Resolution of Anaphoric Reference

Resolution of nominal anaphora

Vašek Němčík

January 31, 2003

WordNet

- still under development at the Princeton University
- initially, meant as a resource to support psychological/psycholinguistic experiments
- George A. Miller and his experiments on human lexical memory and lexical knowledge acquisition
- the underlying psychological ideas influenced the whole area of lexicography

WordNet Characteristics

- the main unit is not *"word"*, but *"concept"*
- a concept is represented as a synonymical set, *"synset"*
- synset is a set of words (which are synonymical in a particular context)
- different *"senses"* of a word belong to different synsets
- includes nouns, verbs, adjectives and adverbs
- synsets are connected by (a *net* of) a number of semantic relations (such as antonymy, hyperonymy, meronymy, etc.)

Selected Semantic relations

- **hyperonymy** – linking a concept and its superconcept
  a *mammal* is a hyperonym of a *dog*
- **hyponymy** – linking a concept and its subconcept
  a *poodle* is a hyponym of a *dog*
- **meronymy** – linking a part to a complexer concept
  a *finger* is a meronym of a *hand*
- **holonymy** – linking the whole to one of its parts
  a *window* is a holonym of a *window pane*
- **troponymy** – linking an action to its manner
  to march is a meronym of *to walk*
- some other relations:
  causes, *is a subevent of*, *has a subevent*, *is an agent of*, etc.

EuroWordNet

- including languages other than only English
- otherwise independent WordNets for individual languages linked using ILI (InterLingual Index)
- language independent **Top Concept Ontology** (most important concepts) and **Domain Ontology** (scripts and scenarios)
- the possibility of comparing H-H subtrees in different languages

EWN Participants

language	*size in synsets*	*institution*
Dutch	30'000	the University of Amsterdam
Spanish	30'000	Fundacíon Universidad Empressa
Italian	30'000	Istituto di Ling. Comp., Pisa
English	120'000	the University of Sheffield
French	7'500-15'000	Université d'Avignon
German	7'500-15'000	Universität Tübingen
Czech	7'500-15'000	Masaryk University, Brno
Estonian	7'500-15'000	University of Tartu

GermaNet

- Division of Computational Linguistics of the Linguistics Department at the University of Tübingen
- one of the extensions of the EuroWordNet
- Statistics (as of 10/2001):

27824	noun synsets
8810	verb synsets
5141	adjective synsets including pertainym synsets
2	adverb synsets
41777	synsets in total
60646	synonyms in synsets
52251	unique word phrases
1.45	synonyms per synset
1.16	word senses per literal

System for resolving definite descriptions

- presented by Masimo Poessio & Renata Vieira in 2000
- **Definite description** – any definite noun phrase with the definite article *"the"*
- other definite NPs, pronouns, demonstratives, possessives, etc. are completely ignored
- assesses a task which is only a subset of resolving the coreference relation in general

Main characteristics of the system

- the system design is based on empirical analysis of corpus data
- compared to human-made annotation of the corpus
- issues forming the backbone of the system:
  - definite description matching techniques
  - discourse segmentation
  - suggesting anchors for bridging descriptions
  - recognition of non-anaphoric use of definite descriptions (definite descriptions are more likely to introduce discourse-new entities than to serve coreferential purposes)

Implications

- almost no chance of using any inference mechanism (the system was tested on a corpus which is too large – there is no knowledge base general enough)
- therefore there was no need to fine-tune the system to a particular domain
- complex cases (where syntax or corpus-inducible information are not enough – mainly bridging relations) are processed using shallow techniques – various heuristics

The corpus used

Penn Treebank I corpus
consisting of Wall Street Journal newspaper articles

Divided into two subsets:

- `Corpus 1` – training data (approx. 1000 definite descriptions) used for development of the system
- `Corpus 2` – test data (approx. 400 definite descriptions) used for evaluating the system

