A Corpus of Tables in Full-Text Biomedical Research Publications

Tatyana Shmanina[1,3], **Ingrid Zukerman**[1], **Ai Lee Cheam**[1], **Thomas Bochynek**[2,3], **Lawrence Cavedon**[4]
[1]Clayton School of Information Technology, Monash University, Australia
[2]Caulfield School of Information Technology, Monash University, Australia
[3]Data61, CSIRO, Melbourne, Australia
[4]School of Science, RMIT University, Australia
[1,2]firstname.lastname@monash.edu, [4]firstname.lastname@rmit.edu.au

Abstract

The development of text mining techniques for biomedical research literature has received increased attention in recent times. However, most of these techniques focus on prose, while much important biomedical data reside in tables. In this paper, we present a corpus created to serve as a gold standard for the development and evaluation of techniques for the automatic extraction of information from biomedical tables. We describe the guidelines used for corpus annotation and the manner in which they were developed. The high inter-annotator agreement achieved on the corpus, and the generic nature of our annotation approach, suggest that the developed guidelines can serve as a general framework for table annotation in biomedical and other scientific domains. The annotated corpus and the guidelines are available at http://www.csse.monash.edu.au/research/umnl/data/index.shtml.

1 Introduction

Biomedical science generates vast quantities of data, which reside in publicly available databases and repositories of structured biomedical information, such as the Catalogue of Somatic Mutations in Cancer (Bamford et al., 2004) and the International Society for Gastrointestinal Hereditary Tumours Database (Plazzer et al., 2013). In order to be useful to researchers, data sources must contain precise and reliable information, and therefore are typically manually curated by biomedical professionals (Campos et al., 2013), which leads to a "curation bottleneck". As a result, automatic information extraction from biomedical literature has become an important task.

The development of approaches for automatic and semi-automatic information extraction requires annotated corpora for training and evaluating text mining systems. To date, biomedical text mining has focused on extracting information from prose, yielding a wealth of diverse annotated corpora for unstructured text. For example, gold and silver standard corpora have been developed for a variety of tasks, such as named entity recognition (Doğan et al., 2014; Kim et al., 2003; Rebholz-Schuhmann et al., 2010; Verspoor et al., 2013), entity linking (Bada et al., 2012; Doğan et al., 2014), and relation and event extraction (Kim et al., 2003; Lee et al., 2016; Rosario and Hearst, 2004; Verspoor et al., 2013). The source of these datasets also varies, e.g., corpora comprising research abstracts (Doğan et al., 2014; Kim et al., 2003; Rebholz-Schuhmann et al., 2010) versus full-text journal articles (Bada et al., 2012; Lee et al., 2016; Verspoor et al., 2013).

In addition to prose, biomedical literature frequently presents information in other forms, such as tables and graphs. Several studies have shown that tables often contain important data and experimental results that are not mentioned in the main text of publications (Jimeno Yepes and Verspoor, 2013; Wong et al., 2009). At the same time, Jimeno Yepes and Verspoor (2013) have shown that text mining techniques developed for prose tend to under-perform when applied to tables, because of the difference in how information is presented in tables and in text. For example, the arrangement of cells, which is meaningful for understanding table contents, is not taken into account by classical prose mining techniques. This calls for the development of specialised approaches to information extraction from tables (*Table IE*) in

Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2016), pages 70–79, Osaka, Japan, December 12th 2016.

the biomedical literature. Advances in Table IE have been made in general and Web domains (Cafarella et al., 2008; Hignette et al., 2009; Hurst, 2000; Jannach et al., 2009; Limaye et al., 2010; Mulwad et al., 2013; Quercini and Reynaud, 2013; Van Assem et al., 2010; Venetis et al., 2011; Wang et al., 2012; Yakout et al., 2012; Yin et al., 2011; Yosef et al., 2011). However, Table IE has received comparatively little attention from the biomedical text mining community, which may be partly attributed to the limited availability of suitable annotated corpora.

In this paper, we introduce a new corpus of tables obtained from full-text biomedical research papers in two areas of genetics: human cancer and mouse — the tables in the human cancer papers cover topics such as genetic aberrations and patient and tumour characteristics; and the tables in the mouse papers include distributions of genotypes and phenotypes, and parameters and outcomes of genetic analyses. This corpus was developed to support our work in Table IE, which focuses on relation extraction and fine-grained named entity recognition. The tables in our corpus were supplemented with the following annotations: (1) concepts in table cells; (2) classification of table cells into homogeneous cell groups; (3) fine-grained cell types of each homogeneous cell group; and (4) relations between cell groups.

