Detecting and Grounding Terms in Biomedical Literature

Kaarel Kaljurand, Fabio Rinaldi, Thomas Kappeler and Gerold Schneider

Institute of Computational Linguistics, University of Zurich
kalju@ifi.uzh.ch, rinaldi@ifi.uzh.ch, kappeler@bluewin.ch,
gschneid@ifi.uzh.ch

Abstract. We present an approach towards the automatic detection of names of proteins, genes, species, etc. in biomedical literature and their grounding to widely accepted identifiers. The annotation is based on a large term list that contains the common expression of the terms, a normalization step that matches the terms with their actual representation in the texts, and a disambiguation step that resolves the ambiguity of matched terms. We describe various characteristics of the terms found in existing term resources and of the terms that are used in biomedical texts. We evaluate our results against a corpus of manually annotated protein mentions and achieve a precision of 57% and recall of 72%.

1 Introduction

The complexity of biological organisms and the success of biological research in describing them, have resulted in a large body of biological entities (genes, proteins, species, etc.) to be indexed, named and analyzed. Probably the most important entities are proteins. They are an essential part of an organism and participate in every process within cells. Most proteins function in collaboration with other proteins, and one of the research goals in molecular biology is to identify which proteins interact.

While the number of different proteins is large, the amount of their possible interactions and combinations is even larger. In order to record such interactions and represent them in a structured way, human curators who work for knowledge base projects, e.g. Molecular INTeraction database (MINT)[1], Human Protein Reference Database (HPRD)[2], IntAct[3] (see [4] for a detailed overview), carefully analyze published biomedical articles. As the body of articles is growing rapidly, there is a need for effective automatic tools to help curators in their work. Such tools must be able to detect mentions of biological entities in the text and tag them with identifiers that have been assigned by existing knowledge bases. As the names that are used to reference the proteins can be very ambiguous, there is a need for an effective ambiguity resolution.

[1] http://mint.bio.uniroma2.it

[2] http://www.hprd.org/

[3] http://www.ebi.ac.uk/intact

© A. Gelbukh (Ed.)
Advances in Computational Linguistics.
Research in Computing Science 41, 2009, pp. 15-26

Received 10/11/08
Accepted 09/12/08
Final version 02/02/09

In this paper, we describe the task of automatically detecting names of proteins, genes, species, and experimental methods in biomedical literature and grounding them to widely accepted identifiers assigned by three different knowledge bases — UniProt Knowledgebase (UniProtKB)[4], National Center for Biotechnology Information (NCBI) Taxonomy[5], and Proteomics Standards Initiative (PSI) Molecular Interactions (MI) Ontology[6].

The term annotation uses a large term list that is compiled on the basis of the entity names extracted from the mentioned knowledge bases. This resulting list covers the common expression of the terms. A term normalization step is used to match the terms with their actual representation in the texts. Finally, a disambiguation step resolves the ambiguity (i.e. multiple IDs proposed by the annotator) of the matched terms.

The work presented in this paper is part of a larger effort undertaken in the OntoGene project[7] aimed at improving biomedical text mining through the usage of advanced natural language processing techniques. The results of the protein detection approach described in this paper feed directly into the process of identification of protein interactions. Our approach relies upon information delivered by a pipeline of NLP tools, including sentence splitting, tokenization, part of speech tagging, chunking, and a dependency-based syntactic analysis of candidate sentences [6]. The syntactic parser takes into account constituent boundaries defined by previously identified multi-word entities. Therefore the richness of the annotation process (including a variety of domain entities) has a direct beneficial impact on the performance of the parser, and thus leads to better recognition of interactions.

This paper is structured in the following way. In section 2 we describe the terminological resources that we have used, in section 3 we describe an automatic annotation of biomedical texts using these resources, in section 4 we describe the evaluation method and results, in section 5 we review related work, and finally, in section 6 we draw conclusions and describe future work.

