Annotating and Evaluating Text for Stem Cell Research

Mariana Neves[1], Alexander Damaschun[2], Andreas Kurtz[2], Ulf Leser[1]

[1]Humboldt-Universität zu Berlin, Knowledge Management in Bioinformatics, Berlin, Germany,
[2]Berlin Brandenburg Center for Regenerative Therapies, Charité, Berlin, Germany
neves@informatik.hu-berlin.de, alexander.damaschun@charite.de,
andreas.kurtz@charite.de, leser@informatik.hu-berlin.de

Abstract

The regeneration of vital organs and tissues remains one of the biggest medical challenges. However, the use of embryonic stem cells and induced pluripotent stem cells allows novel replacement strategies. The CellFinder project aims to create a stem cell data repository by linking information from existing public databases and by performing text mining on the research literature. We present the first version of our corpus which is composed of 10 full text documents containing more than 2,100 sentences, 65,000 tokens and 5,200 annotations for entities. The corpus has been annotated with six types of entities (anatomical parts, cell components, cell lines, cell types, genes/protein and species) with an overall inter-annotator agreement around 80%. Preliminary results using baseline methods based on freely available terminologies and systems have returned a recall which ranges from 48% to 90% for the extraction of the named entities. The high distribution of entities which are representative of the stem cell research, specially cell types, makes our corpus a valuable resource for the stem cell domain.

Keywords: biomedical corpus, names-entity recognition, stem cell research.

1. Introduction

The regeneration of lost vital organ and tissue function after severe injury or end-stage progression of diseases remains one of the biggest unmet medical challenges (Viswanathan and Keating, 2011). Despite pharmacological advances in alleviating the symptoms of compromised vital functions or in slowing disease progression, the only available therapy for permanent impairment or organ loss is organ replacement. However, since there are few indications for which sufficient numbers of donors exist in order to meet the demand for transplant organs (Watson and Dark, 2012), alternative strategies are needed to restore organ and tissue function.

The advent of human embryonic stems cell (hESCs) (Thomson et al., 1998) and human induced pluripotent stem cells (hiPSCs) (Yu et al., 2007) together with the identification of many types of multipotent precursor and adult stem cell (Barile et al., 2011) have opened promising new routes for novel replacement strategies (Atala, 2012). Some of these approaches aim to activate the body's endogenous regenerative capacities, others look at the stem cells' capacity to differentiate into specific cell types for direct application in cell therapy or use them as building blocks in tissue engineering.

All regenerative approaches involving stem cells or their differentiated progeny have one fundamental requirement in common: The cells to be used have to be both effective and safe. Therefore, therapeutic cell populations to be applied in the patient or in engineered tissue have to be well characterized based on reliable measurement and analysis techniques as well as validated by knowledge bases of stem cells and their progeny (Wohlers et al., 2009; Kerrigan and Nims, 2011).

Results of such studies create an ever-rising flood of scientific information and experimental data that is virtually impossible to be registered, analyzed or exploited without the aid of sophisticated bioinformatics applications running on powerful computer infrastructures. This is particularly evident in the rising field of regenerative medicine, in which several specialized scientific disciplines are combined (Viswanathan and Keating, 2011). Alongside clinical cell-directed pathological and cytological data, additional information such as cell-anatomical, cell-biological, genetic and biochemical data as well as potencies and functional interactions are required for the modeling, prediction and analysis of cell-based therapies, as well as for basic cell research. Consequently regenerative medicine will not progress without an integrating, systematic and analytical approach that utilizes adequate shared information and data resources (Hatano et al., 2011; Jung et al., 2010).

CellFinder[1] is based on the idea of establishing a central stem cell data repository, by utilizing and interlinking existing public databases regarding defined areas of human pluripotent stem cell research. Provision of standardized description, registration and interlinking of stem cell data on the above mentioned levels is a prerequisite for the effective exchange of data. One specific aim of CellFinder is to identify processes by which various kinds of stem and precursor cells may differentiate, function and react and subsequently be applied. An important source of knowledge are published research results. In CellFinder, text mining methods are employed to extract knowledge from this scientific literature, which will be further made available in our on-line repository.

