# Resolving Coreferent and Associative Noun Phrases in Scientific Text

**Ina Rösiger**
Institute for Natural Language Processing
University of Stuttgart, Germany
Pfaffenwaldring 5b, 70569 Stuttgart
roesigia@ims.uni-stuttgart.de

**Simone Teufel**
Computer Laboratory
University of Cambridge, UK
15 JJ Thomson Avenue, Cambridge CB3 0FD
sht25@cl.cam.ac.uk

## Abstract

We present a study of information status in scientific text as well as ongoing work on the resolution of coreferent and associative anaphora in two different scientific disciplines, namely computational linguistics and genetics. We present an annotated corpus of over 8000 definite descriptions in scientific articles. To adapt a state-of-the-art coreference resolver to the new domain, we develop features aimed at modelling technical terminology and integrate these into the coreference resolver. Our results indicate that this integration, combined with domain-dependent training data, can outperform the performance of an out-of-the-box coreference resolver. For the (much harder) task of resolving associative anaphora, our preliminary results show the need for and the effect of semantic features.

## 1 Introduction

Resolving anaphoric relations automatically requires annotated data for training and testing. Anaphora and coreference resolution systems have been tested and evaluated on different genres, mainly news articles and dialogue. However, for scientific text, annotated data are scarce and coreference resolution systems are lacking (Schäfer et al., 2012). We present a study of anaphora in scientific literature and show the difficulties that arise when resolving coreferent and associative entities in two different scientific disciplines, namely computational linguistics and genetics.

Coreference resolution in scientific articles is considered difficult due to the high proportion of definite descriptions (Watson et al., 2003), which typically require domain knowledge to be resolved. The more complex nature of the texts is also reflected in the heavy use of abstract entities such as results or variables, while easy-to-resolve named entities are less frequently used. We test an existing, state-of-the-art coreference resolution tool on scientific text, a domain on which it has not been trained, and adapt it to this new domain. We also address the resolution of *associative anaphora* (Clark, 1975; Prince, 1981), a related phenomenon, which is also called bridging anaphora. The interpretation of an associative anaphor is based on the associated antecedent, but the two are not coreferent. Examples 1 and 2 show two science-specific cases of associative anaphora from our data.

(1) Xe-Ar was found to be in *a layered structure* with Ar on **the surface**[1].

(2) We base our experiments on *the Penn treebank*. **The corpus size** is ...

The resolution of associative links is important because it can help in tasks which use the concept of textual coherence, e.g. Barzilay and Lapata (2008)'s entity grid or Hearst (1994)'s text segmentation. They might also be of use in higher-level text understanding tasks such as textual entailment (Mirkin et al., 2010) or summarisation based on argument overlap (Kintsch and van Dijk, 1978; Fang and Teufel, 2014).

Gasperin (2009) showed that biological texts differ considerably from other text genres, such as news text or dialogue. In this respect, our results confirm that the proportion between non-referring and referring entities in scientific text differs from that reported for other genres. The same holds for the type and relative number of linguistic expressions used for reference. To address this issue, we decided to investigate *information status* (Nissim et al., 2004) of noun phrases. Information status tells us whether a noun phrase refers to an already

[1] Anaphors are typed in bold face, their antecedents shown in italics.

*Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 45–55,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

known entity, or whether it can be treated as non-referring. Since no corpus of full-text scientific articles annotated with both information status and anaphoric relations was available, we had to create and annotate our own corpus. The main contributions of this work are (i) a new information status-based annotation scheme and an annotated corpus of scientific articles, (ii) a study of information status in scientific text that compares the distribution of the different categories in scientific text with the distribution in news text, as well as between the two scientific disciplines, (iii) experiments on the resolution of coreferent anaphora: we devise domain adaptation for science and show how this improves an out-of-the-box coreference resolver, and (iv) experiments on the resolution of associative anaphora with a coreference resolver that is adapted to this new notion of "reference" by including semantic features. To the best of our knowledge, this is the first work on anaphora resolution in multi-discipline, full-text scientific papers that also deals with associative anaphora.