2 Term Resources

2.1 Introduction

As a result of the rapidly growing information in the field of biology, the research community has realized the need for consistently organizing the discovered information — assign identifiers to biological entities, enumerate the names by which the entities are referred to, interlink different resources (e.g. existing knowledge bases and literature), etc. This has resulted in large and ever-growing knowledge bases (lists, ontologies, taxonomies) of various biological entities (genes, proteins, species, etc.). These resources can be treated as linguistic resources

[4] http://www.uniprot.org

[5] http://www.ncbi.nlm.nih.gov/Taxonomy/

[6] http://psidev.sourceforge.net/mi/psi-mi.obo

[7] http://www.ontogene.org

which can function as the basis of large term lists that can be used to annotate existing biomedical publications in order to identify the entities mentioned in these publications. In the following we describe three resources: UniProtKB, NCBI Taxonomy, and PSI-MI Ontology.

2.2 UniProtKB

The UniProt Knowledgebase (UniProtKB)[8] assigns identifiers to 397,539 proteins and describes their amino-acid sequences. The identifiers come in two forms: numeric accession numbers (e.g. `P04637`), and mnemonic identifiers that make visible the species that the protein originates from (e.g. `P53_HUMAN`). In the following we always use the mnemonic identifiers for better readability.

In addition to enumerating proteins, possible names used in the literature to refer to the proteins are listed in UniProtKB. UniProt sees as one of its functions to help with the standardization of protein nomenclature and thus tries to cover all the common ways of referring to a protein[9], while at the same time specifying a single name as "recommended name", following certain naming guidelines[10]. In addition, the names of functional domains and components of proteins, and also names of genes that encode the proteins are provided. The set of names covers names with large lexical difference (e.g. both 'Orexin' and 'Hypocretin' can refer to protein `OREX_HUMAN`), but usually not names with minor spelling variations (e.g. replacing a space with a hyphen). UniProtKB attempts to cover proteins of all species. The top five species ranked by the number of their different proteins are *Homo sapiens* (Human) with 20,325 proteins, *Mus musculus* (Mouse) with 15,915, *Rattus norvegicus* (Rat) with 7170, *Arabidopsis thaliana* (Mouse-ear cress) with 6970, and *Saccharomyces cerevisiae* (Baker's yeast) with 6553.[11]

We extracted 626,180 (different) names from the UniProtKB XML file, using the XPath expressions listed in table 1. The ambiguity of a name can be defined as the number of different UniProtKB entries that contain the name. UniProtKB names can be very ambiguous. This follows already from the naming guideline which states that "a recommended name should be, as far as possible, unique and attributed to all orthologs"[12]. Thus, a protein that is found in several similar species has one name but each of the species contributes a different ID. For UniProtKB, the average ambiguity is 2.61 IDs per name. If we discard the species labels, then the average ambiguity is 1.05 IDs. Ambiguous names (because the respective protein occurs in multiple species) are e.g. 'Cytochrome b' (1770 IDs), 'Ubiquinol-cytochrome-c reductase complex cytochrome b subunit' (1757), 'Cytochrome b-c1 complex subunit 3' (1757). Ambiguous names (without species

[8] We use the manually annotated and reviewed Swiss-Prot section of UniProtKB version 14, in its XML representation

[9] `http://www.uniprot.org/faq/9`

[10] `http://www.uniprot.org/docs/nameprot`

[11] The amount of proteins for a species reflects the amount of research done on the given species, rather than the amount of proteins that the species has.

[12] `http://www.uniprot.org/docs/nameprot`

Table 1. Frequency ranking of paths to XML elements that contain terms in UniProtKB.

Frequency	XPath (starting with `/uniprot/entry/`)
752,019	`gene/name`
397,539	`protein/recommendedName/fullName`
284,782	`protein/alternativeName/fullName`
90,397	`protein/recommendedName/shortName`
65,500	`protein/alternativeName/shortName`
16,400	`protein/component/recommendedName/fullName`
8913	`protein/domain/recommendedName/fullName`
6339	`protein/component/alternativeName/fullName`
5269	`protein/domain/alternativeName/fullName`
5023	`protein/component/recommendedName/shortName`
1416	`protein/CdAntigenName`
1207	`protein/domain/recommendedName/shortName`
1069	`protein/component/alternativeName/shortName`
787	`protein/domain/alternativeName/shortName`

labels) are e.g. ‘Capsid protein’ (103), ‘ORF1’ (97), ‘CA’ (88). Interestingly, very ambiguous names are not necessarily short, as is usually the case with ambiguous words.