In the last years, we have observed an increase in the availability of corpora for the biomedical domain (Kim et al., 2003; Pyysalo et al., 2008). In the last 10 years, the biomedical natural language community has migrated from sentences-based corpora annotated with one or a couple of

[1]http://www.cell-finder.org/

Sections	Sentences	Tokens	Annotations						
			Anatomy	Cell Comp.	Cell Line	Cell Type	Gene	Species	TOTAL
Abstract	79	2683	88	6	10	151	45	24	324
Introduction	225	6881	155	17	7	302	56	59	596
Methods	539	15540	130	64	101	228	356	109	988
Results	1052	31975	423	91	187	832	1036	191	2760
Discussion	256	7221	99	20	15	245	112	47	538
Conclusion	26	731	18	0	8	19	16	8	69
TOTAL	2177	65031	913	198	328	1777	1621	438	5275

Table 1: Number of sentences, tokens and annotations per entity and per section in the full text document. A total per type of entity and per sections is shown in the last line and last column, respectively.

named-entities (Rosario and Hearst, 2004; Tanabe et al., 2005) to the annotation of abstracts with more than one type of entity (Kim et al., 2003; Klinger et al., 2008; Furlong et al., 2008), relationships between entities (Pyysalo et al., 2008) and biological events (Kim et al., 2008). More recently, also full texts have become popular (Kim et al., 2011; Carreira et al., 2011). Finally, the community-based effort for the construction of the CALBC silver standard corpus composed by a variety of entities (Rebholz-Schuhmann et al., 2010) is certainly helpful for the biomedical natural language processing research.

Studies have shown that the structures of abstract and full text are different (Cohen et al., 2010) and that more valuable information is usually found only in full texts. We have indeed noticed that the data which is relevant for the CellFinder's database is usually present only in the results sections of the publications. Therefore, in order to support the development and evaluation of our text mining methods, some selected full text documents have been annotated with entities and biological processes relevant for the stem cell domain. The annotation schema includes a variety of entities, such as cell lines, anatomical parts and genes/proteins, as well as biological events, such as gene expression and differentiation. We present here the first version of our corpus which is composed of 10 full text documents comprising 2,177 sentences, 65,031 tokens and 5,275 annotations of entities.

2. Overview of the Corpus

We present ongoing work which aims at annotating a corpus on the stem cell domain with semantic entities, biological events as well as associated meta-knowledge (Thompson et al., 2011). Our annotation schema consists of six types of entities:

- anatomical parts (i.e., tissues, organs and body parts): "bone marrow", "adipose tissue";
- cell components: "membrane", "chromosome", "nuclei";
- cell lines: "hESMPC9.1", "H1";
- cell types: "mesenchymal precursors", "skeletal muscle cells";
- genes/proteins: "OCT4", "vimentin";
- species: "human", "mouse".

The importance of each of these entities in the stem cell research is evident. We now give a more detailed description of each of them.

Anatomical parts entities describe the spacio-temporal locations of cell types throughout their existence/development in tissues, organs (and part thereof), body parts and organisms. The annotation of species is necessary in order to map homologies between different organisms and to transfer insights from established animal models to the human organisms and vice versa. Anatomy has also been applied to in vitro anatomies formed by cells, e.g. embryoid bodies, monolayers or rosettes.

Cell component refers to sub-cellular structures or locations within a cell (sometimes specific to a certain cell type) where genetic functions are exerted, proteins are expressed or molecules are detected.

Cell lines describe instances of cells of a certain type that have been modified with biomolecular, genetic, chemical or physical techniques in order to preserve one or several properties of their specific type or to arrest the cells in a certain stage of their development. This enables the cells to be cultivated reproducibly over prolonged periods of time (compared with the naive status), or, in the case of immortalized cells, indefinitely. Designations for cell lines are commonly arbitrary and originate from their providers.

Cell type encompasses all instances of a biological cell (individual cells, colonies or agglomerations in biological tissue) with a distinct set of morphological, biomolecular and functional properties. With the exception of terminally differentiated adult/somatic cells, all cells of a certain type have at least one precursor cell type and at least one progeny cell type.

Genes or proteins refers to instances in the text that mention gene names and functions, RNA that has been transcribed from any particular gene or a protein that has been expressed as the result of gene (up)regulation. The same is true for any mention of the absence of a protein (or its expression) or the suppression of a gene.

Annotations have been performed by two experts from the stem cell domain. Annotator 1 is a biologist with extensive expertise in molecular and stem cell biology (mesenchymal stem cell, hESC, hiPSC), cell generation, characterization and GXP manufacture, systems biology, state-of-the-art analysis techniques, clinical studies and biomedical

ethics. Annotator 2 is a biotechnologist with long-years expertise in stem cell characterization and registration, cell-based knowledge bases and dissemination and tissue engineering.