## 2 Related Work

*Noun phrase coreference resolution* is the task of determining which noun phrases (NPs) in a text or dialogue refer to the same real-world entities (Ng, 2010). Resolving anaphora in scientific text has only recently gained interest in the research community and focuses mostly on the biomedical domain (Gasperin, 2009; Batista-Navarro and Ananiadou, 2011; Cohen et al., 2010). Some work has been done for other disciplines, such as computational linguistics. Schäfer et al. (2012) present a large corpus of 266 full-text computational linguistics papers from the ACL Anthology, annotated with coreference links. The CoNLL shared task 2012 on modelling multilingual unrestricted coreference in OntoNotes (Pradhan et al., 2012) produced several state-of-the-art coreference systems (Fernandes et al., 2012; Björkelund and Farkas, 2012; Chen and Ng, 2012) trained on news text and dialogue, as provided in the OntoNotes corpus (Hovy et al., 2006). Other state-of-the-art systems, such as Raghunathan et al. (2010) and Berkeley's Coreference Resolution System (Durrett and Klein, 2013), also treat coreference as a task on news text and dialogue. We base our experiments on the IMS coreference resolver by Björkelund and Farkas (2012), one of the best publicly available English coreference systems. The resolver uses the decision of a cluster-based decoding algorithm, i.e. one that decides whether two mentions are placed in the same or in different clusters, or whether they should be considered singletons. Their novel idea is that the decision of this algorithm is encoded as a feature and fed to a pairwise classifier, which makes decisions about pairs of mentions rather than clusters. This stacked approach overcomes problems of previous systems that are based on the isolated pairwise decision. The features used are mostly taken from previous work on coreference resolution and encode a variety of information, i.e, surface forms and their POS tags, subcategorisation frames and paths in the syntax tree as well as the semantic distance between the surface forms (e.g. edit distance).

However, none of this work is concerned with associative anaphora. Hou et al. (2013) present a corpus of news text annotated with associative links that are not limited with respect to semantic relations between anaphor and antecedent. Their experiments focus on antecedent selection only, assuming that the recognition of associative entities has already been performed. Information status has been investigated extensively in different genres such as news text, e.g. in Markert et al. (2012). Poesio and Vieira (1998) performed an information status-based corpus study on news text, defining the following categories: coreferential, bridging, larger situation, unfamiliar and doubt. To the best of our knowledge, there is currently no study on information status in scientific text.

In this paper, we propose a classification scheme for scientific text that is derived from Riester et al. (2010) and Poesio and Vieira (1998). We investigate the differences between news text and scientific text by analysing the distribution of information status categories. We hypothesise that the proportion of associative anaphora in scientific text is higher than in news text, making it necessary to resolve them in some form. Our experiments on the resolution of coreferent anaphora concern the domain-adaptation of a coreference resolver to this new domain and examine the effect of domain-dependent training data and features aimed at capturing technical terminology. We also present an unusual setup where we assume that an existing coreference resolver can also be used to identify associative links. We integrate semantic features in the hope of detecting cases where domain knowledge is required to establish the relation between the anaphor and the antecedent.

|  | Category | Example |
|---|---|---|
| COREFERENCE LINKS | GIVEN (SPECIFIC) | We present *the following experiment.* **It** deals with ... |
|  | GIVEN (GENERIC) | We use *the Jaccard similarity coefficient* in our experiments.<br>**The Jaccard similarity coefficient** is useful for ... |
| ASSOCIATIVE LINKS | ASSOCIATIVE | Xe-Ar was found to be in *a layered structure* with Ar<br>on **the surface**. |
| Categories without links | 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 |  |

Table 1: Categories in our classification scheme

## 3 Corpus Creation

We manually annotated a small scientific corpus to provide a training and test corpus for our experiments, using the annotation tool Slate (Kaplan et al., 2012).