Table 2 shows the orthographic/morphological properties of the names in UniProtKB in terms of how much certain types of characters influence the ambiguity. Non alphanumeric characters or change of case, while increasing ambiguity, influence the ambiguity relatively little. But as seen from the last column, digits matter a lot semantically. These findings motivate the normalization that we describe in section 3.2. Table 2 also shows the main cause for ambiguity of the names — the same name can refer to proteins in multiple species. While these proteins are identical in some sense (similar function or structure), the UniProtKB identifies them as different proteins.

Table 2. `ID_ORG` stands for the actual identifiers (which also include the species ID). `ID` stands for artificially created identifiers where we have dropped the qualification to the species. “Unchanged” = no change done to the terms; “No whitespace” = all whitespace is removed; “Alphanumeric” = everything but alphanumeric characters is removed; “Lowercase” = all characters are preserved but lowercased; “Alpha” = only letters are preserved.

	Unchanged	No whitespace	Alphanumeric	Lowercase	Alpha
`ID_ORG`	2.609	2.611	2.624	2.753	10.616
`ID`	1.049	1.050	1.053	1.058	4.145

2.3 NCBI Taxonomy

The National Center for Biotechnology Information provides a widely used resource called NCBI Taxonomy[13], which describes all known species and also lists the various forms of species names (e.g. latin names and common names). As explained in section 2.2, knowledge of these names is essential for effective disambiguation of protein names.

We compiled a term list on the basis of the taxonomy names list[14], but kept only names whose ID mapped to a UniProtKB species "mnemonic code" (such as `ARATH`)[15]. The resulting list has very little ambiguity (one example of an ambiguous term is 'mink' which can refer to both the European and the American Mink, which are classified as different species in the NCBI Taxonomy, and have therefore different identifiers).

The final list contains 31,733 entries where the species name is mapped to the UniProtKB mnemonic code. To this list, 8877 entries were added where the genus name is abbreviated to its initial (e.g. 'C. elegans') as names in such form were not included in the source data. These entries can be ambiguous in general (e.g. 'C. elegans' can refer to four different species), but are needed to account for such frequently occurring abbreviation in biomedical texts. Furthermore, six frequently occurring names that consist only of the genus name were added. In these cases, the name was mapped to a unique identifier (e.g. 'Arabidopsis' was mapped to `ARATH`), as it is expected that e.g. 'Arabidopsis' alone is always used to refer to *Arabidopsis thaliana*, and never to e.g. *Arabidopsis lyrata*.

2.4 PSI-MI Ontology

Proteomics Standards Initiative (PSI) Molecular Interactions (MI) Ontology[16] contains 2207 terms (referring to 2163 PSI-MI IDs) related to molecular interaction and methods of detecting such interactions (e.g. 'western blot', 'pull down'). There is almost no ambiguity in these names in the ontology itself. Several reasons motivate including the PSI-MI names in our term list. First, names of experimental methods are very frequent in biomedical texts. It is thus important to annotate such names as single units in order to make the syntactic analysis of the text more accurate. Second, in some cases a PSI-MI name contains a substring which happens to be a protein name (e.g. 'western blot' contains a UniProtKB term 'blot'). If the annotation program is not aware of this, then some tokens would be mistagged as protein names. Third, some PSI-MI terms overlap with UniProt terms, meaning that the corresponding proteins play an important function in protein interaction detection, but are not the subject of the actual interaction. An example of this is 'GFP' (PSI-MI ID `0367`, UniProtKB ID `GFP_AEQVI`), which occurs in sentences like "interaction between Pop2p and

[13] `http://www.ncbi.nlm.nih.gov/Taxonomy/`

[14] `ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz` (file `names.dmp`)

[15] `http://www.uniprot.org/help/taxonomy`

[16] `http://psidev.sourceforge.net/mi/psi-mi.obo`

GFP-Cdc18p was detected" where the reported interaction is between POP2 and CDC18, and GFP only "highlights" this interaction.