In this first round of annotations, 10 full text documents have been annotated. Papers have been selected based on the work of (Löser et al., 2010) in which publications on the field of human embryonic stem cells have been surveyed (up to November of 2009). A list of 990 publications have been derived from this work (available as supplementary material). From this list, 62 are included in the PubMed Central Open Access Subset, and thus, can be freely used for text mining purposes. Our annotators have selected 10 full papers for the annotation, namely PMIDs: 16316465, 17381551, 17389645, 18162134, 18286199, 15971941, 16623949, 16672070, 17288595 and 17967047. Full texts were obtained in XML format from the Pubmed Central Open Access Subset page[2]. For performing the annotations, we used Brat[3] (brat rapid annotation tool) (Stenetorp et al., 2012). The documents had to be split into sections due to the compromised performance of Brat when dealing with long documents. The number of annotations found for each entity type in the various sections of the full paper is shown in Table 1. An example of some of the annotations for the six entity types is shown in Figure 1.

The gold-standard corpus was created by merging the annotations from both annotators. An automatic consensus was carried out to remove overlapping annotations, such as singular and plural forms (e.g., "stem cell" and "stem cells") and mentions starting with hyphens (e.g., "-H1.3" and "H1.3"). Additionally, we automatically checked those mentions which started or ended with parenthesis, curly or squared brackets, which are certainly due to a mistake when selecting the text of the mention. Finally, the documents were manually checked and some few overlapping inconsistencies have been corrected, such as "mesenchym" and "mesenchymal", by keeping only the larger one.

Entities	Exact	Overlap+Type	Overlap
Anatomy	0.37	0.61	0.76
Cell Comp.	0.33	0.39	0.49
Cell Line	0.75	0.92	0.95
Cell Type	0.30	0.85	0.91
Gene/Protein	0.77	0.81	0.83
Species	0.78	0.81	0.83
TOTAL	0.51	0.80	0.85

Table 2: F-score of the inter-annotator agreement for each of the entities.

The inter-annotator agreement (IAA) was computed as F-score and is shown in Table 2. Annotations which matched exactly regarding the span and the type of entity are shown as "Exact". Overlapping annotations which belonged to the same type of entity were also included in the consensus corpus as alternative synonyms. The agreement when considering these cases is found in Table 2 under the column "Overlap+Type". For instance, one of the annotators identified "human embryonic stem cells" as a cell type, while the other annotated just "stem cells". Alternative synonyms are not unusual in the biomedical domain, such as the GENETAG corpus (Tanabe et al., 2005) which includes synonyms for gene and protein names. We have also decided to add overlapping annotations which belong to different types of entities. For instance, for the same example above, when one of the annotators identified "human embryonic stem cells" as a cell type, the other one annotated "human" as a species and "embryonic" as anatomical part. We show the increment in the IAA when allowing overlapping for annotations of different types in the "Overlap" column of Table 2. Finally, as an ongoing work, entities which have been annotated by only one of the annotators were integrated into the gold standard in this phase of the project.

The corpus was made available in our repository of corpora[4] in its full text version and also split by sections. For the visualization of the corpus, we recommend Chrome, Safari or Opera. The full text and the sections-split versions of the corpus are available for downloading from the corpus web page[5] in the standoff format used by Brat and in the XML format used in (Pyysalo et al., 2008).

3. Preliminary Evaluation

In this section, we present our preliminary results for predicting entities in the stem cell domain. Regarding the recognition of the entities annotated in this corpus, we are more concerned about the recall. If a certain entity cannot be found in the text during the named-entity recognition step, the events in which it participates will not be found either.

As baseline, we decided to use only dictionary-based methods derived from existing ontologies or terminologies and freely available systems. Thus, we did not use the annotated corpus to train a specific tagger for any of the entity types. Details for the methods used in the recognition of each entity type are presented below.

We used Metamap (Aronson and Lang, 2010) for extracting annotations for five entity types: anatomical parts, cell components, cell type, gene/protein and species. We restricted the annotations to certain semantic types using the "-J" parameter. The mapping of the semantic types to our entities is shown in Table 3. Additionally, we used the parameter which allows variants for acronyms and abbreviations ("-a"). As Metamap does not work properly with long texts, we split the full text documents into sentences using the sentence detector available in OpenNLP[6]. Additionally, also due to the inability of Metamap in processing long sentences, we only analyzed those under 1000 characters, resulting in five discarded sentences.

