp/Mer
- Names
- Compounds
- Events
- Discourse Topic
- Inference

Individual investors and professional money managers contend.

They make the argument ...



# Classification of Definite Descriptions based on the information required for their processing (Poesio & Vieira)

- Discourse new
- Anaphoric (same head)
- Syn/Hyp/Mer
- Names
- Compounds
- Events
- Discourse Topic
- Inference

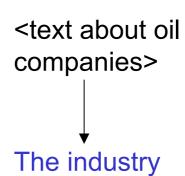
Stock market crash

The markets



# Classification of Definite Descriptions based on the information required for their processing (Poesio & Vieira)

- Discourse new
- Anaphoric (same head)
- Syn/Hyp/Mer
- Names
- Compounds
- Events
- Discourse Topic
- Inference





# Classification of Definite Descriptions based on the information required for their processing (Poesio & Vieira)

- Discourse new
- Anaphoric (same head)
- Syn/Hyp/Mer
- Names
- Compounds
- Events
- Discourse Topic
- Inference

Last week's earthquake

The suffering people



Classification of Definite Descriptions based on the information required for their processing (Poesio & Vieira)

Discourse new

47%

Anaphoric (same head) 30%

Syn/Hyp/Mer

Names

Compounds

20%

Events

Discourse Topic

Inference



Classification of Definite Descriptions based on the information required for their processing (Poesio & Vieira)

Discourse new

Anaphoric	(same head)	19%

• Syn/Hyp/Mer

Names

Compounds 12%

Events

Discourse Topic7%

■ Inference 18%

### WordNet (Miller, 1993)

```
Hyponyms: wordnet, sense 1
```

- => lexical database
  - => electronic database, on-line database,...
    - => database
      - => information, info
        - => message, content,...
          - => communication
            - => social relation
              - =>relation
                - =>abstraction

#### Information in WordNet

- Sorts (for coreference) hypo/hyper
- Synonyms
- Meronyms/holonyms (for bridging)

Poesio, Vieira & Teufel, 1997 – Resolving Bridging References in Unrestricted Text

WordNet Precision – max 28%

Recall – max 46%

#### Problems with WordNet

- Not all the words are covered (Proper Names!)
- Disambiguation problems
- Hierarchy problems
  - 1. jackfruit, jak, jack (immense East Indian fruit resembling breadfruit of ..)
  - 2. jack (an electrical device consisting of a connector socket ..)
  - 3. jack (game equipment ..)
  - 4. jack (small flag indicating a ship's nationality)
  - 5. jack, knave (one of four face cards in a deck bearing a picture of a young prince)
  - 6. jack (tool for exerting pressure or lifting)
  - 7. jack (any of several fast-swimming predacious fishes ..)
  - 8. jack, jackass (male donkey)

#### Problems with WordNet

- Not all the words are covered (Proper Names!)
- Disambiguation problems
- Hierarchy problems

Holonyms: tree\_branch
Sense 1
limb, tree branch
PART OF: tree

#### Problems with WordNet

- Not all the words are covered (Proper Names!)
- Disambiguation problems
- Hierarchy problems
  - 1. (58) cut (separate with or as if with an instrument; "Cut the rope")
  - 4. (2) cut (make an incision or separation; "cut along the dotted line")
  - 29. cut (reap or harvest; "cut grain")
  - 30. cut (fell by sawing; "The Vietnamese cut a lot of timber..")
  - 33. cut (shorten as if by severing the edges or ends of; "cut my hair")
  - 41. cut (...)

#### Problems with WordNet

- Not all the words are covered (Proper Names!)
- Disambiguation problems
- Hierarchy problems

```
Overview: branch
```

- 1. (19) branch, subdivision, arm (an administrative division ..)
- 2. (15) branch (a division of a stem .. of a plant)
- 3. (5) branch, fork, leg (a part of a forked or branching shape)
- 6. (..)

Holonyms: branch

Sense 3

branch, fork, leg

PART OF: furcation, bifurcation, forking

#### Problems with WordNet

- Not all the words are covered (Proper Names!)
- Disambiguation problems
- Hierarchy problems

Hyponyms: geological\_phenomenon geological\_phenomenon

- ⇒ earthquake, quake, temblor, seism
- ⇒ alluvial fan, alluvial cone
- ⇒ catastrophe, cataclysm
- ⇒ continental drift
- ⇒ deposit, sedimentation, alluviation
- ⇒ flood, inundation, deluge, alluvion
- ⇒ frost heave, frost heaving
- ⇒ volcanism

### Knowledge Sources – Corpora

### Selectional constraints and preferences

This book is about the Syberian Tri-colored Rabbit. They eat carrots.

```
They=?
```

⇒ books

⇒ book+rabbit

⇒ .

## Knowledge Sources – Corpora

#### Selectional constraints and preferences

This book is about the Syberian Tri-colored Rabbit. They are carnivorous.

```
They=?
```

- ⇒ books
- ⇒ book+rabbit
- ⇒ .



## Knowledge Sources – Corpora

### Smoothing

- context-based
- class-based (WordNet!)
- alternative



### Knowledge Sources – Internet

### Overcoming data sparseness problem

- Unseen word combinations
- Proper Names classification

#### Problems with Internet

- Noisy unbalanced data
- No possibility of sophisticated search/analysis
- Slow

## Using knowledge

Acquired facts may be unreliable, contradicting, ...

I entered the room.  $\frac{10^3*(f(a,b))^2}{f(a)*f(b)}$ 

The ceiling was high. 10

The size was overhelming. 4.2

The windows looked out to the bay. 3.7

The chandelier sparkled brightly. 0.45



#### Possible solutions

- Probabilistic reasoning
- Nonmonotonic reasoning

## Conclusion

Good and reliable Semantic Knowledge is crucial for coreference resolution systems. Possible knowledge sources:

- WordNet
- Corpora
- Internet
- (Hand-coded) Knowledge Base

#### Current work

- Using Internet for Proper names classification (Geography)
- Baseline algorithm for coreference resolution