Semi-supervised anaphora resolution in biomedical texts

Caroline Gasperin
Computer Laboratory,
University of Cambridge,
15 JJ Thomson Avenue,
Cambridge CB3 0FD, UK
cvg20@cl.cam.ac.uk

Abstract

Resolving anaphora is an important step in the identification of named entities such as genes and proteins in biomedical scientific articles. The goal of this work is to resolve associative and coreferential anaphoric expressions making use of the rich domain resources (such as databases and ontologies) available for the biomedical area, instead of annotated training data. The results are comparable to extant state-of-the-art supervised methods in the same domain. The system is integrated into an interactive tool designed to assist FlyBase curators by aiding the identification of the salient entities in a given paper as a first step in the aggregation of information about them.

1 Introduction

The number of articles being published in biomedical journals per year is increasing exponentially. For example, Morgan et al. (2003) report that more than 8000 articles were published in 2000 just in relation to FlyBase[1], a database of genomic research on the fruit fly *Drosophila melanogaster*.

The growth in the literature makes it difficult for researchers to keep track of information, even in very small subfields of biology. Progress in the field often relies on the work of professional curators, typically postdoctoral-level scientists, who are trained to identify important information in a scientific article. This is a very time-consuming task which first requires identification of gene, allele and protein names and their synonyms, as well as several interactions and relations between them.The information extracted from each article is then used to fill in a template per gene or allele.

To extract all information about a specific biomedical entity in the text and be able to fill in the corresponding template, a useful first step is the identification of all textual mentions that are referring to or are related with that entity. Linking all these mentions together corresponds to the task known as *anaphora resolution* in Natural Language Processing.

In this paper, we are interested in linking automatically all mentions that refer to a gene or are related to it (i.e. its 'products'). For example, in the following portion of text, we aim to link the highlighted mentions:

''... is composed of **five proteins**(1) encoded by **the male-specific lethal genes**(2) ... **The MSL proteins**(3) colocalize to hundreds of sites ... male animals die when they are mutant for any one of **the five msl genes**(4).''

In this work we use the output of a gene name recogniser (Vlachos et al., 2006) and information from the Sequence Ontology (Eilbeck and Lewis, 2004) to identify the entities of interest and the genomic relations among them. We also use RASP (Briscoe and Carroll, 2002), a statistical parser, to identify NPs (and their subconstituents) which may be anaphorically linked. Our system identifies coref-

[1] http://www.flybase.org

Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06, pages 96–103,
New York City, June 2006.

erential relations between biomedical entities (such as (1) and (3), and (2) and (4) above) as well as associative links (relations between different entities, e.g. the link between a gene and its protein as in (2) and (3) above). A previous version of this system was presented in (Vlachos et al., 2006); here we improve its results due to refinements on some of the steps previous to the resolution and to the anaphora resolution process itself.

The large majority of the entities in biomedical texts are referred to using non-pronominal noun phrases, like proper nouns, acronyms or definite descriptions. Hence, we focus on these NPs and do not resolve pronominal references (as pronouns represent only about 3% of the noun phrases in our domain).

In the following section, we detail the different components of the anaphora resolution system. The results are tested against hand-annotated papers, and an extensive evaluation is provided in Section 3, where the performance and errors are discussed.

2 The anaphora resolution system

Our system for anaphora resolution makes use of lexical, syntactic, semantic and positional information to link anaphoric expressions. The lexical information consists of the words themselves. The syntactic information consists of noun phrase boundaries and the distinction between head and premodifiers (extracted from RASP output). The distance (in words) between the anaphoric expression and its possible antecedent is taken into account as positional information. The semantic information comes from the named entity recognition (NER) process and some extra tagging based on features from the Sequence Ontology.

FlyBase is used as source of gene names, symbols and synonyms, giving rise to training data for the gene name recognition system detailed in Section 2.1. The output of this system is tagged named entities that refer to the fruit fly genes.

We then parse the text using RASP in order to extract the noun phrases and their subparts (head and modifiers). Retagging gene names as proper names before parsing improves the parser's performance, but otherwise the parser is used unmodified.