Anatomical parts and cell components were extracted only using Metamap. For cell types, besides Metamap, a dictionary of cell type synonyms was created using the OBO Cell Type ontology (Bard et al., 2005). We used Lingpipe

[2]http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[3]http://brat.nlplab.org

[4]http://corpora.informatik.hu-berlin.de/index.xhtml

[5]http://www.informatik.hu-berlin.de/forschung/gebiete/wbi/resources/cellfinder/

[6]http://incubator.apache.org/opennlp/

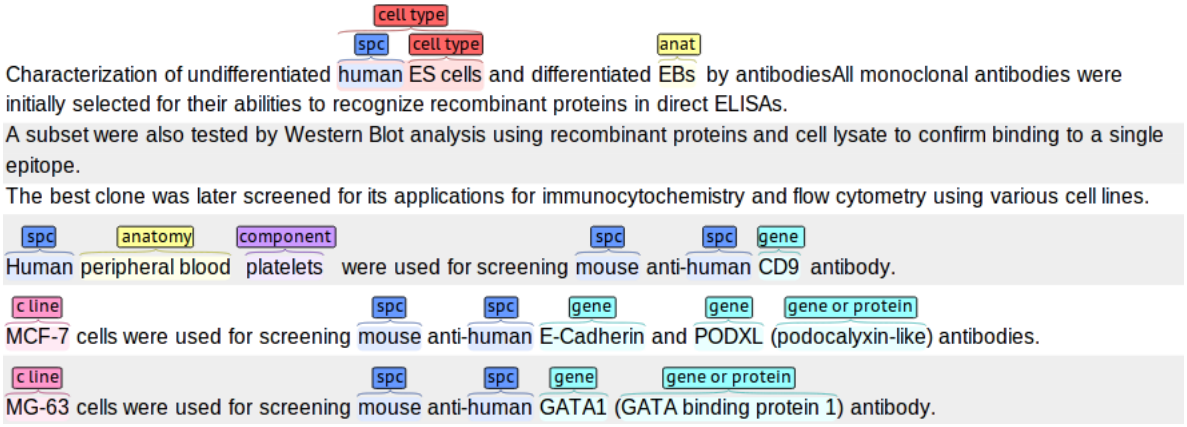

Figure 1: Passage of document "16316465" shows annotations for our six entity types. The following colors and abbreviations are used: "anat" or "anatomy" (yellow) for anatomical parts, "cell type" (red) for cell types, "spc" (dark blue) for species, "component" (purple) for cell components, "gene" and "gene or protein" (light blue) for genes and proteins and "c line" (rose) for cell lines. Visualization of the corpus is provided using Brat annotation tool.

named-entity recognition procedures[7] for case-insensitive matching of the synonyms to the text.

Regarding cell lines, we created a dictionary of synonyms by merging names of cell lines from three different sources: hESCReg (Borstlap et al., 2008), a list of human embryonic cell lines presented as supplementary material in (Löser et al., 2010) and data available on-line in the Cell Line Data Base (Romano et al., 2009). Variations for the cell line synonyms were automatically generated (e.g., "CCTL-6", "CCTL 6", "CCTL6"). The derived dictionary of synonyms was also matched to the text using Lingpipe.

Besides Metamap, genes were extracted using GNAT (Hakenberg et al., 2008; Hakenberg et al., 2011) configured with the default model. We defined a threshold score of 0.2 for dismissing potential false positives. Finally, mentions for species were extracted using Linnaeus tool (Gerner et al., 2010), besides Metamap.

As discussed before, we have not yet made use of our corpus for training specific taggers for the recognition of any type of entity. Instead, we tried to use freely available systems, terminologies, databases and ontologies. For evaluation of our baseline methods, the corpus was split in two groups, five for development (16316465, 17381551, 17389645, 18162134, 18286199) and five for testing (15971941, 16623949, 16672070, 17288595, 17967047). We have used the development dataset for the error analysis while we kept the other dataset for a blind test.

The evaluation results for each entity type for both datasets are presented in Table 4. Results refer only to the recognition of the mentions, with no normalization of entities. We present results for exact matching, i.e., the exact mention as well as the exact type of entity, and for a more flexible strategy in which we consider also as a correct match any overlapping mention belonging to the same type.

Entities	Semantic groups
Anatomy	"Anatomical Structure", "Body Location or Region", "Body Part, Organ, or Organ Component", "Body Space or Junction", "Body Substance", "Body System", "Embryonic Structure", "Fully Formed Anatomical Structure", "Tissue"
Cell Comp.	"Cell Component", "Nucleic Acid, Nucleoside, or Nucleotide"
Cell Type	"Cell"
Gene/protein	"Amino Acid, Peptide, or Protein", "Enzyme", "Receptor", "Amino Acid Sequence", "Carbohydrate Sequence", "Gene or Genome", "Molecular Sequence", "Nucleic Acid, Nucleoside, or Nucleotide", "Nucleotide Sequence"
Species	"Amphibian", "Animal", "Archaeon", "Bacterium", "Bird", "Eukaryote", "Family Group", "Fish", "Fungus", "Group", "Human", "Mammal", "Organism", "Plant", "Population Group", "Reptile", "Vertebrate", "Virus"

Table 3: Semantic types which have been considered for each entity when using Metamap.

