Resolving Coreferent and Associative Noun Phrases in Scientific Text

Ina Rösiger\textsuperscript{1}  Simone Teufel\textsuperscript{2}

\textsuperscript{1}Institute for Natural Language Processing, University of Stuttgart
\textsuperscript{2}Computer Laboratory, University of Cambridge

Student Research Workshop, EACL 2014
Coreferent and Associative Anaphora

Coreferent anaphora

Peter went into the supermarket. He bought a pizza.

Associative (=bridging) anaphora

Tim’s bike needs to be repaired. The tire has a puncture.
Coreferent Anaphora in **Scientific Text**

computational linguistics

In this paper, a general method of maximizing top-down constraints is proposed. The method provides a procedure to dynamically compute ...
We base our experiments on the Penn treebank. The corpus size is ...

Xe-Ar was found to be in a layered structure with Ar on the surface.
Questions

- Theoretical:
  - How does scientific text differ from other genres with respect to the proportion of non-referring and referring entities?
    ⇒ investigate information status
Information Status

- Referring:
  - **textual**: the Penn Treebank ...
    - coreferent: The corpus
    - associative: The corpus size
  - *world/domain knowledge*:
    Noam Chomsky, the F1-measure
  - **deictic**:
    This paper (= the current paper)
Information Status

Noun phrase
- referring
- non-referring

Non-referring:
- new: a sampling method
- descriptions:
  - the fact that the accuracy rises,
  - the most efficient short interfering RNA
- idioms: on the one hand
Questions

- Theoretical:
  - How does scientific text differ from other genres with respect to the proportion of non-referring and referring entities?

- Practical:
  - How does a state-of-the-art coreference resolver perform on scientific text?
  - How can we adapt the resolver to
    - improve performance for coreferent references in the new domain;
    - also be able to resolve associative references?
Hypotheses for scientific text

- Information status
  - higher proportion of associative anaphora

- Coreference resolution can be improved by including
  - domain-dependent training data
  - terminology as a ML feature

- Resolution of associative references
  - using a coreference resolver is possible
  - can be improved by semantic features
Outline

- Research Question and Hypotheses
- Corpus Annotation and Corpus Study
- Adapting a Coreference Resolver ...
  - ... to the Scientific Domain
  - ... for Associative Links
- Experiments
  - Setup
  - Resolving Coreferent References
  - Resolving Associative References
- Conclusion and Future Work
Outline

- Research Question and Hypotheses
- Corpus Annotation and Corpus Study
  - Adapting a Coreference Resolver ...
    - ... to the Scientific Domain
    - ... for Associative Links
- Experiments
  - Setup
  - Resolving Coreferent References
  - Resolving Associative References
- Conclusion and Future Work
Corpus Annotation

- Scientific text
  - shows much variance across the scientific disciplines
    ⇒ we chose two exemplary disciplines
    - computational linguistics (CL): familiar domain
    - genetics (GEN): no particular knowledge
Corpus Annotation

- 8 papers for each of the two disciplines
- Manual annotation of both information status and anaphoric links
- Annotation scheme
  - mark all definite noun phrases with information status
  - antecedent has to be an NP: no event reference
  - coref + assoc links: the anaphor has to be definite, the antecedent can be any type of NP
## Corpus Annotation – Annotation Scheme

| Category                        | Example                                                                 |
|---------------------------------|-------------------------------------------------------------------------|
| **GIVEN (SPECIFIC)**            | We present the following experiment.                                    |
|                                 | It deals with ...                                                       |
| **GIVEN (GENERIC)**             | We use the Jaccard similarity coefficient in our experiments.           |
|                                 | The Jaccard similarity coefficient is useful for ...                   |
| **ASSOCIATIVE**                 | Xe-Ar was found to be in a layered structure with Ar on the surface.    |
| **ASSOCIATIVE SELF-CONTAINING** | The structure of the protein                                            |
| **DESCRIPTION**                 | The fact that the accuracy improves ...                                 |
| **UNUSED**                      | Noam Chomsky introduced the notion of ...                               |
| **DEICTIC**                     | This experiment deals with ...                                          |
| **PREDICATIVE**                 | Pepsin, the enzyme, ...                                                 |
| **IDIOM**                       | On the one hand ... on the other hand ...                               |
| **DOUBT**                       | (for unclear cases)                                                     |
## Corpus Annotation – Resulting Corpus

|                | GEN | CL | GEN+CL |
|----------------|-----|----|--------|
| Papers         | 8   | 8  | 16     |
| Sentences      | 1,834 | 1,637 | 3,471  |
| Words          | 43,691 | 38,794 | 82,485 |
| Annotated NPs  | 3,800 | 4,247 | 8,047  |
Corpus Study

- Question: How does scientific text differ from news text?
- Poesio and Vieira 1998:
  - Corpus study on definite descriptions in news text
  - Slight differences wrt their annotation scheme
### Distribution of categories in different domains, in percent

| Category                        | Scientific | News  |
|--------------------------------|------------|-------|
| Given                          | 50.32      | 44.00 |
| Associative                    | 7.44       | 8.50  |
| Associative self-containing    | 7.02       | –     |
| Description                    | 24.96      | 21.30 |
| Unused                         | 7.13       | 23.50 |
| Deictic                        | 1.01       | –     |
| Predicative                    | 0.35       | –     |
| Idiom                          | 0.52       | (2.00)|
| Doubt                          | 0.78       | 2.60  |
Question

How does scientific text differ from news text?