### 3.1 Annotation Scheme

Two types of reference are annotated, namely COREFERENCE and ASSOCIATIVE LINKS. COREFERENCE LINKS are annotated for all types of nominal phrases; such links are annotated between enitites that refer to the same referent in the real world. ASSOCIATIVE LINKS and information status categories are only annotated for definite noun phrases. In our scheme, ASSOCIATIVE LINKS are only annotated when there is a clear relation between the two entities. As we do not pre-define possible associative relations, this definition is vague, but it is necessary to keep the task as general as possible. Additionally, we distinguish the following nine categories, as shown in Table 1[2]: The category GIVEN comprises coreferent entities that refer back to an already introduced entity. If a coreference link is detected, the referring expression is marked as GIVEN and the link with its referent NP is annotated. The obligatory attribute GENERIC tells us whether the given entity has a generic or a specific reading. ASSOCIATIVE refers to entities that are not coreferent but whose interpretation is based on a previously introduced entity. A typical relation between the two noun phrases is meronymy, but as mentioned above we do not pre-define a set of allowed semantic relations. The category ASSOCIATIVE (SELF-CONTAINING) comprises cases where we identify an associative relation between the head noun phrase and the modifier. ASSOCIATIVE SELF-CONTAINING entities are annotated without a link between the two parts. In scientific text, an entity is considered DEICTIC if it points to an object that is connected to the current text. Therefore, we include all entities that refer to the current paper (or aspects thereof) in this category. Entities that have not been mentioned before and are not related to any other entity in the text, but can be interpreted because they are part of the common knowledge of the writer and the reader are covered by the category UNUSED. DESCRIPTION is annotated for entities that are self-explanatory and typically occur in particular syntactic patterns such as NP complements or relative clauses. Idiomatic expressions or metaphoric use are covered in the category IDIOM. Predicative expressions, including appositions, are annotated as PREDICATIVE. Finally, the category DOUBT is used when the text or the antecedent is unclear. Note that NEW, a category that has been part of most previous classification schemes of information status, is not present as this information status is typically observed in indefinite noun phrases. As we deal exclusively with definite noun phrases[3], we do not include this category in our scheme. In contrast to Poesio and Vieira's scheme, ours contains the additional categories PREDICATIVE, ASSOCIATIVE SELF-CONTAINING, DEICTIC and IDIOM.

[2] The entity being classified is typed in bold face, referring expressions are marked by a box and the referent is shown in italics.

[3] With the exception of coreferring anaphoric expressions, as previously discussed.

| | GEN | CL |
|---|---|---|
| Sentences | 1834 | 1637 |
| Words | 43691 | 38794 |
| Def. descriptions | 3800 | 4247 |

Table 2: Properties of the annotated two subcorpora, genetics (GEN) and computational linguistics (CL)

| | GEN | CL |
|---|---|---|
| Coreference links | 1976 | 2043 |
| Associative links | 328 | 324 |
| Given | 1977 | 2064 |
| Associative | 315 | 280 |
| Associative (sc) | 290 | 272 |
| Description | 810 | 1215 |
| Unused | 286 | 286 |
| Deictic | 28 | 54 |
| Predicative | 9 | 19 |
| Idiom | 9 | 34 |
| Doubt | 39 | 22 |

Table 3: Distribution of information status categories and links in the two disciplines, in absolute numbers

### 3.2 Resulting Corpus

Our annotated corpus contains 16 full-text scientific papers, 8 papers for each of the two disciplines. The computational linguistics (CL) papers cover various topics ranging from dialogue systems to machine translation; the genetics (GEN) papers deal mostly with the topic of short interfering RNAs, but focus on different aspects of it. In total, the annotated computational linguistics papers contain 1637 sentences, 38,794 words and 4247 annotated definite descriptions while the annotated genetics papers contain 1834 sentences, 43,691 words and 3800 definite descriptions; the two domain subcorpora are thus fairly comparable in size. See Table 2 for corpus statistics and Table 3 for the distribution of categories and links.

It is well-known that there are large differences in reference phenomena between scientific text and other domains (Gasperin, 2009). In scientific text, it is assumed that the reader has a relatively high level of background. We would expect this general property of scientific text to have an impact on the distribution of categories with respect to information status.

Table 4 compares the two scientific disciplines in our study with each other. We note that the proportion of entities classified as DESCRIPTION in the CL papers is considerably higher than in the GEN papers. The proportions of the other categories are similar, though the proportion of GIVEN, ASSOCIATIVE and UNUSED entities is slightly higher in the GEN articles.

Table 4 also compares the distribution of categories in news text (Poesio and Vieira, 1998; P&V) with that of ours (as far as they are alignable, using our names for categories). Note that on a conceptual level, these categories are equivalent, but there are some differences with respect to the annotation guidelines.