2.5 Compiled Term List

We compiled a term list of 1,679,483 terms based on the terms extracted from UniProtKB, NCBI, and PSI-MI. The term list has a simple 3-column format listing the term name, the term ID, and the term type in each entry. The type corresponds roughly to the resource the term originates from. For UniProtKB, there are two types — PROT and GEN — first assigned to all the terms from the path /uniprot/entry/protein/, and second to all the terms from /uniprot/entry/gene/. For NCBI, there are six types, distinguishing between common and scientific names, and the rank of the name in the taxonomy. For the PSI-MI Ontology terms there is just one type — MI. The frequency distribution of types is listed in table 3.

Table 3. Frequency distribution of types in the compiled term list.

Frequency	Type	Description
884,641	PROT	UniProtKB protein name
752,019	GEN	UniProtKB gene name
16,979	ocs	NCBI common name, species or below
8877	oss	NCBI scientific name, species or below
8877	ogs2	oss name, genus abbreviated (e.g. 'A. thaliana')
3316	oca	NCBI common name, above species
2561	osa	NCBI scientific name, above species
2207	MI	PSI-MI term
6	ogs1	NCBI selected genus name (e.g. 'Arabidopsis')
1,679,483		Total

In this list, 934,973 of the terms are multi-word units (e.g. 257,379 contain two tokens, 189,751 three tokens, and a few terms even 20 tokens). We did not normalize the names to any canonical representation nor generate all possible spelling variations of the names. One is expected to apply such processing during term annotation to account for differences in spacing, hyphenation etc. with respect to the terms actually occurring in the texts that undergo annotation.

3 Automatic Annotation of Terms

3.1 Introduction

Using the described term list, we can annotate biomedical texts in a straightforward way. First, the sentences and tokens are detected in the input text.

We use the LingPipe[17] tokenizer and sentence splitter which have already been trained on biomedical corpora. The tokenizer produces a granular set of tokens, e.g. words that contain a hyphen (such as 'Pop2p-Cdc18p') are split into several tokens, revealing the inner structure of such constructs which would e.g. allow to discover the interaction mention in "Pop2p-Cdc18p interaction". We slightly modified the sentence splitter to take into account abbreviations common in species names (e.g. 'sp.', 'subsp.').

The processing then proceeds by annotating the longest possible and non-overlapping sequences of tokens in each sentence, and in the case of success, assigns all the possible IDs (as found in the term list) to the annotated sequence. The annotator ignores certain common English function words (we use a list of ∼50 stop words), although, it is possible that some of them are UniProtKB terms. Also, figure and table references (e.g. 'Fig. 3a' and 'Table IV') are detected and ignored.

3.2 Normalization

In order to account for possible orthographic differences between the terms in the term list and the token sequences in the text, a normalization step is included in the annotation procedure. The same normalization is applied to the term list terms in the beginning of the annotation when the term list is read into memory, and to the tokens in the input text. In case the normalized strings match exactly then the input sequence is annotated with the IDs of the term list term. We currently apply the following normalization rules which were developed gradually over a training set. Many are based on similar rules reported in the literature, see e.g. [1,2,9].

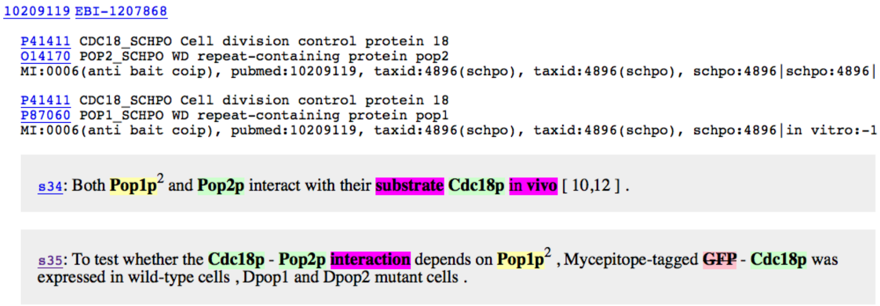
- Remove all characters that are not alphanumeric or space
- Normalize spaces, e.g. remove spaces between letters and numbers
- Normalize Greek letters, e.g. 'alpha' → 'a'
- Normalize Roman numerals, e.g. 'IV' → '4'
- Remove the final 'p' if it follows a number, e.g. 'Pan1p' → 'Pan1'
- Remove lowercase-uppercase distinction

In general, these rules increase the recall of term detection, but can lower the precision. For example, sometimes case distinction is used to denote the same protein in different species (e.g. according to UniProtKB, the gene name 'HOXB4' refers to `HXB4_HUMAN`, 'Hoxb4' to `HXB4_MOUSE`, and 'hoxb4' to `HXB4_XENLA`). However, the gain in recall seems to outweigh the loss of precision.