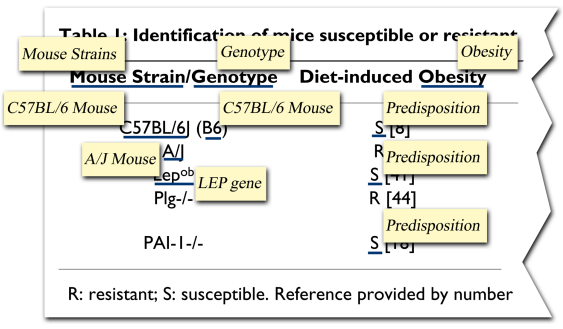
In Section 2, we motivate the design of the corpus, and describe the created corpus, the annotation schema and the annotation guidelines. Section 3 details the corpus construction process. The characteristics of the developed corpus are discussed in Section 4, followed by concluding remarks in Section 5.

2 Corpus Design

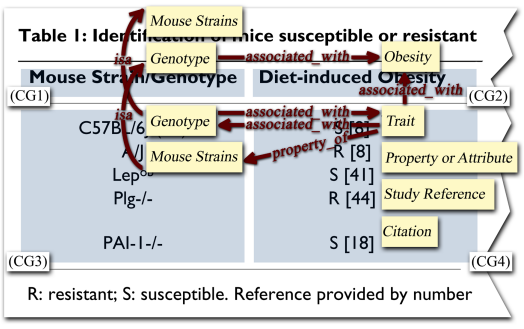
The design of our corpus and associated annotation schemas is closely aligned with the requirements of our project on information extraction from tables in biomedical research papers. However, both the corpus and the schemas are general enough to be of value to the broader biomedical text mining community.

The presented corpus is the gold standard for two information extraction tasks: (1) mapping of table cells into fine-grained entity types, and (2) identification of relations between table cells. Both fine-grained entity types and relations are drawn from a domain vocabulary. The design of the corpus was strongly influenced by the characteristics of the biomedical tables we encountered, which have a variety of structures, and tend to be more complex than the structures typically considered by researchers in information extraction. For example, Limaye et al. (2010) and Mulwad et al. (2013), who worked on a table information extraction task similar to ours, but for a Web domain, assumed that tables have lattice-like structures, and that the objective of information extraction is to identify table column types and relations between columns. However, even for simple lattice-like biomedical tables, we often found that interesting relations could be built between columns and their headers, and between the headers themselves. For instance, the relation *associated_with* can be built between the header "Diet-induced Obesity" and the data cells "S [8]", "R [8]", "S [41]", "R [44]" and "S [18]" in the table in Figure 1b.

This motivated us to view each table as a collection of homogeneous groups of cells, rather than a collection of columns. We assume that (1) all cells within each homogeneous group of cells share the same fine-grained type, and (2) it is possible to define relations among cell groups that hold for each corresponding pair of cells inside the groups. These assumptions motivated the creation of four types of annotation: (1) *cell group*, which splits each table into sets of homogeneous cell groups; (2) *cell*



(a) Concept annotations. The annotated text spans are underlined, concept annotations are pictured in boxes.



(b) Cell group, cell type and relation annotations. Cell groups (CG1)–(CG4) are highlighted, cell types are pictured in boxes, relations are represented as arrows.

Figure 1. Annotation example for a sample biomedical table from (Hoover-Plow et al., 2006)

type, which represents the mapping of all cells in a homogeneous group into a single fine-grained named entity label; (3) *concept*, which represents the mapping of utterances inside table cells (i.e., the syntactic heads of the utterances expanded with their modifiers) into their semantic equivalents from a domain vocabulary; and (4) *relation*, which represents relations between cell groups. Figure 1 illustrates a table annotated with these types of information.

2.1 Annotation Schema

Cell Group Annotation. Each set of homogeneous cells in a table is assigned a unique identifier to distinguish between different cell groups.

Concept, Cell Type and Relation Annotation. To generate fine-grained entity and relation labels, we used the National Cancer Institute (*NCI*) subset (Sioutos et al., 2007) of the Unified Medical Language System® (UMLS®) Metathesaurus (*UMLS-NCI*) and the UMLS Semantic Network (*UMLS SN*) as the basis for our annotation schema (UMLS release 2015AA). To illustrate, using this annotation schema, the text spans "Obesity" and "S" in Figure 1a are mapped to the UMLS-NCI concepts *Obesity* and *Predisposition* respectively to create concept annotations. The single-cell group "Diet-induced Obesity" is assigned the fine-grained cell type *Obesity* (Figure 1b) and the coarser cell type *Disease or Syndrome* from the UMLS SN. Finally, the relation *associated_with* from the UMLS SN is built between the cell group "Diet-induced Obesity" and the cell group comprising the data cells "S [8]", "R [8]", "S [41]", "R [44]" and "S [18]" (Figure 1b). No relations from UMLS-NCI can be built between these cell groups.