The most apparent difference is the proportion of UNUSED entities (6-7 % in science, 23 % in news text) which might be due to the prevalence of named entities in news text. Compared to the distribution of categories in news text, the proportion of GIVEN entities is about 4-8 % higher in scientific text. The proportion of ASSOCIATIVE entities[4] is twice as high in the scientific domain compared to news text. UNUSED entities have a distinctly lower proportion, with about 7%. As our guidelines limit deictic references to only those that refer to (parts of) the current paper, we get a slightly lower proportion than the 2 % in news text, reported by Poesio and Vieira (1998) in an earlier experiment, where no such limitation was present.

| Category | GEN | CL | P&V |
|---|---|---|---|
| Given | 52.03 | 48.60 | 44.00 |
| Associative | 8.29 | 6.59 | 8.50 |
| Associative (sc) | 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 |

Table 4: Distribution of information status categories in different domains, in percent

It has been shown in similar annotation experiments on information status, with similarly fine-grained schemes (Markert et al., 2012; Riester et al., 2010), that it is possible to achieve annotation with marginally to highly reliable inter-annotator agreement. In our experiments, only one person (the first author) performed the annotation, so that we cannot compute any agreement measurements. We are currently performing an inter-annotator study with two additional annotators so that we can better judge human agreement and use the annotations as a reliable gold standard.

[4] The union of categories ASSOCIATIVE and ASSOCIATIVE SELF-CONTAINING.

## 4 Adapting a Coreference Resolver to the Scientific Domain

To show the difficulties that a coreference resolver faces in the scientific domain, we ran, out-of-the-box, a coreference system (Björkelund and Farkas, 2012), that has not been trained on scientific text, on our corpus and perform an error analysis. In particular, we are curious about which of the system's errors are domain-dependent. This analysis motivates a set of terminological features that are incorporated and tested in Section 6.

### 4.1 Error Analysis

**Domain-dependent errors.** The lack of semantic, domain-dependent knowledge results in the system's failure to identify coreferent expressions, e.g. those expressed as synonyms. This type of error can be prevented by implementing domain-dependent knowledge. In Example 3, we would like to generate a link between *treebank* and *corpus* as these terms are used as synonyms. The same is true for *protein-forming molecules* and *amino acids* in Example 4.

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

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

Another common error is that the coreference resolver links all occurrences of demonstrative science-specific expressions such as *this paper* or *this approach* to each other, even if they are several paragraphs apart. In most cases, these demonstrative expressions do not corefer, but refer to an approach or a paper recently described or cited. This type of error is particularly frequent in the computational linguistics domain and might be reduced by a feature that captures this peculiarity. A special case occurs when authors re-use clauses of the abstract in the introduction. The coreference resolver then interprets rather large spans as coreferent which are not annotated in the gold standard. Yet a different kind of error is based on the fact that the coreference resolver has been trained on OntoNotes, i.e. mostly on non-written text. Thus, the classifier has not seen certain phenomena and, for example, links all occurrences of *e.g.* into one equivalence class as it is interpreted as a named entity.

**General errors**. Some errors are general errors of coreference resolvers in the sense that they have very little to do with domain dependence, such as choosing the wrong antecedent or linking non-referential occurrences of *it* (see Examples 5 and 6).

(5) This approach allows *the processes of building referring expressions* and identifying **their** referents.

(6) *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 ...

### 4.2 Terminological Features

This section deals with the design of possible terminological features for our experiments that are aimed at capturing some form of domain knowledge. We create these using the information in 1000 computational linguistics and 1000 genetics papers that are not part of our scientific corpus.

**Non-coreferring bias list.** Our first feature concentrates on nouns which have a low probability to be coreferring (i.e. category GIVEN) if they appear as the head of noun phrase. We assume that the normal case of coreference between definite noun phrases is that of a concept introduced as an indefinite NP and later referred to as a definite NP, and compile a list of lexemes that do not follow this pattern. NPs with those lexemes should be more likely to be of category UNUSED or DESCRIPTION. We find the lexemes by recording head nouns of definite NPs which are not observed in a prior indefinite NP in the same document (local list) or the entire document collection (global list). We create two lists of such head words for every discipline. The lexemes are arranged in decreasing order of their frequency so that we can use both their presence or non-presence on the list and their rank on the list as potential features.

As can be seen in Table 5, *the presence, the beginning* and *the literature* are definite descriptions that are always used without having been introduced to the discourse. These terms are either part of domain knowledge (*the hearer, the reader*) or part of the general scientific terminology (*the literature*). In the local list we see expressions that can be used without having been introduced, but may in some contexts occur in the indefinite form as well, e.g. *the word* or *the sentence*.