Classification of definite descriptions

- **anaphoric DD** (subsequent mention)
  denotes the same discourse entity as the antecedent
  (further classification follows)
- **associative/bridging DD** (inferrable)
  entity is somehow associated with the antecedent
- **DD introducing a new discourse entity**

Types of anaphoric DDs

- direct anaphora – identical head noun
- synonym
- generalization (hypernym)
- specialization (hyponym)
- antecedent lacking the head noun
  "Bill Clinton ... the president"
- having no antecedent at all
  (used on the basis of common knowledge)
  "the pope"

Experimental annotation classes

I.	II.
direct anaphora	direct anaphpora
"bridging descriptions"	
DDs denoting the same entity	DDs denoting the same entity
DDs denoting an associated entity	DDs denoting an associated entity
discourse-new DDs	discourse-new DDs

Classification remarks

- more than half of all the DDs were discourse-new (!)
- most of the annotators agreed on the distinction between first-mention DDs and subsequent-mention DDs
- very low agreement among the annotators was reached on bridging descriptions (they proposed different anchors)
- the agreement rate was not influenced substantially by the choice of the classification (as proposed in the previous slide)

Various impulses

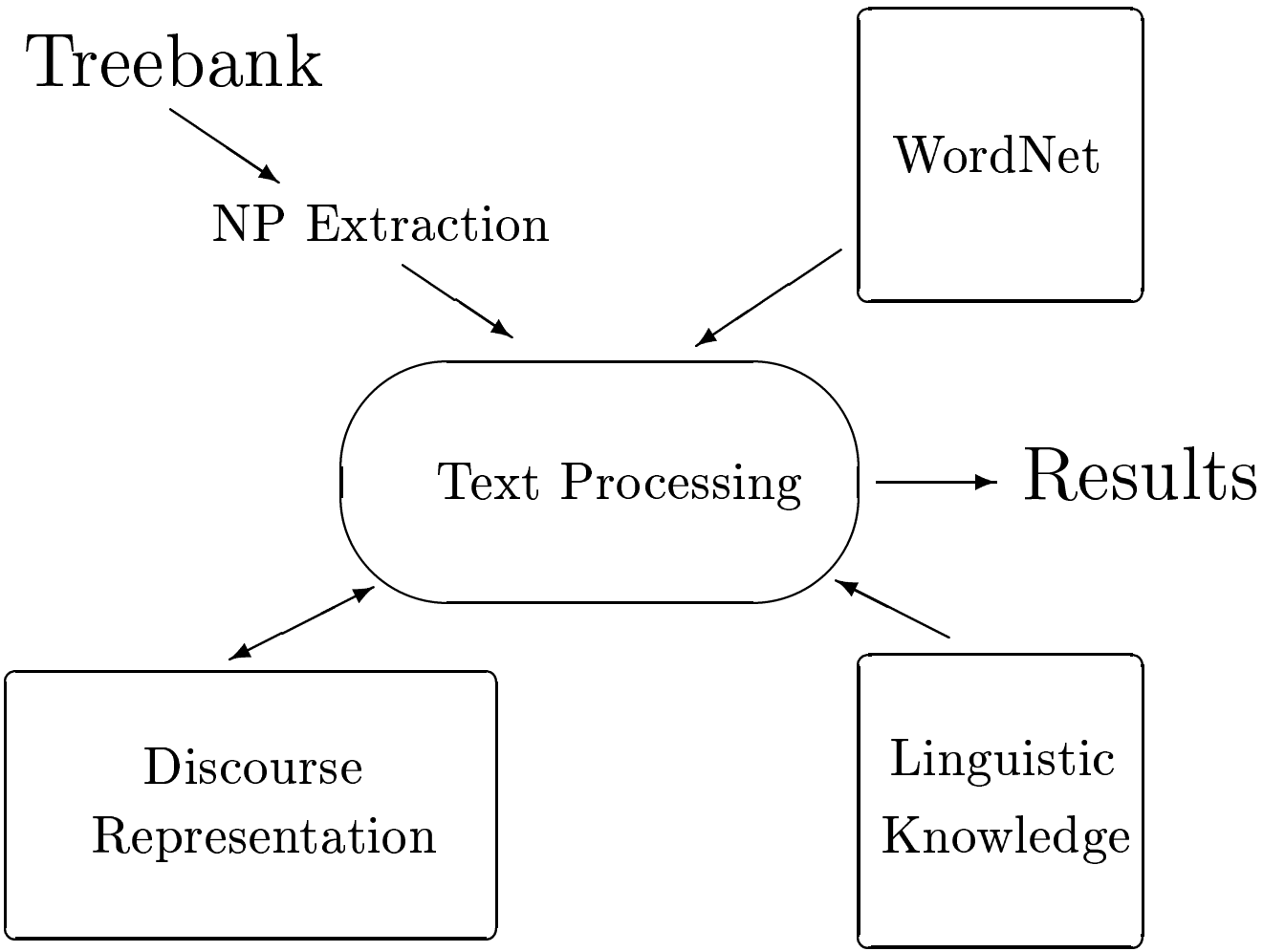
Fraurud and Löbner propose that

- recognition of discourse-new entities is as important as resolution of anaphorical definites
- as such, algorithms handling these problems should be run independently (in parallel)

$$\Longrightarrow$$

The system is based on independent modules whose role in the system can be determined later (based on performance evaluation)

The system architecture



The Heuristics

The sytem is based on a number of various heuristics.
They have been empirically evaluated, configured and given order of their application.
Types of heuristics:

- direct resolution of anaphoric descriptions
  (direct anaphora, segmentation, modification)
- identification of discourse-new descriptions
  (detecting DDs with semantic functions)
- identification of a bridging-description and its anchor
  (detecting semantic relations)