We chose the UMLS because of the lexical and information extraction tools for unstructured text that are distributed with the UMLS. We decided to use one subset of the UMLS Metathesaurus, because (1) the reduced size of the annotation schema reduces the complexity of the annotation tasks; and (2) the UMLS combines over 100 source vocabularies, and does not resolve conflicts among these vocabularies.

The ***UMLS-NCI*** subset was chosen because it is a large, comprehensive and heterogeneous controlled vocabulary, which focuses on genetics. It comprises over 110,000 biomedical Concept Unique Identifiers (*CUI*s) (e.g., *C0028754* for *Obesity* and *C0220898* for *Predisposition*), and more than one million relationships between concepts, drawn from 208 unique relation types (e.g., the generic hierarchical relations *parent-child*, and many specialised relations, such as *is_grade_of_disease* and *gene_mapped_to_disease*), thus providing a large set of fine-grained entity and relation labels. The concepts and relations from UMLS-NCI were used as the primary set of labels for concept, cell-type and relation annotation. However, despite its extensive scope, the UMLS-NCI's coverage of relations between the concepts in our dataset was very sparse. Specifically, it yielded only 222 relations in total for the entire corpus. We therefore expanded the set of relations in the UMLS-NCI schema with SRs from the UMLS SN.

The ***UMLS SN*** provides an alternative set of labels for cell types and relations that is smaller, and hence of lower granularity, than the labels in UMLS-NCI. It consists of (1) a set of Semantic Types (*ST*s) that provide a broad subject categorisation of the concepts represented in the UMLS-NCI; and (2) a set of Semantic Relations (*SR*s) that can hold between STs. The UMLS SN contains 127 STs (e.g., *Organism Attribute* for the concepts *Age* and *Gender*, *Clinical Attribute* for the concepts *Tumour Stage* and *Cellular Differentiation*) and 54 SRs (e.g., *isa*, *causes*, *consists_of*, *interacts_with*, *assesses_effect_of* and *location_of*). Each concept in UMLS-NCI is assigned one or more STs;[1] and SRs defined between the STs in UMLS SN may or may not hold between particular concepts assigned to these STs. The incorporation of UMLS SN into our schema enabled the creation of 1625 additional relation annotations.

3 Corpus Construction

The construction of the dataset involved the following activities, described below: (1) document selection, (2) development of annotation guidelines, (3) document pre-processing and choice of distribution format, (4) annotation tool configuration and development, and (5) actual annotation process.

3.1 Document Selection

The following criteria were applied to select documents for our corpus: (1) the documents must represent full-text biomedical research articles containing at least one table; (2) the corpus must be diverse with

[1] When a concept was linked to several STs, our annotation guidelines required the exclusion of STs that were irrelevant in the context of the source table.

respect to the structure of tables, and representative of their distribution in biomedical research publications, in order to eliminate selection bias based on table structure; (3) the documents must be available in a structured format, preferably XML, to avoid the need to programmatically determine document and table structure; and (4) the articles must be available under non-restrictive licensing terms to enable future public release of our dataset. Finally, we preferred articles that were already included in other corpora with existing concept, named entity or relation annotations of unstructured text, in order to facilitate the future development of text mining tools for the joint analysis of free text and tabular content.

The application of these criteria resulted in the inclusion of papers from the following datasets:[2]

1. CRAFT Corpus (Bada et al., 2012). A subset of the CRAFT Corpus comprising 24 papers (50 tables) was included in our dataset. The CRAFT dataset, which comprises articles drawn from the Open Access subset of PubMed Central, is heterogeneous with respect to the content of the papers, and covers topics related to mouse genetics. The CRAFT dataset contains a mapping of concepts that appear in its free-text parts into seven open biomedical ontologies.
2. The Human Variome Project (HVP) Corpus (Verspoor et al., 2013). Nine out of ten papers (28 tables) from the HVP Corpus were included in our dataset. This corpus covers topics related to the genetics of human colon cancer. The free-text parts of the papers are annotated using a small annotation schema comprising eleven named entity classes and thirteen binary relations between the entity classes.
3. An additional subset of ten papers (22 tables) was randomly sampled from Open Access subsets of three datasets comprising papers about genomic variation (Jimeno Yepes and Verspoor, 2013; Wong et al., 2009). These datasets did not contain annotations of unstructured text, but several of these papers were previously used in (Shmanina et al., 2014).