The Sequence Ontology (SO) can be used to identify words and phrases related to a gene: its subtypes (e.g. oncogene, transposable element), parts (e.g. transcript, regulatory region) and products (e.g. polypeptide, protein). Subsection 2.3 details the information extracted from SO to type the non-gene mentions.

2.1 Gene-name recognition

The NER system we use (Vlachos et al., 2006) is a replication and extension of the system developed by Morgan et al. (2004): a different training set and software were used. For training data we used a total of 16609 abstracts, which were automatically annotated by a dictionary-based gene name tagger. The dictionary consists of lists of the gene names, symbols and synonyms extracted from FlyBase. The gene names and their synonyms that were recorded by the curators from the full paper were annotated automatically in each abstract, giving rise to a large but noisy set of training data. The recognizer used is the open source toolkit LingPipe[2], implementing a 1st-order HMM model using Witten-Bell smoothing. A morphologically-based classifier was used to deal with unknown gene names (that were not present in the training data).

The performance of the trained recogniser on a revised version of the test data used in Morgan et al. (86 abstracts annotated by a biologist curator and a computational linguist) was 80.81% recall and 84.93% precision.

2.2 Parsing and NP extraction

RASP is a pipelined parser which identifies sentence boundaries, tokenises sentences, tags the tokens with their part-of-speech (PoS) and finally parses PoS tag sequences, statistically ranking the resulting derivations. We have made minor modifications to RASP's tokeniser to deal with some specific features of biomedical articles, and manually modified a small number of entries in the PoS tagger lexicon, for example to allow the use of *and* as a proper name (referring to a fruit fly gene). Otherwise, RASP uses a parse ranking module trained on a generic treebank and a grammar also developed from similar resources.

The anaphora resolution system first tags genes

[2] http://www.alias-i.com/lingpipe/

using the gene recogniser. This means that identified gene mentions can be retagged as proper names before the RASP parser is applied to the resulting PoS sequences. This improves parser performance as the accuracy of PoS tagging decreases for unknown words, especially as the RASP tagger uses an unknown word handling module which relies heavily on the similarity between unknown words and extant entries in its lexicon. This strategy works less well on gene names and other technical vocabulary from the biomedical domain, as almost no such material was included in the training data for the tagger. We have not evaluated the precise improvement in performance as yet due to the lack of extant gold standard parses for relevant text.

RASP can output grammatical relations (GRs) for each parsed sentence (Briscoe, 2006). GRs are factored into binary lexical relations between a head and a dependent of the form `(GR-type head dependent)`. We use the following GR-types to identify the head-nouns of NPs (the examples of GRs are based on the example of the first page unless specified otherwise):

- `ncsubj` encodes binary relations between non-clausal subjects and their verbal heads; e.g. (**ncsubj** colocalize proteins).

- `dobj` encodes a binary relation between verbal or prepositional head and the head of the NP to its immediate right; e.g. (**dobj** of sites).

- `obj2` encodes a binary relation between verbal heads and the head of the second NP in a double object construction; e.g. for the sentence "Xist RNA provides a mark for specific histones" we get (**dobj** provides mark) (**obj2** provides histones).

- `xcomp` encodes a binary relation between a head and an unsaturated VP complement; e.g. for the phrase "a class of regulators in Drosophila is the IAP family" we get (**xcomp** is family).

- `ta` encodes a binary relation between a head and the head of a text adjunct delimited by punctuation (quotes, brackets, dashes, commas, etc.); e.g. for "BIR-containing proteins (BIRPs)" we get (**ta** proteins BIRPs).

To extract the modifiers of the head nouns, we search the GRs typed `ncmod` which encode binary relations between non-clausal modifiers and their heads; e.g (**ncmod** genes msl).

When the head nouns take part in coordination, it is necessary to search the `conj` GRs which encode relations between a coordinator and the head of a conjunct. There will be as many such binary relations as there are conjuncts of a specific coordinator; e.g. for "CED-9 and EGL-1 belong to a large family ..." we get (**ncsubj** belong and) (**conj** and CED-9) (**conj** and EGL-1).