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

Table 5: Top five terms of local and global non-coreferring bias lists

**Collocation list.** One of our hypotheses is that the NPs occurring in verb-object collocations are typically not part of any coreference chain. To test this, we use our collection of 2000 scientific papers to extract domain-specific verb-object collocations. We assume that for some collocations, this tendency is stronger (*make use, take place*) than for others that could potentially be coreferring (*see figure, apply rule*). The collocations have been identified with a term extraction tool (Gojun et al., 2012). Every collocation that occurs at least twice in the data is present on the list. Table 6 shows the most frequent terms.

| | |
|---|---|
| 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 |

Table 6: Most frequent occurring collocation candidates in scientific text (unsorted)

**Argumentation nouns, work nouns and idioms.** As mentioned in Section 4.1, the baseline classifier often links demonstrative, science-specific entities, even if they are several paragraphs apart. To prevent this, we combine a distance measure with a set of 182 argumentation and work nouns taken from Teufel (2010), such as *achievement*, *claim* or *experiment*. We also create a small list of idioms as they are never part of a coreference chain.

## 5 Adapting a Coreference Resolver for Associative Links in Science

We now turn to the much harder task of resolving associative anaphora.

### 5.1 Types of Associative Anaphora

To illustrate the different types of associative anaphora, we here show a few examples, mostly taken from the genetics papers. The anaphors are shown in bold face, the antecedents in italics. Many associative anaphors include noun phrases with the same head. In most of these cases, the anaphor contains a different modifier than the antecedent, such as

(8) the negative strain ... **the positive strain**;
(9) three categories ... **the first category**;
(10) siRNAs ... **the most effective siRNAs**.

We assume that these associative relations can be identified with a coreference resolver without adding additional features. Other cases are much harder to identify automatically, such as those where semantic knowledge is required to interpret the relation between the entities:

(11) the classifier ... **the training data**;
(12) this database ... **the large dataset**.

In other cases, the nominal phrase in the antecedent tends to be derivationally related to the head word in the anaphor, as in

(13) the spotty distribution ...**the spots**;
(14) competitor ... **the competitive effect**.

There are also a number of special cases, such as

(15) the one interference process ... **the other interference process**.

We hypothesise that the integration of semantic features discussed in the previous section enables the resolver to cover more than just those cases that are based on the similarity of word forms.

### 5.2 Semantic Features

It is apparent that the recognition and correct resolution of associative anaphora requires semantic knowledge. Therefore, we adapt the coreference resolver by extending the WordNet feature, one of the features implemented in the IMS resolver, to capture more than just synonyms.

We use the following WordNet relations: Hypernymy (*macromolecule* → *protein*), hyponymy (*nucleoprotein* → *protein*), meronymy (*surface* → *structure*), substance meronymy (*amino acid* → *protein*), topic member (*acute, chronic* → *medicine*) and topic ( *periodic table* → *chemistry*).

WordNet's coverage in the scientific domain is surprisingly good: 75,91 % of all common nouns in the GEN papers and 88,12 % in the CL papers are listed in WordNet. Terms that are not covered are, for example, abbreviations of different types of ribonucleic acid in genetics or specialist terms like *tagging, subdialogue* or *SVM* in computational linguistics.

## 6 Experiments

We now compare the performance of an out-of-the-box coreference system with the resolver trained on our annotated scientific corpus (Section 6.2). We also show the effect of adding additional features aimed at capturing technical terminology. In the experiments on the resolution of associative anaphora (Section 6.3), we test the hypothesis that the coreference resolver is able to adjust to the new notion of reference and show the effect of semantic features.

### 6.1 Experimental Setup

We perform our experiments using the IMS coreference resolver as a state-of-the-art coreference resolution system (Björkelund and Farkas, 2012)[5]. The algorithm and the features included have not been changed except where otherwise stated. We use the OntoNotes datasets from the CoNLL 2011 shared task[6] (Pradhan et al., 2012; Hovy et al., 2006), only for training the out-of-the-box system. We also use WordNet version 3.0 as provided in the 2012 shared task[7] as well as JAWS, the Java API for WordNet searching[8]. Performance is reported on our annotated corpus, using 8-fold cross-validation and the official CoNLL scorer (version 5).