3.2 Development of Annotation Guidelines

The table annotation guidelines were developed by the first author, who is a researcher in biomedical text mining. The guidelines contain four parts, each corresponding to a single annotation task: (1) cell group, (2) concept, (3) cell type, and (4) relation annotation.

The initial versions of the guidelines were developed through several iterative attempts to annotate ten tables using the guidelines. After each annotation iteration, we tested whether the strict application of the guidelines yielded consistent and objective annotations, and revised the guidelines as necessary. The final version of the guidelines was 49 pages long.

Cell Group Annotation. Cells were collated into homogeneous cell groups according to two main guidelines: (1) every header cell should form its own cell group, and (2) data cells should be merged into maximum-size cell groups.

Concept Annotation. These annotations were created due to the potential subjectivity of assigning a cell type to a cell group. For example, a cell group comprising the entries "C57/LJ", "AKR/J" and "NZB/BiNJ" can be potentially assigned UMLS-NCI concepts *[C0026809] Mice*, *[C1518614] Organism Strain* or *[C2985604] Biologic Entity Group*. By first annotating concepts for each table cell, it was relatively straightforward to derive cell type annotations. For example, the mentions above can be mapped into the concepts *[C1511387] C57L/J Mouse*, *[C1515841] AKR/J Mouse* and *[C1513862] NZB/BlNJ Mouse* — all of which have a common parent *[C0025927] Inbred Mouse Strains* in the UMLS-NCI concept hierarchy, which is then chosen to represent the cell type.

We based our concept annotation guidelines on those used in the CRAFT Corpus, which resulted in high inter-annotator agreement for concept annotations of free text (Bada et al., 2012). The main characteristics of these guidelines are: (1) a text is mapped into a concept from a vocabulary only if the concept is an exact semantic match for the text; and (2) the rules for the identification of text segments are syntax-based, and specify how to annotate nouns and noun phrases, adjectival and prepositional phrases, nested and overlapping mentions, etc.

We slightly modified the original CRAFT guidelines to better suit our table annotation task. Firstly, for each table cell, we annotated only the syntactic head of an utterance, expanded with as many of

[2] All the articles in our table dataset belong to the Open Access subset of PubMed Central.

its modifiers as possible. For example, given the table entry "Fragment length", UMLS-NCI contains concepts that are semantically equivalent to "Fragment" and "length", but there is no concept that is semantically equivalent to "Fragment length". Therefore, only "length", which is the syntactic head of the phrase, receives a concept annotation, and its modifier is excluded from the annotation. Secondly, the text inside table cells tends to be more concise than unstructured text. For instance, in a column listing colon cancer stages, such as "A", "B" and "C", "B" could stand for "stage B" or "Dukes B rectal cancer". We addressed this problem by stipulating that if it was not possible to annotate a concept using the guidelines from (Bada et al., 2012), its mention should be mapped into the semantically closest concept available in the vocabulary, e.g., "B" should be mapped to *Dukes B rectal cancer*.

Cell Type Annotation. The cell type is the most specific superclass of all the entries in a cell group. It must first be obtained for entries with concept annotations (if they exist) using the UMLS-NCI concept hierarchy, and possibly generalised to entries without concept annotations. However, if there are no concept annotations for the cells in a cell group (54.22% of cell groups in our dataset), or the concept annotations do not have an informative superclass in UMLS-NCI (e.g., when the most specific common ancestor of concept annotations in UMLS-NCI is *Conceptual Entity*, *NCI Administrative Concept* or *NCI Thesaurus*), the annotator must retrieve the most specific concept in UMLS-NCI that best describes the content of the cells in the cell group. For example, a cell group that lists chromosomes (e.g., "1", "2", "18") is annotated with the concept *[C0008633] Chromosome*.

Relation Annotation. The relation annotation guidelines were developed under the assumption that the relation annotation phase would follow the cell type annotation phase. This meant that all relation hypotheses could be automatically pre-computed using the constraints from the UMLS Metathesaurus and the UMLS SN, and suggested to the annotators, who in turn could accept or reject the suggestions. Therefore, the relation annotation guidelines contain the following information: (1) definition and examples of what constitutes a valid relation between two cell groups; (2) a definition and use cases for the *isa* relation; and (3) a list of cases where no relation should be built between two cell groups.