Results

- **Unused**: prevalence of named entities in news text, much reference to time and locations
  - near the San Francisco airport
  - at the East Coast
- **Given**: higher proportion in scientific text
- **Associative**: higher proportion in scientific text
Outline

■ Research Question and Hypotheses
■ Corpus Annotation and Study
■ Adapting a Coreference Resolver ...
  ▪ ... to the Scientific Domain
  ▪ ... for Associative Links
■ Experiments
  ▪ Setup
  ▪ Resolving Coreferent References
  ▪ Resolving Associative References
■ Conclusion and Future Work
Use of Domain Terms as Features

- Idea: capture domain knowledge
- Method: analyse 1000 CL and 1000 GEN papers, extract terminology
- 3 Features:
  - non-coreferring bias list
  - (collocation list)
  - (argumentation and work nouns, idioms)
Non-coreferring Bias List

- Which lexemes do not typically occur in coreference chains in the scientific domain?
Non-coreferring Bias List

- Normal case of coreference: introduce a concept as an indefinite NP, then refer to it as a definite NP
- Compile a list of lexemes that do not follow this pattern
  ⇒ These lexemes are biased to be non-coreferring, i.e. of category UNUSED or DESCRIPTION
- Two versions:
  - global list (entire document collection)
  - local list (at least one document)
## Non-coreferring Bias List

|       | CL (a) global | CL (b) local | GEN (a) global | GEN (b) local |
|-------|---------------|--------------|----------------|--------------|
| presence       | number        | manuscript   | data           |
| beginning      | word          | respect      | region         |
| literature     | sentence      | prediction   | gene           |
| hearer         | training      | monograph    | case           |
| reader         | user          | notion       | species        |
Outline

- Research Question and Hypotheses
- Corpus Annotation and Corpus Study
- Adapting a Coreference Resolver ...
  - ... to the Scientific Domain
  - ... for Associative Links
- Experiments
  - Setup
  - Resolving Coreferent References
  - Resolving Associative References
- Conclusion and Future Work
### Example Types of Associative Anaphora

- **Same head, different modifier**
  
  the negative strain ... the positive strain;

- **Semantic knowledge required**
  
  the classifier ... the training data;

- **Derivationally related words**
  
  the spotty distribution ... the spots.
Semantic Features

- Use of the following WN relations:
  - hypernymy (*macromolecule* → *protein*)
  - hyponomy (*nucleoprotein* → *protein*)
  - meronymy (*surface* → *structure*)
  - substance meronymy (*amino acid* → *protein*)
  - topic member (*acute, chronic* → *medicine*)
  - topic (*periodic table* → *chemistry*)

- WordNet’s coverage in the scientific domain:
  75.91 % of all common nouns in GEN papers and
  88.12 % in CL
Outline

- Research Question and Hypotheses
- Corpus Annotation and Study
- Adapting a Coreference Resolver ...
  - ... to the Scientific Domain
  - ... for Associative Links
- Experiments
  - Setup
    - Resolving Coreferent References
    - Resolving Associative References
- Conclusion and Future Work
Experimental Setup

- IMS coreference resolver [Björkelund and Farkas 2012]
  trained on the OntoNotes dataset from the CoNLL 2011 shared task (mostly non-written text, dialogue)
  ⇒ out-of-the-box system
- Performance is evaluated on our annotated corpus
- 8-fold cross validation
- official CoNLL evaluation metric
Outline

- Research Question and Hypotheses
- Corpus Annotation and Study
- Adapting a Coreference Resolver ...
  - ... to the Scientific Domain
  - ... for Associative Links
- Experiments
  - Setup
  - Resolving Coreferent References
  - Resolving Associative References
- Conclusion and Future Work
Resolving Coreferent References

- **Baseline**: out-of-the-box system, trained on CoNLL shared task OntoNotes dataset 2011 (17,865,211 words)
- **Scientific**: trained on our scientific corpus (82,485 words)
- **Mixed**: trained on both

| Training Set  | GEN | CL  | GEN+CL |
|---------------|-----|-----|--------|
| **Baseline**  | 35.30 | 40.30 | 37.80 |
| **Scientific**| 44.94 | 42.41 | 43.68 |
| **Mixed**     | 47.92 | 47.44 | 47.68 |