4. Discussion

In this work we have presented a first version of our ongoing corpus developed for the CellFinder project. We have annotated more than 5,200 annotations for a corpus of more than 65,000 tokens, which makes the density of our corpus about 8%. This is a satisfactory density provided that the named entities were usually annotated in the context of biological events. Additionally, no text mining has been performed in the documents before its manual annotations, i.e., the annotators have worked over texts free of any pre-annotations. Finally, although Brat provides a way of querying on-line resources during the annotation (e.g., EntrezGene or Uniprot), it does not support for terminologies and ontologies, which might increase the density of annotations as well as the agreement among annotators.

The density of our corpus is comparable to other full text corpora. For instance, the 14 full papers belonging to the Genia Event Task from the BioNLP Shared Task 2011 also

[7] http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.html

Entities	Development				Test			
	Exact matches		Overlapping matches		Exact matches		Overlapping matches	
	Recall	F-score	Recall	F-score	Recall	F-score	Recall	F-score
Anatomy	0.30	0.23	0.48	0.33	0.32	0.29	0.48	0.41
Cell Comp.	0.55	0.31	0.67	0.38	0.75	0.17	0.80	0.19
Cell Line	0.48	0.35	0.48	0.35	0.43	0.43	0.62	0.59
Cell Type	0.46	0.42	0.66	0.59	0.37	0.36	0.57	0.55
Gene/Protein	0.68	0.36	0.78	0.44	0.77	0.29	0.90	0.35
Species	0.90	0.46	0.93	0.48	0.83	0.47	0.86	0.49
TOTAL	0.56	0.31	0.69	0.38	0.50	0.28	0.67	0.37

Table 4: Recall and F-score of each entity type for the development and testing datasets. We present results when evaluating using a exact matching and when allowing overlapping of the annotations.

hold a density of about 8% for the proteins. A corpus on microbial cellular responses reported in the work of (Carreira et al., 2011) contains 130 full text documents annotated with 59,000 annotations of biomedical concepts. Although the number of tokens has not to be provided in the publication, its density seems to be lower that 10%. Finally, in the CRAFT corpus (Cohen et al., 2011), which comprise 97 full text documents, annotations have been performed using available ontologies, such as NCBI Taxonomy or Gene Ontology. It contains 597,000 tokens and 118,783 annotations, i.e, a density of almost 20%.

Regarding the disagreement between our annotators, 80% to 85% is also considered satisfactory for the biomedical domain. However, the distinction among anatomical parts, cell components and cell types still need to be discussed further in the next phase of our project. The overlapping among these entities is certainly due to the granularity of our annotation schema, specially on the anatomical level. By manually checking some of the annotations which were only performed by one of the annotators, we have noticed that they do not usually take part on the biological events, which is the final aim of our ongoing corpus. When compared to the CRAFT corpus (Bada et al., 2010), their inter-annotator agreements ranges from 70% to almost 100%, provided that they have performed various training sessions and that the annotation was supported by available ontologies. Our inter-annotator agreement is also comparable to the microbial cellular response corpus (Carreira et al., 2011) which ranges from 21% to 83% after three training cycles.

However, our corpus has some limitations. As already discussed, we have not used any available terminology or ontology while performing the annotations. Therefore, we only provide text mentions, without any association to an identifier.

Being an ongoing project, an extension of the corpus is already being carried out. We have started the annotation of biological events relevant to the stem cell research, such as cell differentiation and gene expression in cells and in anatomical parts. These are valuable information which we plan to make available to the scientific community in our CellFinder project's database, along with the respective bibliographic reference.

We are also proceeding to the annotation of meta-knowledge according to the work of (Thompson et al., 2011). This information is of great importance regarding the reliability of the data being extracted, whether it describes the existence or not of a certain biological process, its intensity (high or low) and the primary publication for finding further information, which is essential when associating data in CellFinder to its respective publication.

Finally, we also intend to annotate a larger number of abstracts in order to have more diversity of entities and biological events. A larger corpus is also usually necessary for training and evaluating machine learning methods for extracting entities or biological events. Our preliminary results show that training a classifier might be necessary at least for the extraction of cell lines (as discussed below).

Regarding the methods and the evaluation presented here, we performed a brief analysis of the errors for all six entities. This analysis was performed only on the five documents belonging to the development corpus. As we are more concerned about the recall of the system, we focused our error analysis on the false negatives. A discussion of the mistakes is presented below for each entity type.