Evaluation methods

The performance was evaluated by comparing the results to the human annotation of the corpus.
Two versions of the system:

- **Version 1**: do not attempt bridging (because it is not developed properly yet)
- **Version 2**: full version of the system

The performance was evaluated using precision and recall when comparing against the majority version annotation.

Measuring the extent of agreement was done using the kappa statistics.

Precision and recall

Recall describes how many of all the cases were actually detected.

$$R = \frac{number\ of\ correct\ responses}{number\ of\ cases}$$

Precision characterizes the proportion of the detected cases detected correctly.

$$P = \frac{number\ of\ correct\ responses}{number\ of\ responses}$$

F measure characterizes the performance by combining both precision and recall

$$P = \frac{(W+1)RP}{(WR)+P}$$

where W is the relative weight of recall to precision (usually set to 1)

Kappa statistics

The kappa coefficient (K) measures pairwise agreement among the annotators, correcting for expected chance agreement:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

where
$P(A)$ is the proportion of times that the coders agree
$P(E)$ is the proportion of times that we would expect them to agree by chance

$K = 0$ ... expected agreement (by chance) only
$K = 1$ ... total agreement

Evaluation

- DDs are indexed by the system (or annotator, respectively).

  `system:` `dd_class(135,anaphoric).`

  `coder:` `dd_class(135,anaphoric).`

- DDs are marked for coreference

  `coder:` `coref(135,106).`

  `system:` `coref(135,106).`

- **coreference chains** – equivalence classes of discourse entities are checked, whether they correspond to each other

- system coreference chains do not have to be complete

- the indices needn't be identical

Direct Anaphora Heuristics

assess the questions of identifying the antecedent corresponding to the respective direct DD and putting restrictions on such antecedents

Presented heuristics:

- Head noun identification
- Potential antecedents
- Segmentation
- Noun modifiers

Head noun identification

Simply select the rightmost noun of the NP.