3.3 Choice of Corpus Distribution Formats and Pre-processing of the Documents

When choosing formats for the articles and annotation, we considered the following criteria: (1) they must preserve information about table structure; (2) they should preferably preserve information about the structure and formatting of the original paper; and (3) they should be flexible and expressive enough to uniformly encode all the annotation types and schemas described above.

We considered various linguistic annotation formats, such as BRAT stand-off, BioC and XML/JSON stand-off. However, neither BRAT nor BioC satisfy the first two criteria, as they convert tables into plain text, and BRAT also stores documents in plain text. We therefore decided to distribute the articles in the original XML format used by PubMed Central for archiving. Such XML versions of articles use The Journal Archiving and Interchange Tag Set,[3] which preserves the content, structure and format of the articles and the tables within. The created annotations were stored in stand-off JSON format – one JSON annotation file per paper.

To construct the dataset, we downloaded the XML versions of the papers from the FTP service of PubMed Central.[4] We then automatically assigned unique IDs to all the XML tags that contained actual content (as opposed to article meta-data), such as paragraphs `<p>`, section titles `<title>`, article titles `<article-title>`, table headers `<th>` and data cells `<td>`. The templates of the JSON annotation files for each paper in the dataset were automatically generated by a script.

3.4 Annotation Process

The annotation process was conducted in three stages. First, cell groups were annotated, followed by the annotation of concepts and cell types for each cell group; relations between cells were annotated last.

Cell group annotation was carried out by the first author for the 100 tables of the dataset (43 papers) using a text editor with programming language support.

[3] `http://dtd.nlm.nih.gov/archiving/`

[4] `ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/`

Owing to budgetary constraints, the annotation for the second and third stages was done on a subset of 83 tables in 39 papers. The annotation was performed by the first author and two specialist annotators who hold post-graduate degrees in Biology and are familiar with the subject matter. The annotators received extensive annotation guidelines for each stage, were instructed in the use of the annotation tools, and were advised to consider information from the full text of the papers during the annotation process. Thereafter, the annotators performed a double-blind annotation of a subset of five tables (two papers), to assess the level of inter-annotator agreement, and to discover any problems related to the guidelines and tools. The budgetary constraints also led us to perform a single-blind annotation for the remaining 78 tables (37 papers), which were distributed between the first author and the annotators: each participant annotated their assigned tables, which were then passed to another team member for verification. Throughout the annotation process (130 hours per person) the first author and the annotators met after every 6 to 8 hours of annotation, in order to measure the agreement between the original annotator of a table and the "reviewing" team member; annotation disagreements were resolved by consensus, and the annotation guidelines were amended where necessary, which happened rarely. At the end of each annotation stage, the master version of the annotations for the entire corpus was verified by the first author for consistency and compliance with the guidelines. Annotations were corrected where necessary.

Concept and ***cell-type*** annotations were carried out using a modified version of the BRAT annotation tool. Prior to loading the corpus into the tool, the XML files were automatically mapped into plain text (the input format of BRAT). Upon completion of the annotation, the files were mapped back into the XML/JSON distribution format.[5] We employed an in-house Web interface to integrate the NCI subset of the UMLS Metathesaurus into our BRAT installation. The annotators used BRAT to select text spans for annotation, and to query the Web interface for concepts related to these text spans. The Web interface allowed annotators to browse the lists of returned concepts, and to look up information about these concepts, such as name, STs, definition and position in the NCI concept hierarchy. If a suitable concept was found, the annotation was sent back to BRAT.

Finally, ***relations between cells*** were annotated using an in-house online Relation Annotation Tool, which suggested relation annotations drawn from UMLS-NCI and the UMLS Semantic Network on the basis of existing cell type annotations.

4 Results and Discussion

4.1 Corpus and Annotation Statistics

83 tables from our corpus were manually annotated with cell groups, ***C***oncepts, ***C***ell ***T***ypes and ***R***elations (denoted *CCTR-83*); 17 additional tables were annotated with cell groups only. Table 1 details the composition of the dataset. As seen in Table 1, the corpus is evenly split between two main topics, human cancer and mouse genetics (43 and 40 tables in CCTR-83 respectively), offering interesting opportunities for cross-domain training and testing. Statistics about the dimensions of the tables (average, median, minimum and maximum per table, and total counts) appear in Table 2. Table 3 shows statistics of cell-group, concept, cell-type and relation annotations, both for all annotations (left-hand side) and unique annotations (right-hand side).