Last but not least, to identify definite descriptions, we search the `det` GR for a definite specifier, e.g. (**det** proteins The). By using the GR representation of the parser output we were able to improve the performance of the anaphora resolution system by about 10% over an initial version described in (Vlachos et al., 2006) that used the RASP tree output instead of GRs. GRs generalise more effectively across minor and irrelevant variations in derivations such as the X-bar level of attachment in nominal coordinations.

2.3 Semantic typing and selecting NPs

To identify the noun phrases that refer to the entities of interest, we classify the head noun as belonging to one of the five following classes: "part-of-gene", "subtype-of-gene", "supertype-of-gene", "product-of-gene" or "is-a-gene". These classes are referred to as biotypes.

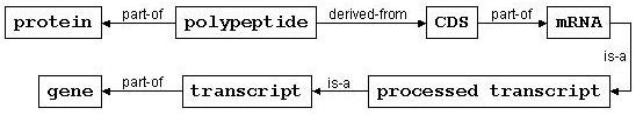

Figure 1: SO path from gene to protein.

The biotypes reflect the way the SO relates entities to the concept of the gene using the following relations: derives_from, member_of, part_of, and is_a, among others.[3] We extracted the unique path

[3] We consider the member_of relation to be the same as the part_of relation.

of concepts and relations which leads from a gene to a protein. The result is shown in Figure 1.

Besides the facts directly expressed in this path, we also assumed the following:[4]

1. Whatever is-a transcript is also part-of a gene.
2. Whatever is part-of a transcript is also part-of a gene.
3. An mRNA is part-of a gene.
4. Whatever is part-of an mRNA is also part-of a gene.
5. CDS is part-of a gene.
6. A polypeptide is a product (derived-from) of a gene.
7. Whatever is part-of a polypeptide is also a product of a gene.
8. A protein is a product of a gene.

We then used these assumptions to add new derivable facts to our original path. For example, an *exon* is a part of a transcript according to the SO, therefore, by the 2nd assumption, we add the fact that an *exon* is a part of a gene. We also extracted information about gene subtypes that is included in the ontology as an entry called "gene class". We consider NPs as supertypes of a gene when they refer to nucleotide sequences that are bigger than but include the gene.[5]

Finally, we tagged every NP whose head noun is one of the items extracted from the SO with its biotype. For instance, we would tag "the third exon" with "part-of-gene".

The NPs whose head noun is a gene name tagged in the NER phase also receive the "is-a-gene" biotype. Other NPs that still remain without biotype info are tagged as "other-bio" if any modifier of the head is a gene name.

This typing process achieves 75% accuracy when evaluated against the manually annotated corpora described in Section 3. The majority of the errors (70%) are on typing NPs that contain just a proper name, which can refer to a gene or to a protein. At the moment, all of these cases are being typed as "is-a-gene".

The biotyped NPs are then selected and considered for anaphora resolution. NPs with the same biotype can be coreferent, as well as NPs with is-a-gene and subtype-of-gene biotypes. The anaphoric relation between an is-a-gene NP and a part-of-gene or product-of-gene NP is associative rather than coreferential.

2.4 Resolving anaphora cases

We take all proper namer (PNs) and definite descriptions (DDs) among the filtered NPs as potential anaphoric expressions (anaphors) to be resolved. As possible antecedents for an anaphor we take all bio-typed NPs that occur before it in the text. For each anaphor we look for its antecedent (the closest previous mention that is related to it). For linking anaphors to their antecedents we look at:

- $head_{an}$: anaphor head noun
- $head_a$: antecedent head noun
- mod_{an}: set of anaphor pre-modifiers
- mod_a: set of antecedent pre-modifiers
- $biotype_{an}$: anaphor biotype
- $biotype_a$: antecedent biotype
- d: distance in sentences from the anaphor

The pseudo-code to find the antecedent for the DDs and PNs is given below:

- Input: a set A with all the anaphoric expressions (DDs and PNs); a set C with all the possible antecedents (all NPs with biotype information)
- For each anaphoric expression A_i:
  - Let antecedent 1 be the closest preceding NP C_j such that head(C_j)=head(A_i) and biotype(C_j)=biotype(A_i)

[4] A curator from FlyBase was consulted to confirm the validity of these assumptions.