But, this is not a sure-fire rule:

- headless NPs
  "the highest in the southern region"
- coordinated NPs
  "the reporters and editors"

Potential antecedents

All potential antecedents carry some information with them – index, NP structure, head noun and NP type (definite/indefinite/bare plural/possessive)

We have a choice how to define the class of all potential antecedents:

- indefinite NPs (indefinite sg., bare pl.)
- indefinite NPs $\cup$ definite descriptions (beginning with "the")
- indefinite NPs $\cup$ definite descriptions $\cup$ possessive NPs
- all NPs

We can get various trade-offs between precision and recall depending on the choice of the class.

Segmentation

The number of antecedents can be reduced by segmentation (ignoring or penalizing antecedents outside the current segment).

Window-based techniques represent a useful approximation (only antecedents within a fixed-size window are considered).

Two versions (both require the identity of head nouns):

- **loose segmentation** (a relaxed window-based technique)
  - antecedent within the established window *or*
  - antecedent itself is an subsequent mention *or*
  - antecedent and DD are identical NPs (not only heads)
- **recency** – indexing NPs by their heads (and considering only the last occurence for each index)

Noun modifiers

Identical heads ⇒ antecedent ???

a blue car ... the red car
the population ... the voting population

Taking **premodifiers** into consideration:

- the premodifiers are a subset of the antecedent's premodifiers
  BUT: *the pixie-like clarinetist ... the soft-spoken clarinetist*
- antecedent without premodifiers matches any definite with the same head
  a check ... the lost check

The rule for **postmodifiers**:
if both are postmodified, the postmodifications must be the same

Evaluation of Anaphora Resolution

Loose segmentaion/recency	*R*	*P*	*F*
Segmentation, 1-sentence window	**71.79%**	86.48%	78.45%
Segmentation, 4-sentence window	**76.92%**	82.75%	79.73%
Segmentation, 8-sentence window	**78.20%**	80.26%	79.22%
Recency, all sentences	**80.76%**	78.50%	79.62%
Strict segmentation			
1-sentence window	29.48%	**89.32%**	44.33%
4-sentence window	57.69%	**88.23%**	69.76%
8-sentence window	67.94%	**84.46%**	75.31%
Combined Heuristics			
4-sentences + recency	75.96%	87.77%	**81.44%**
8-sentences + recency	77.88%	84.96%	**81.27%**

Evaluation of Anaphora Resolution

Antecedents selections	*R*	*P*	*F*
Indefinites, DDs, and possessives	**75.96%**	**87.77%**	**81.44%**
All NPs	77.88%	86.17%	81.81%
Indefinites and DDs	73.39%	88.41%	80.21%
Indefinites only	12.17%	77.55%	21.05%
Premodification constraints			
1. Antecedent-set/DD-subset	69.87%	91.21%	79.12%
2. Antecedent-empty	55.12%	88.20%	67.85%
3. Antecedent-subset/DD-set	64.74%	88.59%	74.81%
1 & 2	**75.96%**	**87.77%**	**81.44%**
1 & 3	75.96%	87.13%	81.16%
None	78.52%	81.93%	80.19%

Resulting configuration

- four-sentence window
- combined loose segmentation and recency
- indefinites, DDs, and possessives as potential antecedents
- the premodification of the description must be contained in the premodification of the antecedent or when the antecedent has no premodifiers

Resulting configuration performance

Anaphora classification	#	+	−	R	P	F
Training data	492	368	60	75%	86%	80%
Test data	218	151	58	69%	72%	70%
Anaphora resolution	#	+	−	R	P	F
Training data	312	237	33	76%	88%	81%
Test data	154	96	19	62%	83%	71%

Main Problems

- data errors (typos)
- heuristics limit the search range - precision/recall
  (locality, antecedent type, ...)
- resolving pl. references to objects introduced as a number of sg. NPs

Heuristics for Discourse-New DDs

Discourse-new definite descriptions have usually (but not always) certain syntactic and/or lexical features ...