3.3 Disambiguation

A marked up term can be ambiguous for two reasons. First, the term can be assigned an ID from different term types, e.g. a UniProtKB ID and a PSI-MI

[17] `http://alias-i.com/lingpipe/`



```
10209119 EBI-1207868

  P41411 CDC18_SCHPO Cell division control protein 18
  O14170 POP2_SCHPO WD repeat-containing protein pop2
  MI:0006(anti bait coip), pubmed:10209119, taxid:4896(schpo), taxid:4896(schpo), schpo:4896|schpo:4896|

  P41411 CDC18_SCHPO Cell division control protein 18
  P87060 POP1_SCHPO WD repeat-containing protein pop1
  MI:0006(anti bait coip), pubmed:10209119, taxid:4896(schpo), taxid:4896(schpo), schpo:4896|in vitro:-1
```

s34: Both **Pop1p**^{2} and **Pop2p** interact with their **substrate** **Cdc18p** **in vivo** [10,12] .

s35: To test whether the **Cdc18p** - **Pop2p** **interaction** depends on **Pop1p**^{2} , Mycepitope-tagged ~~**GFP**~~ - **Cdc18p** was expressed in wild-type cells , Dpop1 and Dpop2 mutant cells .

Fig. 1. Visualization of the annotation results. Terms of different type are highlighted with a different background color. The terms that were rejected by the disambiguator are crossed out. For ambiguous terms, the number of different IDs is shown in the superscript. At the top of the screenshot, the actual interaction information is show. This information originates from the IntAct protein-protein interaction knowledge base.

ID. This situation does not occur often and usually happens with terms that are probably not interesting as protein mentions (such as ‘GFP’ discussed in section 2.4). We disambiguate such terms by removing all the UniProtKB IDs. (Similar filtering is performed in [8].) Second, the term can be assigned several IDs from a single type. This usually happens with UniProtKB terms and is typically due to the fact that the same protein occurs in many different species. Such protein names can be disambiguated in various ways. We have experimented with two different methods: (1) remove all the IDs that do not reference a species ID specified in a given list of species IDs; (2) remove all IDs that do not “agree” with the IDs of the other protein names in the same textual span (e.g. sentence, or paragraph) with respect to the species IDs.

For the first method, the required species ID list can be constructed in various ways, either automatically, on the basis of the text, e.g. by including species mentioned in the context of the protein mention, or by reusing external annotations of the article (e.g. it might be possible to exploit MeSH annotations). We are developing and evaluating separately an approach to the detection of species names mentioned in the article. The species mentions are used to create a ranked list, which will then be used to disambiguate other entities in the text, such as the protein mentions. This recently emerged task, which is sometimes called TX task (“Taxonomy task”), is attracting growing interest as a crucial task in biomedical text mining. Currently our experimental results in this task are above 70% F-Score.

The second method is motivated by the fact that according to the IntAct database, interacting proteins are usually from the same species: less than 2% of the listed interactions have different interacting species. Assuming that proteins that are mentioned in close proximity often constitute a mention of interaction, we can implement a simple disambiguation method: for every protein mention, the disambiguator removes every UniProtKB ID that references a species that is not among the species referenced by the IDs of the neighboring protein mentions. Only in case the intersection of proposed species is empty, should all the IDs be kept — this would cover the case when the textual span contains unambiguous protein mentions which do not agree with each other with respect to their species. The neighborhood can be defined to be a textual unit such as a phrase, sentence, paragraph, etc. We currently use a sentence as the unit, as sentence splitting information is easily obtained from our linguistic pre-processing. We note that this form of disambiguation might be better applied after syntactic analysis, when we have a more granular information about potentially interacting proteins. For example, after syntactic analysis, the textual span that constitutes the neighborhood can be defined to be a relative clause or a predicate-argument structure.

It should be noted that the disambiguation result is not always a single IDs, but often just a reduced set of IDs which must be disambiguated by a possible subsequent step. Also, it can happen that none of the IDs matches a listed species. In this case all the IDs are removed. Thus, the disambiguation step can revert the decision made by the annotation step.