[5] In the SO a gene holds an is-a relation to "sequence" and "region" entries.

- Let antecedent 2 be the closest preceding NP C_j such that biotype(C_j)$\neq$ biotype(A_i), but head(C_j)=head(A_i) or head(C_j)=mod(A_i) or mod(C_j)=head(A_i) or mod(C_j)=mod(A_i)
- Take the closest candidate as antecedent, if 1 and/or 2 are found; if none is found, the DD/PN is treated as non-anaphoric

- Output: The resolved anaphoric expressions in A linked to their antecedents.

As naming conventions usually recommend gene names to be lower-cased and protein names to be upper-cased, our matching among heads and modifiers is case-insensitive, allowing, for example, `msl gene` to be related to `MSL protein` due to their common modifiers.

Antecedent 1, if found, is considered coreferent to A_i, and antecedent 2, associative. For example, in the passage:

``Dosage compensation, which ensures that the expression of **X-linked genes**:C_j is equal in males and females ... the hypertranscription of **the X-chromosomal genes**:A_j in males ...''

the NP in bold font which is indexed as antecedent C_j is taken to be coreferential to the anaphor indexed as A_j. Additionally, in:

``... the role of **the roX genes**:C_k in this process ... which MSL proteins interact with **the roX RNAs**:A_k ...''

C_k meets the conditions to form an associative link to A_k. The same is true in the following example in which there is an associative relation between C_j and A_j:

``The expression of **reaper**:C_j has been shown to be regulated by distinct stimuli ... it was shown to bind a specific region of **the reaper promoter**:A_j ...''

If we consider the example from the first page, mention (1) is returned by the system as the coreferent antecedent for (3), as they have the same biotype and a common head noun. In the same example, (2) is returned as a coreferent antecedent to (4), and (3) as an associative antecedent to (4).

3 Evaluation

We evaluated our system against two hand-annotated full papers which have been curated in FlyBase and were taken from PubMed Central in XML format. Together they contain 302 sentences, in which 97 DDs and 217 PNs related to biomedical entities (out of 418 NPs in total) were found.

For each NP, the following information was manually annotated:

- NP form: definite NP, proper name, or NP.
- biotype: gene, part-of-gene, subtype-of-gene, supertype-of-gene, product-of-gene, other-bio, or a non-bio noun.
- coreferent antecedent: a link to the closest previous coreferent mention (if there is one).
- associative antecedent: a link to the closest previous associative anaphoric mention (if there is one, and only if there is no closer coreferent mention).

All coreferent mentions become linked together as a coreference chain, which allows us to check for previous coreferent antecedents of a mention besides the closest one.

Table 1 shows the distributions of the anaphoric expressions according to the anaphoric relations they hold to their closest antecedent.

	coreferent	associative	no ant.	Total
DDs	34	51	12	97
PNs	132	62	23	217
Total	166	113	35	314

Table 1: Anaphoric relation distribution

DDs and PNs in associative relations account for 27% of all NPs in the test data, which is almost double the number of bridging cases (associative plus coreferent cases where head nouns are not the same) reported for newspaper texts in Vieira and Poesio (2000).

Table 2 shows the distribution of the different biotypes present in the corpus.

gene	part	subtype	supertype	product
67	62	1	7	244

Table 2: Biotype distribution

	coreferent	associative	no ant.
DDs	20/+2	14/+13	7
PNs	115/+9	11/+22	16

Table 3: Resolved anaphoric relations

3.1 Results

The anaphora resolution system reaches 58.8% precision and 57.3% recall when looking for the closest antecedent for DDs and PNs, after having been provided with hand-corrected input (that is, perfect gene name recognition, NP typing and selection). If we account separately for coreference and associative relations, we get 59.47% precision and 81.3% recall for the coreferent cases, and 55.5% precision and 22.1% recall for the associative ones.