### 6.2 Resolving Coreferent References

**IMS coreference resolver unchanged.** To be able to judge the performance of an existing coreference resolver on scientific text, we first report performance without making any changes to the resolver whatsoever, using different training data. The BASELINE version is trained on the OntoNotes dataset from the CoNLL 2011 shared task. In the SCIENTIFIC version, we only use our annotated scientific papers. MIXED contains the entire OntoNotes dataset as well as the scientific papers, leading to a larger training corpus which compensates for the rather small size of the scientific corpus[9]. Table 7 shows the average CoNLL scores[10] of the two subdomains genetics and computational linguistics.

| 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 |

Table 7: Resolving coreferent references: CoNLL metric scores for different training sets

The BASELINE achieves relatively low results in comparison to the score of 61.24 that was reported in the shared task (Björkelund and Farkas, 2012). Even though our scientific corpus is only 7% the size of the OntoNotes dataset, it inceases performance of the BASELINE system by 15,6%. The SCIENTIFIC version outperforms the BASELINE version for all of the GEN papers and for 6 out of 8 CL papers. MIXED, the version that combines the scientific corpus with the entire OntoNotes dataset, proves to work best (47.92 for GEN and 47.44 for CL). In THE BASELINE version, the performance on the CL papers is better than on the GEN papers. Interestingly, this is not true for the SCIENTIFIC version, where the performance on the GEN papers is better. However, as the main result here, we can see that training on scientific text was successful. The increase in score in both the SCIENTIFIC and the MIXED version over BASELINE is statistically significant[11]

[5] See: www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/IMSCoref.html
We follow their strategy to use the AMP decoder as the first decoder and the PCF decoder, a pairwise decoder, as a second. The probability threshold is set to 0.5 for the first and 0.65 for the second decoder.

[6] http://conll.cemantix.org/2011/data.html

[7] http://conll.cemantix.org/2012/data.html

[8] http://lyle.smu.edu/~tspell/jaws/

[9] We also experimented with a balanced version, which contains an equal amount of sentences from the OntoNotes corpus and our scientific corpus. The results are not reported here as this version performed worse.

[10] The CoNLL score is the arithmetic mean of MUC, $B^3$ and CEAFE.

[11] We compute significance using the Wilcoxon signed rank test (Siegel and Castellan, 1988) at the 0.01 level unless otherwise stated.

(+9.64 and +12.62 in the GEN papers, +2.11 and +7.14 in the CL papers, absolute in CoNLL score).

**IMS coreference resolver, adapted to the domain.** We show the results from the expansion of the feature set in Table 8. Each of the single features is added to the version in the line above the current version. Compared to the MIXED version, adding the features to the resolver results in an increase in performance for both of the scientific disciplines. However, when adding the collocation feature to the version including the bias lists, the argumentation nouns as well as idioms, performance drops slightly. This might indicate the need for a revised collocation list where those nouns are filtered out that could potentially be coreferring, e.g. *see figure*. For the best version of the CL papers, the increase in CoNLL score, compared with the MIXED version, is +1.08; for the GEN papers it is slightly less, namely +0.22. This increase in score is promising, but the data is too small to show significance.

| | 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 |

Table 8: Resolving coreferent references: CoNLL scores for the extended feature sets

However, compared with the BASELINE version, the final version (marked bold) performs significantly better and outperforms the out-of-the-box run by 36.47 % absolute on the CoNLL metric for the GEN papers and by 20.40 % for the CL papers. The results also show that, in our experiments, the effect of using domain-specific training material is larger than the effect of adding terminological features.

### 6.3 Resolving Associative References

**IMS coreference resolver unchanged.** As associative references are not annotated in the OntoNotes dataset, the only possible baseline we can use is the system trained on the scientific corpus. Average CoNLL scores were 33.52 for GEN and 32.86 for CL (33.14 overall). As expected, the performance on associative anaphora is worse than on coreferent anaphora. We have not made any changes to the resolver, so it is interesting to see that the resolver is indeed able to adjust to the new notion of reference and manages to link associative references.