Presented heuristics:

- presence of special predicates
- restrictive modification
- appositives
- copular constructions
- "larger situation" definites

Presence of special predicates

- **unexplanatory modifiers** – turn the head noun into a function

  `superlatives` ∪

  ∪ {`first, last, best, most, maximum, minumum, only`}

  The first German to win a Grand Slam

- head noun has a semantic function and requires a complement

  {`fact, result, conclusion, idea, belief, saying, remark`}

  the fact that there is life on Earth

Restrictive (post)modification

- **Postmodification**

  75% of all complex phrases are first-mention

  (restrictive postmodification is the most frequent reason)

  – Relative clauses

    {`who, whom, which, where, when, why, that`}

    The pen you bought yesterday.

  – Nonfinite postmodifiers

    The man writing the letter is my friend.

    The man to consult is Wilson.

  – Prepositional phrases and of-clauses

    (PPs are the most common type of postmodification)

    the advantages of the new system

    there is a small grammar for this

Restrictive (pre)modification

- **Premodification** (not as common as postmodification)
  mainly **time modifiers** (dates, eras, etc.)
  the 1987 stock market crash

- **Note:** it is important to distinguish restrictive from
  non-restrictive postmodification
  The substance, discovered almost by accident, is very important.

  commas $\Rightarrow$ non-restrictive postmodification

Appositives and copulas

- **appositives**

  Maurice Ravel, the author of Bolero

  the Sandhills Luncheon Care, a tin building in midtown

- **copular constructions**

  with verb phrase heads of {`to be, to seem, to become`}

  [*The man most likely to win the Wimbledon*] *is Andre Aggassi.*

  What the investors object to most is [*the effect they say the proposal would have on their ability to spot telltale "clusters" of trading activity*].

  Note: the complement is an adjective ⇒

  ⇒ the subject is, in majority of cases, used referentially

"larger situation" definites

- with proper names in premodification
  the Iraq war
  (detection – "the" comes before a proper name)
- referring to time
  the morning
- other
  {`the sun, the moon, the sky, the pope, the weather`}

Evaluation of Discourse-new DD Heuristics

Heuristics tested on unseen data	Total found	Errors	Precision
Special predicates	16	2	87%
Appostions	10	2	80%
Copula	6	4	33%
Postmodifications	95	22	77%
Total	127	30	76%

Discourse-new DD identification	#	+	−	R	P	F
Training data	492	368	60	75%	86%	80%
Test data	218	151	58	69%	72%	70%

Drawbacks of discourse-new DD heuristics

- **Appositions** – coordinated NPs match as well
- **Copula** – copula can be also seen as bridging
- **Restrictive premodification** – name-premodifiers can be used in anaphoric constructions as well
- **Restrictive postmodification** – classifies certain cases as discourse-new inappropriately

Heuristics for resolving bridging descriptions

Consists of two subtasks – finding the anchor, and determining the link between the DD and the anchor.
A very complex task:

- the link can be arbitrarily complicated
- there may be several ("correct") anchors

$\Longrightarrow$ the biggest poser for a shallow parser

Heuristics for resolving bridging descriptions

Heuristics classification (based on information needed)

- based on lexical relations retrievable from WN
- antecedent = proper name, description = common noun
- the anchor is a (non-head) noun modifying the antecedent
- the antecedent is not introduced by an NP, but by a VP
  Kadane oil is currently drilling two oil wells. The activity ...
- antecedent is not explicitly mentioned in the text
  (it is implicitly available because it is a discourse topic)
- relation with the anchor is based on more general commonsense knowledge

Heuristics based on WN

Lexical relations retrievable from the WN are relatively well-defined – can be feasible to detect. Relations used for the heuristics:

- **synonymy** (nouns are in the same synset)
- **hyponymy/hyperonymy** *(dog – animal)*
- **meronymy/holonymy** *(hand – finger)*
- **coordinate sisters** (share the same hyperonym)
  (home – house)

Considering more complex relations did not bring convincing results. Moreover, the computational cost is usually very high.