The performance of the system is improved if we consider that it is able to find an antecedent other than the closest, which is still coreferential to the anaphor. These are cases like the following:

``**five proteins** encoded by *the male-specific lethal genes* ... **The MSL proteins** ...''

where the system returns "five proteins" as the coreferent antecedent for "the MSL proteins", instead of returning "the male-specific lethal genes" as the closest (in this case, associative) antecedent. Treating these cases as positive examples we reach 77.5% precision and 75.6% recall[6]. It conforms with the goal of adding the anaphor to a coreferential chain rather than simply relating it to the closest antecedent.

Table 3 reports the number of coreferent and associative DDs and PNs that could be resolved. The numbers on the left of the slash refer to relations with the closest antecedent, and the numbers on the right refer to additional relations found when links with another antecedent are considered (all the new positive cases on the right are coreferent, since our evaluation data just contain associative links to the closest antecedent).

Most of the cases that could be resolved are coreferent, and when the restriction to find the closest antecedent is relaxed, the system manages to resolve 35 cases of DD coreference (64.7% recall). It achieves very high recall (93.9%) on coreferential PNs. All the associative relations that are hand annotated in our evaluation corpus are between an anaphor and its closest antecedent, so when the recency preference is relaxed, we get coreferent instead of associative antecedents: we got 35 coreferent antecedents for anaphors that had a closest associative antecedent that could not be recovered. This conforms to the goal of having coreference chains that link all the mentions of a single entity.

The system could resolve around 27% of the associative cases of DDs, although fewer associative antecedents could be recovered for PNs, mainly due to the frequent absence of head-noun modifiers and different forms for the same gene name (expanded vs. abbreviated).

Although associative anaphora is considered to be harder than coreference, we believe that certain refinements of our resolution algorithm (such as normalizing gene names in order to take more advantage of the string matching among NP heads and modifiers) could improve its performance on these cases too.

The anaphora resolution system is not able to find the correct antecedent when there is no head or modifier matching as in the anaphoric relation between ``Dark/HAC-1/Dapaf-1'' and ``The Drosophila homolog''.

The performance rates drop when using the output of the NER system (presented in Section 2.1), RASP parsing (Section 2.2) and SO-based NP typing (Section 2.3), resulting in 63% precision and 53.4% recall.

When the NER system fails to recognise a gene name, it can decrease the parser performance (as it would have to deal with an unknown word) and influences the semantic tagging (the NP containing such a gene name won't be selected as a possible antecedent or anaphor unless it contains another word that is part of SO). When just the NER step is corrected by hand, the system reaches 71.8% precision

[6] We are able to compute these rates since our evaluation corpus includes also a coreferent antecedent for each case where an associative antecedent was selected.

and 64.1% recall.

4 Related work

Previous approaches to solve associative anaphora have made use of knowledge resources like WordNet (Poesio et al., 1997), the Internet (Bunescu, 2003) and a corpus (Poesio et al., 2002) to check if there is an associative link between the anaphor and a possible antecedent.

In the medical domain, Castaño et al. (2002) used UMLS (Unified Medical Language System)[7] as their knowledge source. They treat coreferential pronominal anaphora and anaphoric DDs and aim to improve the extraction of biomolecular relations from MEDLINE abstracts. The resolution process relies on syntactic features, semantic information from UMLS, and the string itself. They try to resolve just the DDs that refer to relevant biotypes (corresponding to UMLS types) such as amino acids, proteins or cells. For selecting the antecedents, they calculate salience values based on string similarity, person/number agreement, semantic type matching and other features. They report precision of 74% and recall of 75% on a very small test set.

Yang et al. (2004) test a supervised learning-based approach for anaphora resolution, evaluating it on MEDLINE abstracts from the GENIA corpus. They focus only on coreferent cases and do not attempt to resolve associative links. 18 features describe the relationship between an anaphoric expression and its possible antecedent - their source of semantic knowledge is the biotype information provided by the NER component of GENIA. They achieved recall of 80.2% and precision of 77.4%. They also experiment with exploring the relationships between NPs and coreferential clusters (i.e. chains), selecting an antecedent based not just on a single candidate but also on the cluster that the candidate is part of. For this they add 6 cluster-related features to the machine-learning process, and reach 84.4% recall and 78.2% precision.

