An Investigation into the Reusability of Biomedical Terminologies for the Resolution of Referential Expressions

Michael Poprat & Udo Hahn
Jena University Language and Information Engineering (JULIE) Lab
D-07743 Jena, Germany
{*poprat,hahn*}@*coling-uni-jena.de*

Abstract

We investigate the coverage of (domain specific) terminological resources (taxonomies, ontologies) for the resolution of referential expressions in biomedical abstracts. We focus on anaphora that require text-external knowledge for their resolution. The results show that, for the resolution of nominal and subgrouping anaphora, the resources are useful while they are insufficient for resolving bridging anaphora.

1 Introduction

Reference relations in documents between a previously mentioned discourse entity and its anaphoric referent are a wide-spread linguistic phenomenon in discourse. Lacking recognition of such relations is likely to have negative effects on the proper representation of the content of a document — its knowledge remains incomplete, incoherent, and even invalid [5].

In order to resolve referential expressions different types of increasingly complex knowledge are brought to bear. Some varieties of reference relations (e.g., repetitions) can be dealt with using text-internal evidence without any recourse to external knowledge sources. Others (e.g., pronominal anaphora) require modest forms of linguistic knowledge such as consistent gender or number features. However, some complex kinds of reference relations (e.g., nominal and bridging anaphora) rely on profound domain knowledge (e.g., taxonomic). In the biomedical domain, such resources are large in numbers, but at the same time, they are highly fragmented and usually lack semantic interoperability. Hence, the question turns up whether these complex forms of anaphora occur at a reasonably high rate and, if so, how to deal with them given the available resources.

In this exploratory study, we first derive some basic (and still preliminary) corpus statistics on the distribution of various types of reference expressions in biomedical corpora in order to assess whether biomedical terminologies are really needed for large-scale reference resolution (i.e., whether knowledge-intensive anaphora occur at a high rate or not). Second, we focus on the knowledge-intensive anaphoric expressions in the two corpora and try to reuse existing biomedical resources for reference resolution in order to assess their usability for this task.

2 Related Work

A lot of work has been done on the resolution of pronominal anaphora which relies on simple morpho-syntactic features, syntactic constraints and ordering constraints in so-called discourse memories. When it comes to knowledge-intensive forms of referential relations considerably less effort has been spent. For the general-language use case in newspaper documents, WORDNET [3] is taken as a standard resource for accessing taxonomic or partonomic relation information (cf. e.g., [9] for bridging anaphora). Recently, [12] reported good resolution results when using the categorization of articles in the folksonomy WIKIPEDIA. As an alternative to these knowledge-driven approaches pattern-oriented [8] and corpus-based methods [11] have also been tried to by-pass lacking knowledge resources.

Also in the biomedical domain, knowledge-poor approaches to pronominal anaphora resolution have already been tried [14], while the knowledge-intensive varieties have only rarely been studied. In [1] and [6], the authors deal with nominal anaphora using the UMLS [13] semantic classes as features for their salience measure-based resolution algorithm. They find that these classes are too coarse-grained.

The resolution of bridging anaphora requires more than a hierarchically organized taxonomic terminology because relations such as PART-OF or DERIVES-FROM are essential. Hence, [2] resolves bridging anaphora with a home-grown ontology. [4], on the other hand, makes use of and slightly extends a small fragment of the SEQUENCE ONTOLOGY[1] that describes the path from a gene to a protein.

3 Corpora and Resources

3.1 Anaphora Corpora

For our study, we consider two document collections. The first one is the MEDSTRACT corpus [1], which contains 32 abstracts (6052 tokens). In this corpus, entities, events, acronyms and also anaphora (pronominal, nominal and subgrouping anaphora; the latter is explained below) are annotated. Second, we compiled our own corpus (the JENREF corpus) with abstracts selected from MEDLINE[2]

[1] `http://www.sequenceontology.org/`

[2] MEDLINE is the largest bibliographic database in the life sciences.