When performing an extra evaluation and allowing overlapping mentions between different types of entities, the recall for the anatomical parts increases from 48% to 65% (result not shown) for the development dataset. Most of these new matches are with annotations which have been extracted by Metamap as cell types, such as "neural" or "myotubes". However, using only Metamap seems to be not enough for achieving a satisfactory recall. Alternative tools, which we plan to use in the next phase, include the recent work of (Pyysalo et al., 2011) on the recognition of anatomical entities using open biomedical ontologies. We are also aware that we cannot expect a high recall from the existing available tools for those entity types which still have a low agreement among the annotators, such as anatomical parts, cell components and cell types.

Our recall is also not enough for the recognition of cell types, whose extraction is based on Metamap and the Cell Type ontology. However, 64% of our false negatives correspond only to the plural forms of common abbreviations in the stem cell types, such as "hNSC" (human neural stem cells) and "hESCs" (human embryonic stem cells). The use of abbreviation resolution methods (Schwartz and Hearst, 2003) in the next phase of the project may help to overcome

this problem.

However, Metamap returns a high recall when used for extracting cell components. We consider the recall of 67%-80% as satisfactory for an entity type which plays a secondary role in our annotation schema, as it is not usually associated to a biological event. Additionally, the mapping of the semantic type "Nucleic Acid, Nucleoside, or Nucleotide" to this entity type has increased its recall from 44% (result not shown) to 67% for the development dataset, due to the recognition of annotations such as "DNA", "cDNA" and "mRNA".

Likewise, Linnaeus and Metamap perform very well when extracting species, as they provide a recall of 93% and 86% for the development and test datasets, respectively. Although Linnaeus might be enough for retrieving species, we also consider Metamap because it increased the recall for the test dataset from 82% (result not shown) to 86%. The mentions that are missed are mostly due to problems in the parsing of tables, when columns are concatenated into a single token, such as "hPODXLYesNoNoMouse".

On the other hand, the recall for one of the most important entities in our annotation schema, the cell lines, is still rather low, and about half of the annotations are missed. The two more frequent false negative mentions are "SD56" and "NTERA-2". The first one is not present in any of the three dictionaries. Regarding the "NTERA-2", other cell lines related to it could be found in one of our dictionaries (Cell Line Database) as "NTERA-2 clone D1", but it could not be matched using just a case-insensitive matching strategy. For the cell lines, our baseline approach, which considers only freely available dictionaries and tools, does not seem to address the diversity of the nomenclature. The lack of an integrated cell line database frustrates the hopes of having a more complete terminology of cell line names. The use of a machine learning algorithm trained with some of our annotated documents seems to be inevitable in next phases of the project.

Finally, regarding the extraction of genes and proteins, our recall still need to be improved, as 78% (development dataset) might not be enough for an entity which directly participates in many biological events. On the other hand, GNAT and Metamap have achieved the highest recall for all entities for the test dataset. Surprisingly, Metamap increased the recall from 75% (results not shown) to 90% for the test dataset. By analyzing the false negatives, most of them are never found by GNAT (e.g., "eMyHC" and "TuJ1"). However, some mentions have been missed due to the same problem experienced by Linnaeus, i.e., due to the parsing of the tables, such as the token "hSOX17NoYesYesMs" which contains the gene "SOX17". We plan to try some additional available tools for the extraction and normalization of genes and entities, such as ABNER (Settles, 2005), BANNER (Leaman and Gonzalez, 2008), GeneTuKit (Huang et al., 2011), as well as other resources discussed in (Kabiljo et al., 2009).

In our curation process for the CellFinder project, data extracted using text mining methods will be validated by experts before being included into the database. Therefore, we expect false negatives to be curated manually and false positives to be dismissed or corrected by the curators. Nevertheless, a more precise and high-recall text mining approach will certainly reduce the human effort in the validation step.

Regarding the limitations of our methods and evaluation, as discussed before, our corpus only provides the textual mentions for the annotations. Therefore, we did not consider the normalization of the entities in this phase of the project. However, when mapping data extracted using text mining methods to CellFinder's database, which is completely based on ontologies, the availability of an identifier associated to each entity will become an important issue.

5. Conclusion

In this work we have presented the first version of the corpus which has been annotated in the scope of the CellFinder project. This is an ongoing work which aims to annotate biological processes relevant to the stem cell research. This first version of the corpus includes annotations for six types of semantical entities: anatomical parts (e.g., tissues and organs), cell components, cell lines, cell types, genes/proteins and species. This corpus is composed of 10 full papers which contain around 65,000 tokens and more than 5,200 annotations with an inter-annotator agreement around 80% to 85%. We hope it can be a valuable resource for the stem cell research as well as for evaluation of named-entity recognition methods for a variety of entities.