4 Evaluation

We evaluated the accuracy of our automatic protein name detection and grounding method on a corpus provided by the IntAct project[18]. This corpus contains a set of 6198 short textual snippets (of 1 to about 3 sentences), where each snippet is mapped to a PubMed identifier (referring to the article the snippet originates from), and an IntAct interaction identifier (referring to the interaction that the snippet describes). In other words, each snippet is a "textual evidence" that has allowed the curator to record a new interaction in the IntAct knowledge base. By resolving an interaction ID, we can generate a set of IDs of interacting proteins and a set of species involved in the interaction, for the given snippet. Using the PubMed identifiers, we can generate the same information for each mentioned article. By comparing the sets of protein IDs reported by the IntAct corpus providers, and the sets of protein IDs proposed by our tool, we can calculate the precision and recall values.

We annotated the complete IntAct corpus by marking up token sequences that the normalization step matched with an entry in the term list. Each resulting annotation includes a set of IDs which was further reduced by the two disambiguation methods described in 3.3, i.e. some or all of the IDs were removed. Figure 1 shows the visualization of the annotation output on IntAct snippets together with the actual interaction as specified in IntAct.

Results before and after disambiguation are presented in table 4. The results show a relatively high recall which decreases after the disambiguation. This

[18] `ftp://ftp.ebi.ac.uk/pub/databases/intact/current/various/data-mining/`

Table 4. Results obtained on the IntAct snippets, with various forms of disambiguation, measured against PubMed IDs. The evaluation was performed on the complete IntAct data (*all*), and on a 5 times smaller fragment of IntAct (*subset*) for which we automatically extracted the species information. Three forms of disambiguation were applied: IntAct = species lists from IntAct data; TX = species lists from our automatic species detection; span = the species of neighboring protein mentions must match. Additionally, combinations of these were tested: e.g. IntAct & span = IntAct disambiguation followed by span disambiguation. The best results in each category are in boldface.

Disamb. method	Corpus	Precision	Recall	F-Score	True pos.	False pos.	False neg.
No disamb.	all	0.03	0.73	0.05	2237	81,662	848
IntAct	all	0.56	0.73	0.63	2183	1713	804
span	all	0.03	0.71	0.06	2186	68,026	899
IntAct & span	all	0.57	0.72	**0.64**	2147	1599	840
span & IntAct	all	0.57	0.72	0.64	2142	1631	821
No disamb.	subset	0.02	0.69	0.04	424	20,344	188
IntAct	subset	0.51	0.71	0.59	414	397	170
span	subset	0.02	0.67	0.05	407	16,319	205
IntAct & span	subset	0.53	0.69	**0.60**	404	363	180
span & IntAct	subset	0.52	0.69	0.59	399	369	177
TX	subset	0.42	0.59	0.49	340	478	241
TX & span	subset	0.43	0.57	**0.49**	332	445	249
span & TX	subset	0.42	0.57	0.48	329	457	244

change is small however, compared to the gain in precision. False negatives are typically caused by missing names in UniProtKB, or sometimes because the normalization step fails to detect a spelling variation. A certain amount of false positives cannot be avoided due to the setup of task. The tool is designed to annotate all proteins contained in the sentences, but not all of them necessarily participate in interactions, and thus are not reported in the IntAct corpus.

5 Related Work

There is a large body of work in named entity recognition in biomedical texts. Mostly this work does not cover grounding the detected named entities to existing knowledge base identifiers. Recently, however, as a result of the BioCreative workshop, more approaches are extending from just detecting entity mentions to "normalizing" of the terms. In general, such normalization handles gene names (by grounding them to EntrezGene[19] identifiers). [5] gives an overview of the BioCreative II gene normalization task.

A method of protein name grounding is described in [9]. It uses a rule-based approach that integrates a machine-learning based species tagger to disambiguate protein IDs. The reported results are similar to ours.

[19] http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene

There exist also two publicly available systems that return annotations together with UniProtKB identifiers. In the BioCreative Meta Server (BCMS)[20] [3], 2 out of 13 gene/protein taggers annotate using UniProtKB protein identifiers. The Whatizit[21] webservice annotates input texts with UniProtKB, Gene Ontology[22], and NCBI terms. A preliminary comparison showed that our approach gives results of similar quality.