Our system makes use of extant biomedical resources focused on the relevant microdomain (fruit fly genomics), and attempts to tackle the harder problem of associative anaphora, as this constitutes a significant proportion of cases and is relevant to the curation task. Our performance rates are lower than the ones above, but did not rely on expensive training data.

5 Concluding remarks

Our system for anaphora resolution is semi-supervised and relies on rich domain resources: the FlyBase database for NER, and the Sequence Ontology for semantic tagging. It does not need training data, which is a considerable advantage, as annotating anaphora by hand is a complicated and time-demanding task, requiring very precise and detailed guidelines.

The resulting links between the anaphoric entities are integrated into an interactive tool which aims to facilitate the curation process by highlighting and connecting related bio-entities: the curators are able to navigate among different mentions of the same entity and related ones in order to find easily the information they need to curate.

We are currently working on increasing our evaluation corpus; we aim to make it available to the research community together with our annotation guidelines.

We intend to enhance our system with additional syntactic features to deal with anaphoric relations between textual entities that do not have any string overlap. We also intend to add different weights to the features. The performance of the fully-automated version of the system can be improved if we manage to disambiguate between gene and protein names and infer the correct biotype for them. The performance on associative cases could be improved by normalizing the gene names in order to find more matches among heads and modifiers.

Acknowledgements

This work is part of the BBSRC-funded FlySlip[8] project. Caroline Gasperin is funded by a CAPES award from the Brazilian government. Thanks to Nikiforos Karamanis and Ted Briscoe for their comments and help with this manuscript.

[7] http://www.nlm.nih.gov/research/umls/

[8] http://www.cl.cam.ac.uk/users/av308/Project_Index/Project_Index.html

References

Ted Briscoe and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of LREC 2002*, pages 1499–1504, Las Palmas de Gran Canaria.

Ted Briscoe. 2006. Tag sequence grammars. Technical report, Computer Laboratory, Cambridge University.

Razvan Bunescu. 2003. Associative anaphora resolution: A web-based approach. In *Proceedings of EACL 2003 - Workshop on The Computational Treatment of Anaphora*, Budapest.

José Castaño, Jason Zhang, and James Pustejovsky. 2002. Anaphora resolution in biomedical literature. In *Proceedings of International Symposium on Reference Resolution for NLP 2002*, Alicante, Spain.

Karen Eilbeck and Suzanna E. Lewis. 2004. Sequence ontology annotation guide. *Comparative and Functional Genomics*, 5:642–647.

Alex Morgan, Lynette Hirschman, Alexander Yeh, and Marc Colosimo. 2003. Gene name extraction using FlyBase resources. In *Proceedings of ACL 2003 Workshop on Natural Language Processing in Biomedicine*, Sapporo, Japan.

Alex Morgan, Lynette Hirschman, Mark Colosimo, Alexander Yeh, and Jeff Colombe. 2004. Gene name identification and normalization using a model organism database. *J. of Biomedical Informatics*, 37(6):396–410.

Massimo Poesio, Renata Vieira, and Simone Teufel. 1997. Resolving bridging descriptions in unrestricted texts. In *Proceedings of the Workshop on Operational Factors in the Practical, Robust, Anaphora Resolution for Unrestricted Texts*, Madrid.

Massimo Poesio, Tomonori Ishikawa, Sabine Schulte im Walde, and Renata Vieira. 2002. Acquiring lexical knowledge for anaphora resolution. In *Proceedings of LREC 2002*, Las Palmas De Gran Canaria.

Renata Vieira and Massimo Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):525–579.

Andreas Vlachos, Caroline Gasperin, Ian Lewin, and Ted Briscoe. 2006. Bootstrapping the recognition and anaphoric linking of named entities in Drosophila articles. In *Proceedings of the PSB 2006*, Hawaii.

Xiaofeng Yang, Jian Su, Gouodong Zhou, and Chew Lim Tan. 2004. An NP-cluster based approach to coreference resolution. In *Proceedings of COLING 2004*, Geneva, Switzerland, August.