We have also presented here our baseline methods for the prediction of the entities present in the corpus. We have used only freely available systems, terminologies and ontologies. We have obtained a recall which ranges from 48% to 90%, depending on the entity type. Although some improvements are still necessary regarding the agreement between the annotators and the text mining methods for the prediction of the annotations, the work presented here is promising. We believe that this is a unique corpus in the stem cell domain and the data extracted using literature mining will be a valuable source of information once available in CellFinder's database.

6. Acknowledgements

Mariana Neves was supported by DFG (LE 1428/3-1). We thank Dr. Peter Löser (Robert Koch Institute) and his assistants for the support in the selection of relevant publications for the corpus. We thank Pontus Stenetorp (University of Tokyo) and Sampo Pyysalo (NaCTeM and University of Manchester) for the support with Brat annotation tool and Illés Solt (Budapest University of Technology) for helping in the installation of the tool. We also thank Philippe Thomas and Michael Weidlich, our colleagues in the Humboldt-Universität zu Berlin, for their helpful feedback and support with text mining tools. Finally, we want to acknowledge the other members of the CellFinder project (Jean-Fred Fontaine, Fritz Lekschas, Harald Stachelscheid and Stefanie Werner) for their valuable contribution to this publication.

7. References

Alan R Aronson and Franois-Michel Lang. 2010. An overview of metamap: historical perspective and recent

advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Anthony Atala. 2012. Regenerative medicine strategies. *Journal of Pediatric Surgery*, 47(1):17 – 28.

Michael Bada, Lawrence E. Hunter, Miriam Eckert, and Martha Palmer. 2010. An overview of the craft concept annotation guidelines. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 207–211, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jonathan Bard, Seung Rhee, and Michael Ashburner. 2005. An ontology for cell types. *Genome Biology*, 6(2):R21.

Lucio Barile, Claudia Altomare, and Antonio Zaza. 2011. Induced pluripotent stem cells: progress towards a biomedical application. *Expert Review of Cardiovascular Therapy*, 9(10):1265–1269.

Joeri Borstlap, Glyn Stacey, Andreas Kurtz, Anja Elstner, Alexander Damaschun, Begoa Arn, and Anna Veiga. 2008. First evaluation of the european hescreg. *Nature Biotechnology*, 26:859 – 860.

Rafael Carreira, Sonia Carneiro, Rui Pereira, Miguel Rocha, Isabel Rocha, Eugenio Ferreira, and Analia Lourenco. 2011. Semantic annotation of biological concepts interplaying microbial cellular responses. *BMC Bioinformatics*, 12(1):460.

K Bretonnel Cohen, Helen Johnson, Karin Verspoor, Christophe Roeder, and Lawrence Hunter. 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11(1):492.

K. Bretonnel Cohen, Tom Christiansen, William A. Baumgartner, Jr., Karin Verspoor, and Lawrence E. Hunter. 2011. Fast and simple semantic class assignment for biomedical text. In *Proceedings of BioNLP 2011 Workshop*, BioNLP '11, pages 38–45, Stroudsburg, PA, USA. Association for Computational Linguistics.

Laura Furlong, Holger Dach, Martin Hofmann-Apitius, and Ferran Sanz. 2008. Osirisv1.2: A named entity recognition system for sequence variants of genes in biomedical literature. *BMC Bioinformatics*, 9(1):84.

Martin Gerner, Goran Nenadic, and Casey Bergman. 2010. Linnaeus: A species name identification system for biomedical literature. *BMC Bioinformatics*, 11(1):85.

Jörg Hakenberg, Conrad Plake, Robert Leaman, Michael Schroeder, and Graciela Gonzalez. 2008. Inter-species normalization of gene mentions with gnat. *Bioinformatics*, 24(16):i126–i132.

Jörg Hakenberg, Martin Gerner, Maximilian Haeussler, Illés Solt, Conrad Plake, Michael Schroeder, Graciela Gonzalez, Goran Nenadic, and Casey M. Bergman. 2011. The gnat library for local and remote gene mention normalization. *Bioinformatics*, 27(19):2769–2771.

Akiko Hatano, Hirokazu Chiba, Harry Amri Moesa, Takeaki Taniguchi, Satoshi Nagaie, Koji Yamanegi, Takako Takai-Igarashi, Hiroshi Tanaka, and Wataru Fujibuchi. 2011. Cellpedia: a repository for human cell information for cell studies and differentiation analyses. *Database*, 2011.

Minlie Huang, Jingchen Liu, and Xiaoyan Zhu. 2011. Genetukit: a software for document-level gene normalization. *Bioinformatics*, 27(7):1032–1033.

