Quantitative Data on Referring Expressions in Biomedical Abstracts

Michael Poprat **Udo Hahn**
Jena University Language and Information Engineering (JULIE) Lab
Fürstengraben 30, 07743 Jena, Germany
{poprat|hahn}@coling-uni-jena.de

Abstract

We report on an empirical study that deals with the quantity of different kinds of referring expressions in biomedical abstracts.

1 Problem Statement

One of the major challenges in NLP is the resolution of referring expressions. Those references can be established by repeating tokens or by pronominal, nominal and bridging anaphora. Experimental results show that pronominal anaphora are easier to resolve than nominal ones because the resolution of nominal anaphora requires an IS-A-taxonomy as knowledge source. The resolution of bridging anaphora, however, proves to be awkward because encyclopedic knowledge is necessary.[1] But in practice, are all of these phenomena equally important? A look at the publications reveals that a comprehensive overview of the quantity and distribution of referring expressions in biomedical abstracts is still missing. Nevertheless, some scattered data can be found: Castaño et al. (2002) state that 60 of 100 anaphora are nominal anaphora. Sanchez et al. (2006) confirm this proportion (24 pronominal and 50 nominal anaphora in 74 anaphoric expressions). Kim and Park (2004), however, detect 53 pronominal and 26 nominal anaphora in 87 anaphoric expressions. But Gawronska and Erlendsson (2005), on the other hand, claim that pronominal anaphora are rare and nominal anaphora are predominant. Studies on bridging anaphora in the biomedical domain are really still missing. Only Cimiano (2003) states that 10% of definite descriptions are bridging anaphora.

This contradictoriness and the lack of statistics on referring expressions induced us to collect our own data in order to obtain a consistent and meaningful overview. This picture helps to decide where to start if one wants to build a resolution component for the biomedical domain.

2 Empirical Study

For our study we selected articles from MEDLINE for stem cell transplantation and gene regulation. Out of these articles, 11 stem cell abstracts and 9 gene regulation abstracts ($\sim$ 12,000 tokens) were annotated by a team of one biologist and one computational linguist. The boundaries for annotations were neither limited to nominal phrases (NPs) nor on their heads because NPs in biomedical abstracts are often complex and hide relations between nouns (e.g., a *"p53 protein"* is a protein called *"p53"*, a *"p53 gene"* is a gene that codes the *"p53 protein"* and a *"p53 mutation"* is a mutation in the *"p53 gene"*). Furthermore, we annotated anaphoric expressions referring to biomedical entities and to processes.

We distinguished the following referring expressions: As repetitions, we counted string-identical, string-variants and abbreviated token sequences in NPs, identical in their meaning (e.g. *"Mesenchymal stem cells"* - *"MSCs"* - *"MSC inhibitory effect"*). For the time being, modifiers have not been considered. Anaphora comprise pronominal[2], nominal (IS-A relations, e.g., *"B-PLL"* IS-

[1] However, even the resolution of pronouns can benefit from extra-textual information (Castaño et al., 2002).

[2] Without "we" as it always refers to the authors.

BioNLP 2007: Biological, translational, and clinical language processing, pages 193–194,
Prague, June 2007.

Type of Referring Expression	Number
Repetitions	388
Pronominal Anaphora	48 (sent. internal) 6 (sent. external)
Nominal Anaphora	79
Bridging Anaphora	42
Subgrouping Anaphora	91
all	**654**

Table 1: Number of Referring Expressions

A *"B-cell malignancy"*) and bridging anaphora (all other semantic relations, e.g., *"G(1) progression"* PART-OF-PROCESS *"M-G(1) transition"*). Furthermore, we detected a high number of subgrouping anaphora that often occur when a group of entities (e.g., *"Vascular endothelial growth factor receptors"*) are mentioned first and certain subgroups (e.g., *"VEGFR1"* etc.) are discussed later.

In our abstracts we detected 654 referring expressions (see Table 1). Repetitions are predominant with 59%. Within the group of 266 anaphora, subgrouping anaphora contributed with 34%, nominal anaphora with 30%, pronominal anaphora with 20% and bridging anaphora with only 16%. The most common bridging relations were PART-OF-AMOUNT (14) and PART-OF (11). The remaining 17 are held by 8 other semantic relations such as RESULTS-FROM, MUTATED-FROM, etc.

3 Open Issues and Conclusion

In biomedical abstracts we are confronted with numerous repetitions, mainly containing biomedical entities. Their reference resolution within an abstract seems to be easy at first glance by just comparing strings and detecting acronyms. Some examples will 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 NP heads only. But not every modifier influences an identity relation. 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 influencing the semantics of the modified expression must be made. Hence, we have to be careful with assumed repetitions and we have to handle all kinds of modifiers.

In this study we present the first comprehensive overview of various kinds of referring expressions that occur in biomedical abstracts. Although our corpus is still small, we could observe the strong tendency that repetitions play a major role (20 per abstract). Anaphora occur less frequently (13 per abstract). For a sound semantic interpretation, both types must be handled. For knowledge-intensive anaphora resolution, the existing biomedical resources must be reviewed for adequacy. To the best of our knowledge, although dominant in our study, subgrouping anaphora have not been considered in any anaphora resolution systems and suitable resolution strategies must be found. The annotation process (with more than one annotation team) will be continued. The main result of this study, however, is the observation that modifiers play an important role for referencing. Their treatment for semantic interpretation requires further investigations.

Acknowledgements: We thank Belinda Würfel for her annotation work. This study was funded by the EC (BOOTStrep, FP6-028099), and by the German Ministry of Education and Research (StemNet, 01DS001A - 1C).

References

J. Castaño, J. Zhang, and J. Pustejovsky. 2002. Anaphora resolution in biomedical literature. In *Proc. of the Symp. on Reference Resolution for NLP*.

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

B. Gawronska and B. Erlendsson. 2005. Syntactic, semantic and referential patterns in biomedical texts: Towards in-depth comprehension for the purpose of bioinformatics. In *Proc. of the 2nd Workshop on Natural language Understanding and Cognitive science*, pages 68–77.

J.-J. Kim and J. C. Park. 2004. BioAR: Anaphora resolution for relating protein names to proteome database entries. In *Proc. of the ACL 2004: Workshop on Reference Resolution and its Applications*, pages 79–86.

O. Sanchez, M. Poesio, M. A. Kabadjov, and R. Tesar. 2006. What kind of problems do protein interactions raise for anaphora resolution? A preliminary analysis. In *Proc. of the 2nd SMBM 2006*, pages 109–112.