As seen in Table 3, the unique concept and cell-type annotations constitute a relatively high percentage of their total counts (528 out of 3042 concepts, and 375 out of 2545 cell types based on UMLS CUIs). However, the distributions of the unique annotations are skewed. For example, the top-three most frequent cell-type annotations based on UMLS-NCI (*Count*, *Percent* and *Biologic Entity Group Quantity*) together constitute 36% of the 2545 cell types based on UMLS CUIs, while 178 cell-type annotations appear only once in the corpus. The most frequent cell-type annotation based on UMLS STs is *Quantitative Concept*, comprising 49.7% of the 2089 UMLS ST cell-type annotations; followed by the label *Organism Attribute*, which constitutes only 6% of the annotations. Such a strong bias towards *Quantitative Concept* may be explained by the predominantly quantitative nature of biomedical tables: 43.5% of the cell groups in our corpus contain numbers and numerical expressions, while 25.8% and 23.7% of the cell groups contain free text (terms and phrases) and abbreviations respectively; the remaining 7% of the

[5]To our knowledge, currently there is no annotation tool that natively supports table annotation. Due to our budgetary constraints, we were unable to develop such an annotation tool ourselves, and had to resort to partial suboptimal solutions.

Source Dataset	Domain	Full dataset		CCTR-83	
		Papers	Tables	Papers	Tables
CRAFT	Mouse genetics	24	50	22	40
HVP Corpus	Human colorectal cancer genetics	9	28	8	24
Jimeno Yepes and Verspoor (2013)	Human cancer genetics	8	17	8	17
Wong et al. (2009)	Human cancer genetics	2	5	1	2

Table 1. Article and table counts and domains

Table Element	Full dataset					CCTR-83				
	Total #	# per table				Total #	# per table			
		Avg.	Med.	Min.	Max.		Avg.	Med.	Min.	Max.
Cells	13061	130.61	78	12	1300	9753	117.51	64	12	1300
Rows	1929	19.29	13	3	100	1500	18.07	10	3	100
Columns	631	6.31	5	2	16	500	6.02	5	2	13

Table 2. Counts of table cells, rows and columns in the dataset

Annotation Type	All annotations					Unique annotations				
	Total #	# per table				Total #	# per table			
		Avg.	Med.	Min.	Max.		Avg.	Med.	Min.	Max.
Cell Group	2443	24.43	15.0	4	163	–	–	–	–	–
CCTR-83	2134	25.71	15.0	4	163	–	–	–	–	–
Concept	3042	36.65	23	1	296	528	13.83	11	1	37
Cell Type										
UMLS CUI	2545	30.66	20	4	148	375	11.28	10	3	25
UMLS ST	2089	25.17	16	4	122	52	6.75	6	2	15
Relation										
All labels	1847	22.52	9	1	113	31	3.22	3	1	12
UMLS MetaTh.	222	2.71	0	0	32	4	0.61	0	0	2
UMLS SN	1625	19.82	9	1	109	27	2.61	2	1	10

Table 3. Annotation counts (including ambiguous annotations)

cell groups have mixed content (e.g., "MSI-H (n = 19)"). The breakdown for cells is 52.4% numerical, 14% text, 26.7% abbreviations and 6.9% mixed content.

Another noteworthy observation is the relatively modest corpus coverage provided by our concept and relation annotations. Concept annotations were assigned to only 30.47% of the non-empty table cells, which corresponds to 45.78% of the cell groups. This may be explained by (1) the quantitative nature of many biomedical tables combined with the difficulty of mapping numbers to concepts; and (2) the insufficient coverage of table entries by UMLS-NCI for non-numerical concepts such as specific mutations (e.g., "c.1886 A $>$ G"), base sequences (e.g., "5'-dT20-ACTGGC...GAAAAC-3'") and patient IDs (e.g., "IC628"). With regard to relations, even after we expanded the relation annotation schema with labels from UMLS SN, we annotated only 1847 relations between the available 2134 cell groups in the CCTR-83 dataset, yielding a small set of only 31 unique labels, and a median of nine relations per 15-cell-group table (column 4 in Table 3). The distribution of the relation labels in the corpus is even more skewed than the distribution of the concept and cell-type labels, with about 75% of the relations being quite general: the *isa* relation from UMLS SN constitutes 51% of the relation labels, followed by the labels *associated_with* from UMLS SN (11.3%), and *isa* and *inverse_isa* from UMLS-NCI (each 6%). This imbalance may be attributed to the discrepancy between the information in our tables and the relations available in UMLS, which may be mitigated by employing a different annotation schema.