for stem cell transplantation and gene regulation. Out of these articles, 11 stem cell abstracts and 9 gene regulation abstracts (altogether about 12,000 tokens) were annotated by a team of one biologist and one computational linguist (the first author). We annotated the following types of referring expressions: As *repetitions*, we counted string-identical, string-variant and abbreviated token sequences in nominal phrases, all identical in their meaning (e.g., *"CYP1B1"* – *"CYP1B1"*). Anaphora comprise *pronominal* anaphora (e.g., *"it"*, *"they"*),[3] *nominal* anaphora (relying on IS-A relations, such as *"p58 MAPK phosphorylation"* IS-A *"reaction"*) and *bridging* anaphora (relying on semantic relations other than IS-A, e.g., *"M-G(1) transition"* PART-OF *"cell cycle"*). We also annotated *subgrouping* anaphora. They occur when a group of entities (e.g., *"vascular endothelial growth factor receptors"*) are mentioned first and certain subgroups (such as *"VEGFR1"*, *"VEGFR3"*, etc.) are discussed later on in the document. The quantitative distribution of these different types of referring expressions in both corpora is summarized in Table 1. Note that repetitions and bridging anaphora are not annotated in the MEDSTRACT corpus.

Type	JenRef	Medstract
Repetitions	388	–
Pronominal Anaphora	48 (INT) 6 (EXT)	22 (INT) 6 (EXT)
Nominal Anaphora	79	44
Bridging Anaphora	42	–
Subgrouping Anaphora	91	13

Table 1: *Number of Referring Expressions*

In the remainder of this paper, we largely ignore pronominal anaphora and repetitions because we assume that no text-external knowledge is required for their resolution.

3.2 Terminological Resources

In our experiments, the background knowledge needed for the resolution of knowledge-intensive referring expressions is taken from the following terminological resources:

UMLS. The Unified Meta Language System is an umbrella system comprising more than 100 (multilingual) terminologies for the clinical, medical and biological domain. In its current version 2007AA it contains more than a million different concepts that are classified by 134 semantic classes and interlinked by various types of relations. For our purposes we used the UMLS Browser[4] and limited the relationships we searched for to PARENT, CHILD, BROADER, NARROW. We also used the built-in function to find the shortest path between two terms with a maximum path length of 3.

BioPortal. This is the Web interface[5] to the library of the 52 OBO[6] ontologies and to 10 external resources (about 281,000 entries). It allows browsing and searching them. Besides these ontologies, we also incorporated our lab-grown ontology for major histocompatibility complex alleles and molecules.

BioThesaurus. The BIOTHESAURUS [7] is a Web-based thesaurus of protein and gene names based on UNIPROT[7] linking more than 90 molecular databases with IPROCLASS descriptors.[8] The collection of databases comprises nearly six million unique entries.

WordNet. WORDNET is an electronic lexical database. Its entries are organized in terms of their part-of-speech category (nouns, verbs, adjectives, adverbs). Nouns are grouped in sets of synonyms that define their meaning. Most terms are interconnected via taxonomic and partonomic relations. Although WORDNET 3.0 is not specific for our domain of interest, it contains some basic biomedical terms at its very bottom layer and, thus may cover missing links to general biological knowledge.

The preference of use for the resources involved was determined by their coverage and diversity of biomedical terms. Accordingly, UMLS was consulted first, then we searched the ontologies via BIOPORTAL. BIOTHESAURUS was browsed when we found no results in the other resources. If none of them yielded a result, WORDNET came into play.

4 Approach

4.1 Some (Crucial) Assumptions

For our manually conducted analysis, we make quite strong assumptions as far as the conduct of analysis is concerned. Nevertheless, they seem justified because our goal is to assess the quality and coverage of different terminological resources for anaphora resolution in biomedical documents rather than the evaluation of a particular resolution method. Hence, the following sources of errors which automatic procedures for anaphora resolution usually face are not an issue here:

Antecedent Finding. We assume a perfect anaphora detector and a perfect antecedent finder.

Mapping into Resources. We also ignore the difficult task of mapping passages of a text into entries of a terminology. Not only variations at a string level complicate such a matching process, but also the problems of synonymy (in particular in the biomedical domain!) and ambiguity. We always select the right concept in the terminologies (manually).