Proper name antecedents

antecedent = proper name

definite description = common noun

Bach ... the composer

- named-entity recognition – detecting the name and determining its type
- WN lookup
- special backtracking component used to type underspecified names which are specified later in the text

Non-head noun anchors

We do not generally consider modifying nouns as potential referents. We only look for head-nouns.

But we would like to match also:

- head of a DD with the the premodifiers of the previous NP
  the company has been selling discount packages ... the discounts
  stock market crash ... the markets
- the premodifiers of a DD with premodifiers of its antecedents
  his art business ... the art gallery
- the premodifiers of a DD with the head of the previous NP
  a 15-acre plot and main home ... the home site

Results of bridging descriptions heuristics

Using a five-sentence window proved to be optimal for bridging.

WN bridging classification	Relations found	Right anchors	% Right
Synonymy	11	4	36%
Hyponymy	59	18	30%
Meronymy	6	2	33%
Sister	30	6	20%
Total	106	30	28%

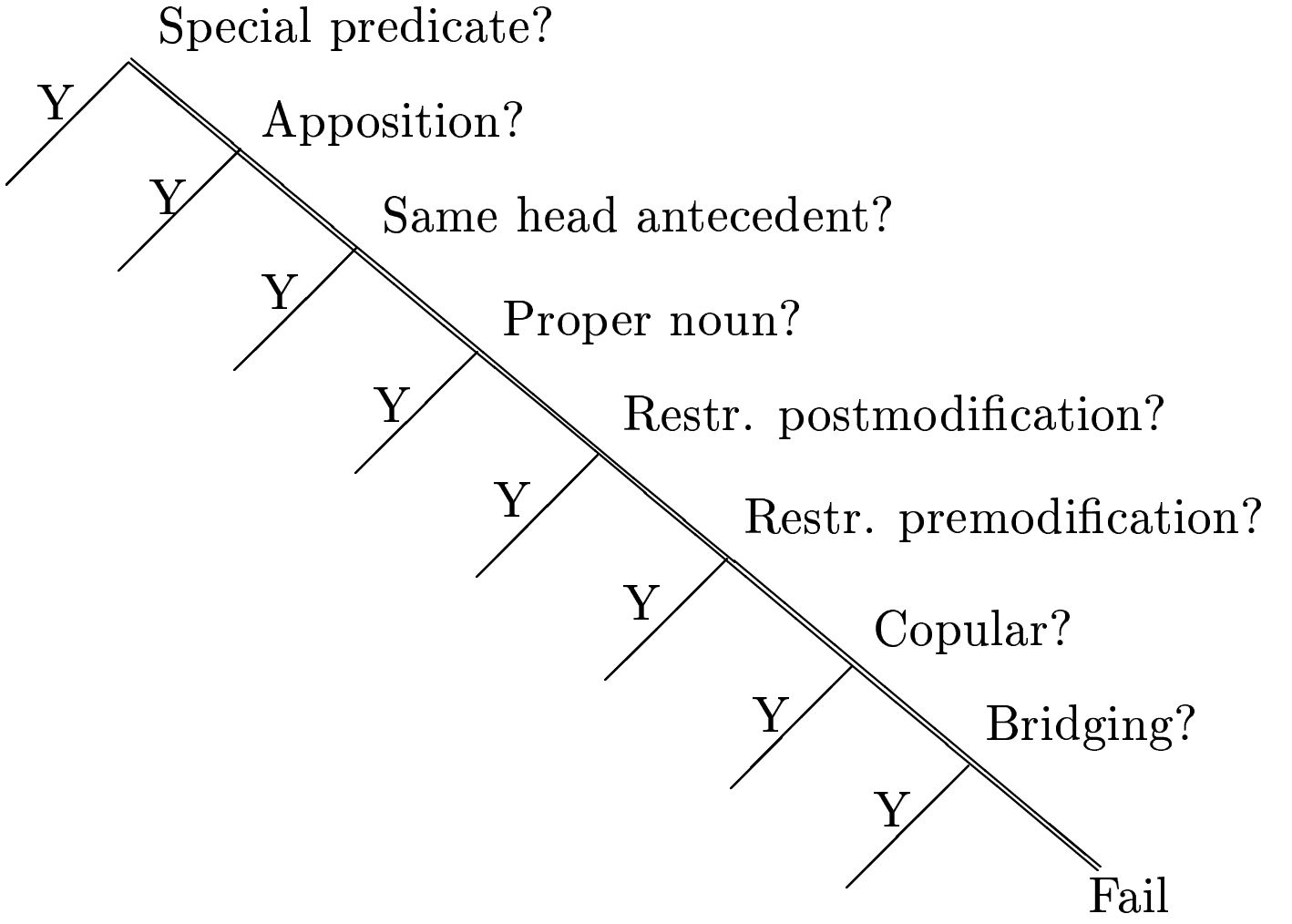
Weaknesses of bridging descriptions heuristics