- Comparison: 61.24 reported for the out-of-the-box system on the OntoNotes shared task dataset
Adding Terminological Features

|                  | GEN  | CL   | GEN+CL |
|------------------|------|------|--------|
| **Mixed**        | 47.92| 47.44| 47.68  |
| + Bias Lists     | 48.04| 47.79| 47.94  |
| + Arg. Nouns and Idioms | 48.14| 48.52| 48.33  |
| + Collocations   | 48.03| 48.12| 48.08  |

- Small gain, yet not significant
Resolving Coreferent References: Summary

- Training on our small scientific corpus better than large out-of-domain training data (OntoNotes)
- Effect of domain-dependent training data larger than that of terminological features
- Final system outperforms the out-of-the-box system by 36.47 % for GEN and 20.40 % for CL

|       | GEN  | CL  | GEN+CL |
|-------|------|-----|--------|
| Baseline | 35.30 | 40.30 | 37.80  |
| Final   | 48.14 | 48.52 | 48.33  |
Outline

- Research Question and Hypotheses
- Corpus Annotation and Study
- Adapting a Coreference Resolver ...
  - ... to the Scientific Domain
  - ... for Associative Links
- Experiments
  - Setup
  - Resolving Coreferent References
  - Resolving Associative References
- Conclusion and Future Work
Resolving Associative References

- best (final) version: statistically significant
  - GEN: synonyms, hyponyms, hypernyms
  - CL: synonyms, hyponyms, topics and meronyms

|                | GEN     | CL      | GEN+CL  |
|----------------|---------|---------|---------|
| **Baseline**   | 33.52   | 32.86   | 33.19   |
| **Final version** | **35.54** | **34.21** | **34.88** |

(domain-dependent)
Outline

- Research Question and Hypotheses
- Corpus Annotation and Study
- Adapting a Coreference Resolver ...
  - ... to the Scientific Domain
  - ... for Associative Links
- Experiments
  - Setup
  - Resolving Coreferent References
  - Resolving Associative References
- Conclusion and Future Work
Results

- Information status:
  - corpus of over 8000 annotated definite noun phrases
  - marked difference between scientific and news text

- Coreference resolution
  - domain adaptation was successful
    ⇒ outperforms state-of-the-art system by about 28%
      - domain-dependent training data
      - terminological features

- Resolution of associative anaphora
  - possible to resolve them with a coreference resolver
  - effect of semantic features
Ongoing Work

- Inter-annotator agreement study:
  - manual agreement
  - annotated data will be made available
- Improve and add new features
- Train and test on other scientific disciplines
- Train a classifier to predict information status
Thanks for your attention!
References

Björkelund and Farkas, 2012
*Data-driven multilingual coreference resolution using resolver stacking*

Poesio and Vieira, 1998
*A corpus-based investigation of definite description use*
Distribution of categories in different domains, in percent

| Category                           | GEN  | CL  | news |
|------------------------------------|------|-----|------|
| Given                              | 52.03| 48.60| 44.00|
| Associative                        | 8.29 | 6.59 | 8.50 |
| Associative self-containing        | 7.63 | 6.40 | –    |
| Description                        | 21.31| 28.61| 21.30|
| Unused                             | 7.53 | 6.73 | 23.50|
| Deictic                            | 0.74 | 1.27 | –    |
| Predicative                        | 0.24 | 0.45 | –    |
| Idiom                              | 0.24 | 0.80 | (2.00)|
| Doubt                              | 1.03 | 0.52 | 2.60 |
Error Analysis

- Domain-dependent errors:
  - missing links due to lack of domain-dependent knowledge

Experiments were performed with the clean part of the treebank. The corpus consists of 1 million words.

Amino acids are organic compounds made from amine (-NH2) and carboxylic acid (-COOH) functional groups. The protein-forming molecules...
Error Analysis

- Domain-dependent errors:
  - effects of non-written training data:
    all occurrences of *e.g.* put into one coreference chain
  - demonstrative, science-specific expressions:
    links every occurrence of *this approach*
  - same sentences used in abstract and in introduction:
    large spans interpreted as coreferent
Error Analysis

- General errors, mostly wrong links

This approach allows the processes of building referring expressions and identifying their referents.

The issue of how to design siRNAs that produce high efficacy is the focus of a lot of current research. Since it was discovered that ...
Collocation list

- Hypothesis: NPs in verb-object collocations are not co-referring
- Extraction of domain-specific collocations using a term extraction tool [Gojun et al. 2012]

| make + use     | take + place       |
|----------------|--------------------|
| give + rise    | silence + activity |
| derive + form  | refashion + plan   |
| parse + sentence | predict + sirna    |
| sort + feature | match + predicate  |
| see + figure   | use + information  |
| silence + efficiency | follow + transfection |
| embed + sentence | apply + rule      |
| focus + algorithm | stack + symbol    |
Argumentation and work nouns, idiom

- Argumentation and work nouns:  
  achievement, claim, experiment, ...

- Idioms:
  on the one hand, ... on the other hand

[Teufel 2010]