We found that the resolver generally identifies very few associative references and so the most common error of the system is that it fails to recognise associative relations, particularly if the computed edit distance, one of the standard features in the coreference resolver, is very high. The easiest associative relations to detect are those which have similar surface forms. For example, the coreference resolver correctly links *RNAI* and *RNAI genes*, *the sense strand* and *the anti-sense strand* or *siRNAs* and *efficacious siRNAs*. However, for most of the associative references, the lack of easily identifiable surface markers makes the task difficult. Ironically, the system also falsely classifies many coreference links as associative, although it has this time of course been trained only on associative references. This is not surprising, given that the tasks are so similar that we are able to use a coreference resolver for the associative task in the first place.

**IMS coreference resolver using semantic features.** Table 9 gives the results of the extended feature set that includes the semantic features described in Section 5.2. Each of the respective semantic features shown in the table is added to the version in the line above the current version.
It can be seen that the different WordNet relations have different effects on the two scientific disciplines. For the genetics papers, the inclusion of synonyms, hyponyms and hypernyms results in the highest increase in performance (+2.02). For the computational linguistics papers, the inclusion of synonyms, hyponyms, topics and meronyms obtains the best performance (+1.19). As the effect of the features is discipline-dependent, we create two separate final feature sets for the two disciplines. The GEN version contains synonyms, hyponyms and hypernyms while the CL version contains synonyms, hyponyms, topics and meronyms. The highest increase in performance for the CL feature set (and the one resulting in the final system) was achieved by dropping topic members and hypernyms. In the final CL system, the increase in performance compared to the baseline version is +1.35. Both final versions significantly outperform the baseline.

When comparing the output of the extended system to the baseline system, it can be seen that

|  | GEN | CL | GEN+CL |
|---|---|---|---|
| Baseline | 33.52 | 32.86 | 33.19 |
| + Synonyms | 33.95 | 32.87 | 33.41 |
| + Hyponyms | 34.04 | 32.94 | 33.49 |
| + Hypernyms | **35.54** | 31.35 | 33.45 |
| + Topic members | 34.61 | 30.61 | 32.61 |
| + Topics | 34.09 | 32.88 | 33.49 |
| + Meronyms | 33.70 | **34.05** | **33.88** |
| + Substance meronyms | 33.57 | 32.40 | 32.99 |
| Final version (domain-dependent) | **35.54** | **34.21** | **34.88** |

Table 9: Resolving associative references: CoNLL metric scores for the extended feature sets

the resolver now links many more mentions (5.7 times more in the GEN papers, 3.8 times more in the CL papers). The reason why this does not lead to an even larger increase in performance lies in the large number of false positives. However, when looking at the data it becomes apparent that the newly created links are mostly links that potentially could have been annotated during the annotation, but are not part of the gold standard because the associative antecedent is not absolutely necessary in order to interpret the anaphor or because the entity has been linked to a different entity where the associative relation is stronger. The existence of more-or-less acceptable alternative associative links casts some doubt on using a gold standard as the sole evaluation criterion. An alternative would be to ask humans for a rating of the sensibility of the links determined by the system.

## 7 Conclusion and Future Work

We have presented a study of information status in two scientific disciplines as well as preliminary experiments on the resolution of both coreferent and associative anaphora in these disciplines. Our results show a marked difference in the distributions of information status categories between scientific and news text. Our corpus of 16 full-text scholarly papers annotated with information status and anaphoric links, which we plan to release soon, contains over 8000 annotated definite noun phrases. We demonstrate that the integration of domain-dependent terminological features, combined with domain-dependent training data, outperforms the unadjusted IMS system (Björkelund and Farkas, 2012) by 36.47 % absolute on the CoNLL metric for the genetics papers and by 20.40 % absolute for the computational linguistics papers. The effect of domain-dependent training material was stronger than the integration of terminological features. As far as the resolution of associative anaphora is concerned, we have shown that it is generally possible to adapt a coreference resolver to this task, and we have achieved an improvement in performance using novel semantic features. We are currently performing an inter-annotator study with two additional annotators, which will also lead to a better understanding of the relative difficulty of the categories. Furthermore, we plan to convert the coreference-annotated ACL papers by Schäfer et al. (2012) into CoNLL format and use them for training the coreference resolver. As we have annotated our corpus with information status, it might also be interesting to train a classifier on the information status categories and use its predictions to improve the performance on anaphora resolution tasks. To do so, we will create a separate corpus for testing, annotated solely with coreference and associative links.