6 Conclusions and Future Work

The main goal of the work described in this paper is to reliably identify protein mentions in order to identify protein-protein interactions in a subsequent processing step. We propose a method that uses large term lists extracted from various sources, and a set of normalization rules that match the token sequences in the input sentences against the term lists. Each matched term is assigned all the IDs that are possible for this term. The following disambiguation step tries to remove most of the IDs on the basis of the term context and knowledge about the species that the article discusses. The evaluation shows that a reasonably performing entity annotation system can be implemented in this way. For the evaluation, we have used the freely available IntAct corpus of snippets of textual evidence for protein-protein interactions. To our knowledge, this corpus has not been used in a similar evaluation before.

In the future, we would like to include more terminological resources in the annotation process. While the described three resources (UniProtKB, NCBI Taxonomy, PSI-MI Ontology) seem to contain the most important names used in biomedical texts, there exist also other names that are frequently used but that are not covered by these resources, e.g. cell line names (listed e.g. in CLKB [7]), names of certain chemical compounds, diseases, drugs, tissues.

We also intend to more conclusively evaluate our system against similar systems, such as BCMS and Whatizit.

Acknowledgements. This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1). Additional support is provided by Novartis Pharma AG, NITAS, Text Mining Services, CH-4002, Basel, Switzerland. The authors would like to thank the two anonymous reviewers of CICLing 2009 for their valuable feedback.

[20] http://bcms.bioinfo.cnio.es/

[21] http://www.ebi.ac.uk/webservices/whatizit/

[22] http://www.geneontology.org/

References

1. Jörg Hakenberg. What's in a gene name? Automated refinement of gene name dictionaries. In *Proceedings of BioNLP 2007: Biological, Translational, and Clinical Language Processing; Prague, Czech Republic*, 2007.
2. Jörg Hakenberg, Conrad Plake, Loic Royer, Hendrik Strobelt, Ulf Leser, and Michael Schroeder. Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome Biology*, 9(Suppl 2):S14, 2008.
3. F Leitner, M Krallinger, C Rodriguez-Penagos, J Hakenberg, C Plake, C-J Kuo, C-N Hsu, RT-H Tsai, H-C Hung, WW Lau, CA Johnson, R Saetre, K Yoshida, YH Chen, S Kim, S-Y Shin, B-T Zhang, WA Baumgartner, L Hunter, B Haddow, M Matthews, X Wang, P Ruch, F Ehrler, A Ozgur, G Erkan, DR Radev, M Krauthammer, T Luong, and R Hoffmann. Introducing meta-services for biomedical information extraction. *Genome Biology*, 9(Suppl 2):S6, 2008.
4. Suresh Mathivanan, Balamurugan Periaswamy, TKB Gandhi, Kumaran Kandasamy, Shubha Suresh, Riaz Mohmood, YL Ramachandra, and Akhilesh Pandey. An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics*, 7(Suppl 5):S19, 2006.
5. AA Morgan, Z Lu, X Wang, AM Cohen, J Fluck, P Ruch, A Divoli, K Fundel, R Leaman, J Hakenberg, C Sun, H-h Liu, R Torres, M Krauthammer, WW Lau, H Liu, C-N Hsu, M Schuemie, KB Cohen, and L Hirschman. Overview of BioCreative II gene normalization. *Genome Biology*, 9(Suppl 2):S3, 2008.
6. Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon. OntoGene in BioCreative II. *Genome Biology*, 9(Suppl 2):S13, 2008.
7. Sirarat Sarntivijai, Alexander S. Ade, Brian D. Athey, and David J. States. A bioinformatics analysis of the cell line nomenclature. *Bioinformatics*, 24(23):2760–2766, 2008.
8. Lorraine Tanabe and W. John Wilbur. Tagging gene and protein names in biomedical text. *Bioinformatics*, 18(8):1124–1132, 2002.
9. Xinglong Wang. Rule-Based Protein Term Identification with Help from Automatic Species Tagging. In *Computational Linguistics and Intelligent Text Processing, 8th International Conference, CICLing 2007*, pages 288–298, Mexico City, Mexico, 2007.