Linkage between Resources. We respect the identity of names (for entities and relations) in the different terminologies as identity in meaning (e.g., the mention of *"T cell"* in the CELL ONTOLOGY (CL:0000084) as well as in UMLS (C0039194) and in WORDNET refers to the same concept). Furthermore, we assume an equivalence class for the different hierarchical relations such as IS-A, BROADER (UMLS), PAR (UMLS) and also for the inverse relations (IS-SUBGROUP-OF, NARROW (UMLS), CHD (UMLS)). These assumptions allow us to combine

[3] Excluding *"we"* since it always refers to the authors of the abstract. We also distinguish between sentence-internal (INT) and sentence-external (EXT) use of a pronominal anaphora.

[4] http://imhotep.ucsd.edu:7873/knowme/umls.html

[5] http://www.bioontology.org/tools/portal/bioportal.html

[6] http://obo.sourceforge.net

[7] http://www.expasy.uniprot.org/

[8] http://pir.georgetown.edu/iproclass/

the resources in case a single resource is not sufficient. At the moment, these requirements are being realized for the ontologies that are curated by the OBO FOUNDRY,[9] also searchable via the BIOPORTAL.

4.2 Decision Procedure

In order to decide whether or not a knowledge-intensive anaphoric expression can be resolved with the given resources and assumptions being made, we define the following decision procedure. Ideally, we find both the anaphoric expression and its antecedent (with all modifications) in the resources enumerated in Section 3.2. Furthermore, anaphoric expression and antecedent are related adequately or map to the same entry in the knowledge resources. The anaphoric expression can thus be resolved immediately. However, it is often the case that the anaphoric expression or the antecedent are supplemented by adjective phrases or prepositional phrases (e.g., *"a neuronally enriched protein"*). If this leads to an unsuccessful mapping, we *generalize* the anaphoric expression or the antecedent by stripping off modifiers until we match the corresponding term in the knowledge resources (here, *protein*). If anaphoric expression and antecedent are both found, we sometimes map to the same entry in the resources (e.g., *"increases in releases"* and *"the increase"* map both to 'C0442805' in UMLS). If we cannot determine either the anaphoric expression or the antecedent, or if the relation between them is inappropriate, we try a purely *text-internal resolution* cutting off (sub)strings from the expressions at hand (e.g., the terms *"carcinoma cells"* and *"the cells"* can be found in the resources, but they are not linked). Since we are aware of the complications by such a simplified approach (see also Section 6), we only decide on a resolution at this level as a last resort.

5 Results

To classify the resolution results of knowledge-intensive anaphora, we distinguish four positive cases (PC) for success and two negative ones (NC) for failure according to the decision procedure (cf. Section 4.2). A successful anaphora resolution occurs if:

PC 1: The anaphoric expression and the antecedent can be *found* in the resources and are adequately *related* or map to the *same entry* in the resources.

PC 2: The anaphoric expression and the antecedent (one or both *generalized*) can be *found* in the resources and are adequately *related* or map to the *same entry* in the resources.

PC 3: The anaphoric expression and the antecedent (one or both generalized, or not at all) can be *found* in the resources, but their relation is *inadequate* or not existent; however, a *text-internal resolution* is possible.

PC 4: Either the anaphoric expression or the antecedent (one or both generalized, or not at all) or both *cannot be found* in the resources; however, a *text-internal resolution* is possible.

[9] http://obofoundry.org

Case	JenRef			Medstract		
	Nom	Sub	Brdg	Nom	Sub	Brdg
PC 1	19	44	2	8	3	-
PC 2	28	23	2	17	5	-
PC 3	5	1	0	0	0	-
PC 4	4	5	3	5	0	-
NC 1	11	6	8	9	4	-
NC 2	12	12	27	5	1	-

Table 2: *Number of Positive Cases (PC) and Negative Cases (NC) of Resolutions (Nom=Nominal Anaphora, Sub=Subgroup, Brdg=Bridging Anaphora)*

Note that we do not require external knowledge for a text-internal resolution (PC 3 & 4) at this level. The following reasons are behind a failed resolution:

NC 1: Both the anaphoric expression and the antecedent (one or both generalized, or not at all) can be *found* in the resources, but their relation is *inadequate* or not existent; a *text-internal resolution* is also *not possible*.