## Acknowledgements

We would like to thank Jonas Kuhn, Ulrich Heid and Arndt Riester for their valuable comments as well as the anonymous reviewers for their insightful remarks. We gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft (DFG) in the framework of project B3 and D7 of the SFB 732. This research is partially supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D11PC20153. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

## References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Riza T. Batista-Navarro and Sophia Ananiadou. 2011. Building a coreference-annotated corpus from the domain of biochemistry. In *Proceedings of BioNLP 2011 Workshop*, pages 83–91. Association for Computational Linguistics.

Anders Björkelund and Richárd Farkas. 2012. Data-driven multilingual coreference resolution using resolver stacking. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 49–55. Association for Computational Linguistics.

Chen Chen and Vincent Ng. 2012. Combining the best of two worlds: A hybrid approach to multilingual coreference resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 56–63, Jeju Island, Korea, July. Association for Computational Linguistics.

Herbert H. Clark. 1975. Bridging. In *Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, pages 169–174. Association for Computational Linguistics.

Kevin B. Cohen, Arrick Lanfranchi, William Corvey, William A. Baumgartner Jr, Christophe Roeder, Philip V. Ogren, Martha Palmer, and Lawrence Hunter. 2010. Annotation of all coreference in biomedical text: Guideline selection and adaptation. In *Proceedings of BioTxtM 2010: 2nd workshop on building and evaluating resources for biomedical text mining*, pages 37–41.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, October. Association for Computational Linguistics.

Yimai Fang and Simone Teufel. 2014. A summariser based on human memory limitations and lexical competition. In *Proceedings of the EACL*. Association for Computational Linguistics. (to appear).

Eraldo Fernandes, Cícero dos Santos, and Ruy Milidiú. 2012. Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 41–48, Jeju Island, Korea, July. Association for Computational Linguistics.

Caroline V. Gasperin. 2009. Statistical anaphora resolution in biomedical texts. Technical Report UCAM-CL-TR-764, University of Cambridge, Computer Laboratory, December.

Anita Gojun, Ulrich Heid, Bernd Weissbach, Carola Loth, and Insa Mingers. 2012. Adapting and evaluating a generic term extraction tool. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC)*, pages 651–656.

Marti A. Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16.

Yufang Hou, Katja Markert, and Michael Strube. 2013. Global inference for bridging anaphora resolution. In *Proceedings of NAACL-HLT*, pages 907–917.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics.

Dain Kaplan, Ryu Iida, Kikuko Nishina, and Takenobu Tokunaga. 2012. Slate – a tool for creating and maintaining annotated corpora. *Journal for Language Technology and Computational Linguistics*, pages 89–101.

Walter Kintsch and Teun A. van Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review*, 85(5):363–394.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 795–804. Association for Computational Linguistics.

Shachar Mirkin, Ido Dagan, and Sebastian Padó. 2010. Assessing the role of discourse references in entailment inference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL 2010, pages 1209–1219. Association for Computational Linguistics.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411. Association for Computational Linguistics.

Malvina Nissim, Shipra Dingare, Jean Carletta, and Mark Steedman. 2004. An annotation scheme for information status in dialogue. *LREC 2004*.

Massimo Poesio and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational linguistics*, 24(2):183–216.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Joint Conference on EMNLP and CoNLL: Shared Task*, pages 1–40.

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. In *Radical Pragmatics*, pages 223–55. Academic Press.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of EMNLP 2010*.

Arndt Riester, David Lorenz, and Nina Seemann. 2010. A recursive annotation scheme for referential information status. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 717–722.

Ulrich Schäfer, Christian Spurk, and Jörg Steffen. 2012. A fully coreference-annotated corpus of scholarly papers from the acl anthology. In *Proceedings of the 24th International Conference on Computational Linguistics. International Conference on Computational Linguistics (COLING-2012), December 10-14, Mumbai, India*, pages 1059–1070.

Sidney Siegel and N. John Jr. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Berkeley, CA, 2nd edition.

Simone Teufel. 2010. *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. CSLI Publications.

Rebecca Watson, Judita Preiss, and Ted Briscoe. 2003. The contribution of domain-independent robust pronominal anaphora resolution to open-domain question-answering. In *Proceedings of the Symposium on Reference Resolution and its Applications to Question Answering and Summarization. Venice, Italy June*, pages 23–25.