Annotation	Kappa values (entire corpus)	Kappa values (per paper)			
		Avg.	Med.	Min.	Max.
Concept	0.88	0.90	0.96	0.13	1.0
Cell Type					
UMLS CUI	0.87	0.82	0.92	0.15	1.0
UMLS ST	0.87	0.86	0.94	0.23	1.0
Relation	0.82	0.87	0.91	0.49	1.0

Table 4. Inter-annotator agreement for concept, cell-type and relation annotation

4.2 Inter-Annotator Agreement

To enable the prompt resolution of problems in the annotation guidelines, and to map the progression of inter-annotator agreement (IAA) over time, per-paper IAA was measured at every conflict-resolution meeting throughout the annotation process. We used Cohen's Kappa statistic (Cohen, 1960) to evaluate IAA for all types of annotations.

Two concept annotations were deemed to match if their UMLS CUIs and text spans were equal; two UMLS CUI cell-type annotations matched if their CUIs were equal, and similarly, two UMLS ST cell-type annotations matched if their STs were equal; and two relation annotations were deemed to match if they had the same relation label, direction and arguments. It is worth noting that every cell group received at least one cell-type label; if there was more than one label (ambiguous annotation), each label was considered separately when computing IAA. In contrast, some table-entries did not have concept annotations, and similarly, some pairs of cell groups did not have relation annotations. In order to handle these cases, as well as ambiguous annotations, we added the label *No_Annotation*, so that IAA could be computed between absent and present labels.

IAAs computed over the entire corpus and per-paper IAAs appear in Table 4. For all annotation types, the average and median IAA values exceeded 0.82 and 0.91 respectively. This shows that, for most papers in our dataset, the application of our annotation guidelines yielded highly consistent annotations. However, the low minimum IAA values indicate that a few papers posed a significant challenge. This was variously due to (1) ambiguities in the annotation guidelines, which were fixed after discussing the relevant part of the guidelines;[6] (2) erroneous annotations caused by lack of clarity and ambiguity of some concepts and relations in the UMLS; and (3) absence of a concept annotation from which a cell-type annotation could be derived — these cases were less consistent across annotators than those where concept annotations were available.

5 Conclusion

We have offered a corpus comprising 100 tables sourced from 43 biomedical journal articles on the topic of genetics. All the tables in the corpus were manually annotated with information about homogeneous cell groups, and a subset of 83 tables was annotated with a total of more than 3000 concepts, 2000 cell types and 1800 relations, drawn from the Unified Medical Language System®. Our annotation schema was designed to accurately capture fine-grained semantic classes of table entries and the relationships between them. This annotation schema, combined with the stringent table annotation guidelines we developed, enabled a high average inter-annotator agreement of over 0.82 for all annotation types. This makes both the annotated corpus and the guidelines used to create it a valuable resource for the development and evaluation of tools for information extraction from biomedical tables. Furthermore, although our guidelines were developed for a particular biomedical corpus, they may be adapted to tables from other scientific fields, thus providing a general framework for table annotation.

Acknowledgements

This research was supported in part by National ICT Australia (NICTA), CSIRO and the Monash University FIT Graduate Research Candidature Funding Scheme.

[6] After updating the guidelines at a meeting, the papers were not returned to the corresponding annotators for re-annotation, hence the IAA was not re-calculated. Only the master version of the annotations in the entire corpus was reviewed for consistency and compliance with the guidelines at the end of each annotation stage.

References

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A. Baumgartner, K. Bretonnel Cohen, Karin Verspoor, Judith A. Blake, and Lawrence E. Hunter. 2012. Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13(1):161.

Sally Bamford, Todd E. Dawson, Simon A. Forbes, Jody Clements, Roger Pettett, Ahmet Dogan, Adrienne M. Flanagan, Jon Teague, P. Andrew Futreal, and Michael R. Stratton. 2004. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British Journal of Cancer*, 91(2):355–358.

Michael J. Cafarella, Alon Y. Halevy, Daisy Zhe Wang, Eugene Wu, and Yang Zhang. 2008. WebTables: Exploring the power of tables on the Web. *Proceedings of the VLDB Endowment*, 1(1):538–549.