NC 2: Either the anaphoric expression or the antecedent (one or both generalized, or not at all) or both *cannot be found* in the resources; a *text-internal resolution* is also *not possible*.

Given this categorization, the positive cases 1 and 2 mirror the real benefit of the resources considered. NC 1 reflects inappropriate relationships between terms that were found in the resources, while NC 2 reflects missing terms.

Table 2 summarizes our findings. Out of 79 nominal anaphora in the JENREF corpus (PC 1-4) 56 (71%) have been resolved successfully (68% in the MEDSTRACT corpus). Out of these, for 47 nominal anaphora/antecedent pairs (25 in MEDSTRACT), the resources were suitable with respect to term coverage and IS-A relations (PC 1-2). Due to a text-internal resolution, we could handle the missing 9 (5 in MEDSTRACT) nominal anaphora (PC 3-4). For the subgrouping anaphora, 80% (73 out of 91) were successfully resolved (MEDSTRACT: 8 of 13 (61%)). We here also observe a very high coverage of 74% (100% in MEDSTRACT) of cases (PC 1-2), in which the use of resources is successful. An explanation might be that such relations are often based on (sub)categorizations of biomedical entities. Note that for 64 out of 147 of all nominal anaphora and subgrouping PCs 1 & 2 in both corpora, also a text-internal resolution would have been possible. But as we want to assess the reusability of the resources, we preferred a match into knowledge resources to the text-internal resolution strategy (see also Section 6).

However, a different picture emerges when we turn to the bridging anaphora in the JENREF corpus: only 17% (7 out of 42) were successfully resolved (PC 1-2). And out of those, the knowledge available in the resources was helpful only in 4 cases (10%).

6 Discussion

For both corpora, UMLS provided a broad coverage for terms and adequate relations between them (51% for JENREF and 79% MEDSTRACT). For 24 resolution examples (22%) in the JENREF corpus,

we had to explore the OBO MOLECULAR ROLE ontology and in 12 cases (12%), we identified terms and relations in the NCI THESAURUS[10]. The need for inter-linkage between different resources can obviously be down-played because only in 2 cases a combination of two resources was required. WORDNET proved to be useful in 4 cases (4%).

A closer look at the cases that could not be resolved (NC 1 & 2) reveals, on the one hand, an insufficient coverage of the terminological resources involved as far as for general (both encyclopedic and biological) knowledge is concerned. On the other hand, the sloppiness of the authors' use of terminology and the gaps that have to be bridged conflicts with the claim of providing consistent ontologies.

Most of the NC 1 examples, where neither a relation could be found in the knowledge resources nor a text-internal resolution was possible, can be traced back to the fact, that the anaphoric expression contains a general-language term. As an example, the *"dramatic upregulation of FasL"* is resumed by *"this effect"*. Furthermore, an anaphoric expression can provide new functional or relational information not already coded in any of the resources. *"1-(1-propynyl)pyrene"*, e.g., is a hydrocarbon and is referred by *"the inhibitor"* that describes its role in the particular context.

Moreover, basic biological knowledge is not or only indirectly described in the knowledge resources considered. The simple fact, e.g., that a gene encodes a protein, is often needed to bridge anaphorical gaps. To the best of our knowledge, this relation is not explicitly encoded in any of the resources considered. In the SEQUENCE ONTOLOGY, the relationship between gene and protein needs five intermediate concepts and three different relations. In the NCI THESAURUS, one can find the relation that a "Gene Product" (C26548) HAS-SUBCLASS "Protein" (C17021), but the relation needed is hidden in the class "Gene Product". Last but not the least, encyclopedic knowledge about amounts, durations, times, places and dimensions is completely lacking in the resources we used. The fact, e.g., that a certain sequence *"[-1614;+20]"* contains *"the [-210;-176] fragment"* and that the *"the [-247;-210] region"* precedes the latter, cannot be inferred with the help of any available resources (22 of all 57 NC 2 examples).