Marc Jung, Hedi Peterson, Lukas Chavez, Pascal Kahlem, Hans Lehrach, Jaak Vilo, and James Adjaye. 2010. A data integration approach to mapping oct4 gene regulatory networks operative in embryonic stem cells and embryonal carcinoma cells. *PLoS ONE*, 5(5):e10709, 05.

Renata Kabiljo, Andrew Clegg, and Adrian Shepherd. 2009. A realistic assessment of methods for extracting gene/protein interactions from free text. *BMC Bioinformatics*, 10(1):233.

Liz Kerrigan and Raymond W. Nims. 2011. Authentication of human cell-based products: the role of a new consensus standard. *Regenerative Medicine*, 6(2):255–260.

J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:10.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of bionlp shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 1–6, Portland, Oregon, USA, June. Association for Computational Linguistics.

Roman Klinger, Corinna Kolik, Juliane Fluck, Martin Hofmann-Apitius, and Christoph M. Friedrich. 2008. Detection of iupac and iupac-like chemical names. *Bioinformatics*, 24(13):i268–i276.

Robert Leaman and Graciela Gonzalez. 2008. Banner: An executable survey of advances in biomedical named entity recognition. In *Pacific Symposium of Biocomputing*, pages 652–663.

Peter Löser, Jacqueline Schirm, Anke Guhr, Anna M. Wobus, and Andreas Kurtz. 2010. Human embryonic stem cell lines and their use in international research. *STEM CELLS*, 28(2):240–246.

Sampo Pyysalo, Antti Airola, Juho Heimonen, Jari Bjorne, Filip Ginter, and Tapio Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9(Suppl 3):S6.

Sampo Pyysalo, Tomoko Ohta, Jun'ichi Tsujii, and Sophia Ananiadou. 2011. Anatomical entity recognition with open biomedical ontologies. In *Proceedings of the Fourth International Symposium on Languages in Biology and Medicine (LBM 2011)*.

Dietrich Rebholz-Schuhmann, Antonio Jimeno-Yepes, Erik M. van Mulligen, Ning Kang, Jan A. Kors, David Milward, Peter Corbett, Ekaterina Buyko, Elena Beisswanger, and Udo Hahn. 2010. Calbc silver standard corpus. *J. Bioinformatics and Computational Biology*, 8(1):163–179.

Paolo Romano, Assunta Manniello, Ottavia Aresu, Massimiliano Armento, Michela Cesaro, and Barbara Parodi. 2009. Cell line data base: structure and recent improvements towards molecular authentication of human cell lines. *Nucleic Acids Research*, 37(suppl 1):D925–D932.

Barbara Rosario and Marti A. Hearst. 2004. Classifying

semantic relations in bioscience texts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ariel S. Schwartz and Marti A. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Pacific Symposium on Biocomputing*, pages 451–462.

Burr Settles. 2005. Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192, Jul.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France, April. Association for Computational Linguistics. (to appear).

Lorraine Tanabe, Natalie Xie, Lynne Thom, Wayne Matten, and W John Wilbur. 2005. Genetag: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6(Suppl 1):S3.

Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou. 2011. Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, 12(1):393.

James A. Thomson, Joseph Itskovitz-Eldor, Sander S. Shapiro, Michelle A. Waknitz, Jennifer J. Swiergiel, Vivienne S. Marshall, and Jeffrey M. Jones. 1998. Embryonic stem cell lines derived from human blastocysts. *Science*, 282(5391):1145–1147.

Sowmya Viswanathan and Armand Keating. 2011. Overcoming the challenges of conducting translational research in cell therapy. *Frontiers of Medicine*, 5:333–335. 10.1007/s11684-011-0166-2.

C. J. E. Watson and J. H. Dark. 2012. Organ transplantation: historical perspective and current practice. *British Journal of Anaesthesia*, 108(suppl 1):i29–i42.

Inken Wohlers, Harald Stachelscheid, Joeri Borstlap, Katrin Zeilinger, and Jrg C. Gerlach. 2009. The characterization tool: A knowledge-based stem cell, differentiated cell, and tissue database with a web-based analysis frontend. *Stem Cell Research*, 3(23):88 – 95.

Junying Yu, Maxim A. Vodyanik, Kim Smuga-Otto, Jessica Antosiewicz-Bourget, Jennifer L. Frane, Shulan Tian, Jeff Nie, Gudrun A. Jonsdottir, Victor Ruotti, Ron Stewart, Igor I. Slukvin, and James A. Thomson. 2007. Induced pluripotent stem cell lines derived from human somatic cells. *Science*, 318(5858):1917–1920.