- **WordNet**
  - ambiguity is a frequent source of errors
  - on the other hand, incompleteness of the relations and synsets in WN
  - typographic variations
- **Proper names** and their typing
  WordNet hasn't been designed for such a task
  (doesn't include many proper names ⇒ poor recall)
- **Bridging based on compound nouns**
  recall of only about 36% (the problem being segmentation)

The overall system architecture

- Based on a decision tree which decides which heuristics should be applied and in which order.
- The decision tree can be hand-crafted (using the method of trial-and-error) or can be obtained automatically.
- meta-classifier structure (strong heuristics first)
- Hand-crafted decision tree strategy:
  1. safe elimination of non-anaphoric cases
  2. same head antecedent matching
  3. discourse-new entities
  4. bridging (in Version 2 only)
     order: proper names – compound nouns – WordNet

Diagram of the hand-crafted system



Overall hand-crafted system performance

System performance	R	P	F
V.1 Overall	53%	76%	62%
V.2 Overall	57%	70%	62%

Agreement among the system and the annotators

Version 1 – The agreement is relatively high.
Comparable to the agreement among the three annotators.
⇒ differences mostly in cases where even people might disagree with each other

Version 2 – Worse results.
Better recall, but much worse precision ...

Deriving the Decision Tree Automatically

The training set was used as an input to the ID3 machine learning algorithm – which builds the desired decision-tree.
These features were used:

- Special Predicates
- Direct Anaphora
- Apposition
- Proper Noun
- Restrictive Postmodification

Other heuristics were excluded because their performance had not been good enough.

Performance of the machine learned algorithm

The results are very comparable – actually, differences between the two decision-trees were only marginal (the order of only several classifiers was changed).

Overall performance	R	P	F
2-class model	69%	69%	69%
3-class model	75%	75%	75%
hand-crafted algorithm	63%	86%	71.7%

Conclusions

- domain-independent
- based on empirical study uf DD usage
- some heuristics rely on syntax and semantic networks (WordNet)
- shallow parse architecture $\Rightarrow$ problems when commonsense reasoning needed

Building upon the 2000 system

- better corpora with more sophisticated annotation systems (MATE, GNOME)
- need to enlarge lexical and common sense knowledge to improve bridging resolution
- some syntactic structures and collocations might provide needed knowledge (meronymy etc.) ...
- possibility to extend WordNet (still under development)
- vector-based semantics – might give good results if fed with good data

Other ideas

- a robust parser would help
- more research about coreference chains needed
- focussing – no reliable focussing mechanism available to support the bridging resolution
- more facts about discourse-new descriptions

References

[1] Massimo Poesio. *Bridging Descriptions, Lexical Information, and Focusing.* CoLi, Saarbrücken 23.1.2003 Colloquim, The University of Essex, 2003.

[2] Massimo Poesio and Renata Vieira. *An Empirically Based System for Processing Definite Descriptions.* Computational Linguistics, 26(4), 2000.

[3] Piek Vossen. *Introduction to EuroWordNet.* University of Amsterdam, 1998.