In most cases of positively resolved anaphora, a generalization was necessary to map successfully into the given knowledge resources (PC 2). In some cases, actually, the resolution could only be based on string comparisons (PC 3 & 4). The big picture reveals that the approach seems to be very simple and efficient, *viz.* just removing modifiers or comparing (sub)strings. Some examples show that this is tricky though. In *"the VEGFR3-transfected ECs exhibited high expression level of LYVE-1."*, this statement on ECs only holds if the modifier *"VEGFR3-transfected"* is taken into account. Furthermore, transfected ECs are not identical with non-transfected ECs which would be the result if considering strings only. But not every modifier has the same influence. For example, the purification in *"... when priming with purified CD34(+) cells"* has no influence on the CD34 cells and statements about these cells keep their generality. A classification of such modifiers adding information with or without an influence on the semantics of the modified expression must be made. We conclude that an interpretation of (NP) expressions as proposed, e.g., by [10] is unavoidable for proper anaphora resolution.

[10] UMLS contains only a subset of the NCI THESAURUS.

7 Conclusion and Future Work

We conducted a quantitative study on the distribution of various types of anaphora in two different biomedical corpora. We then assessed the contribution of existing biomedical terminologies for the different types of anaphora. We have gathered empirical evidence that for nominal and subgrouping anaphora the existing resources are reasonably competent, while for bridging their performance drops dramatically.

Among the open issues we plan to tackle in the future is the role of modifiers for entity matches. From a technical perspective, an anaphora resolution module is missing up until now. Having such a module, we will have to deal with the tricky issues involved in accessing the resources. Then, not only the coverage of resources will influence the resolution results but also the quality of strategies how to avoid false positive resolutions given a set of possible antecedents.

Acknowledgements: This study was funded by the EC (BOOTStrep, FP6-028099), and by the German Ministry of Education and Research (StemNet, 01DS001A - 1C).

References

[1] J. Castaño, J. Zhang, and J. Pustejovsky. Anaphora resolution in biomedical literature. In *Proc.of The Int. Symposium on Reference Resolution for Natural Language Processing*. Alicante, Spain, 2002.

[2] P. Cimiano. On the research of bridging references within information extraction systems. Diploma thesis, University of Karlsruhe, 2003.

[3] C. Fellbaum, editor. WORDNET*: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

[4] C. Gasperin. Semi-supervised anaphora resolution in biomedical texts. In *Proc. of the BioNLP Workshop at HLT-NAACL 06*, pages 96–103, 2006.

[5] U. Hahn, M. Romacker, and S. Schulz. Discourse structures in medical reports – watch out! The generation of referentially coherent and valid text knowledge bases in the MEDSYNDIKATE system. *International Journal of Medical Informatics*, 53(1):1–28, 1999.

[6] Y.-H. Lin and T. Liang. Pronominal and sortal anaphora resolution for biomedical literature. In *Proc. of ROCLING XVI*, Taipei, Taiwan, 2004.

[7] H. Liu, Z.-Z. Hu, J. Zhang, and C. Wu. BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, 22(1):103–105, 2006.

[8] K. Markert and M. Nissim. Comparing knowledge sources for nominal anaphora resolution. *Computational Linguistics*, 31(3):367–402, 2005.

[9] J. Meyer and R. Dale. Using the Wordnet hierarchy for associative anaphora resolution. In *Proc. of SemaNet'02 at COLING 2002*, pages 1–8. Taipei, Taiwan, 2002.

[10] C. J. Mugall. Obol: integrating language and meaning in bio-ontologies. *Comparative and Functional Genomics*, 5:509–520, 2004.

[11] M. Poesio and R. Vieira. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216, 1998.

[12] S. P. Ponzetto and M. Strube. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proc. of the HLT-NAACL 2006*, pages 192–199. Brooklyn, New York, 2006.

[13] UMLS. *Unified Medical Language System*. Bethesda, MD: National Library of Medicine, 2007.

[14] G. Z. Xiaofeng Yang, Jian Su and C. L. Tan. An NP-cluster based approach to coreference resolution. In *Proc. of the COLING 2004*, pages 226–232. Geneva, Switzerland, 2004.