David Campos, Sérgio Matos, and José Luís Oliveira. 2013. A modular framework for biomedical concept recognition. *BMC Bioinformatics*, 14:281.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Gaëlle Hignette, Patrice Buche, Juliette Dibie-Barthélemy, and Ollivier Haemmerlé. 2009. Fuzzy annotation of web data tables driven by a domain ontology. In *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*, pages 638–653, Heraklion, Crete, Greece. Springer-Verlag.

Jane Hoover-Plow, Aleksey Shchurin, Erika Hart, Jingfeng Sha, Annie E. Hill, Jonathan B. Singer, and Joseph H. Nadeau. 2006. Genetic background determines response to hemostasis and thrombosis. *BMC Hematology*, 6(1):1.

Matthew Francis Hurst. 2000. *The interpretation of tables in texts*. Ph.D. thesis, University of Edinburgh.

Dietmar Jannach, Kostyantyn Shchekotykhin, and Gerhard Friedrich. 2009. Automated ontology instantiation from tabular web sources – the AllRight system. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):136–153.

Antonio Jimeno Yepes and Karin Verspoor. 2013. Towards automatic large-scale curation of genomic variation: improving coverage based on supplementary material. In *Proceedings of BioLINK SIG 2013: Roles for text mining in biomedical knowledge discovery and translational medicine*, pages 39–43, Berlin, Germany.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. 2003. GENIA corpus – a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182.

Kyubum Lee, Sunwon Lee, Sungjoon Park, Sunkyu Kim, Suhkyung Kim, Kwanghun Choi, Aik Choon Tan, and Jaewoo Kang. 2016. BRONCO: Biomedical entity relation oncology corpus for extracting gene-variant-disease-drug relations. *Database: The Journal of Biological Databases and Curation*, 2016.

Girija Limaye, Sunita Sarawagi, and Soumen Chakrabarti. 2010. Annotating and searching web tables using entities, types and relationships. *Proceedings of the VLDB Endowment*, 3(1-2):1338–1347.

Varish Mulwad, Tim Finin, and Anupam Joshi. 2013. Semantic message passing for generating linked data from tables. In *Proceedings of the 12th International Semantic Web Conference (ISWC 2013)*, pages 363–378, Sydney, Australia. Springer-Verlag.

John-Paul Plazzer, Rolf H. Sijmons, Michael O. Woods, Päivi T. Peltomäki, Bryony A. Thompson, Johan T. Den Dunnen, and Finlay Macrae. 2013. The InSiGHT database: utilizing 100 years of insights into Lynch Syndrome. *Familial Cancer*, 12(2):175–180.

Gianluca Quercini and Chantal Reynaud. 2013. Entity discovery and annotation in tables. In *Proceedings of the 16th International Conference on Extending Database Technology (EDBT '13)*, pages 693–704, Genoa, Italy. ACM.

Dietrich Rebholz-Schuhmann, Antonio José Jimeno Yepes, Erik M. van Mulligen, Ning Kang, Jan A. Kors, David Milward, Peter T. Corbett, Ekaterina Buyko, Katrin Tomanek, Elena Beisswanger, and Udo Hahn. 2010. The CALBC silver standard corpus for biomedical named entities – a study in harmonizing the contributions from four independent named entity taggers. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC '10)*, pages 568–573, Valletta, Malta. European Language Resources Association.

Barbara Rosario and Marti A. Hearst. 2004. Classifying semantic relations in bioscience texts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL '04)*, pages 430–437, Barcelona, Spain. Association for Computational Linguistics.

Tatyana Shmanina, Lawrence Cavedon, and Ingrid Zukerman. 2014. Challenges in information extraction from tables in biomedical research publications: A dataset analysis. In *Proceedings of the Australasian Language Technology Association Workshop 2014 (ALTA 2014)*, pages 118–122, Melbourne, Australia. Association for Computational Linguistics.

Nicholas Sioutos, Sherri de Coronado, Margaret W. Haber, Frank W. Hartel, Wen-Ling Shaiu, and Lawrence W. Wright. 2007. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, 40(1):30–43.

Mark Van Assem, Hajo Rijgersberg, Mari Wigham, and Jan Top. 2010. Converting and annotating quantitative data tables. In *Proceedings of the 9th International Semantic Web Conference (ISWC 2010)*, pages 16–31, Shanghai, China. Springer-Verlag.

Petros Venetis, Alon Y. Halevy, Jayant Madhavan, Marius Paşca, Warren Shen, Fei Wu, Gengxin Miao, and Chung Wu. 2011. Recovering semantics of tables on the Web. *Proceedings of the VLDB Endowment*, 4(9):528–538.

Karin Verspoor, Antonio Jimeno Yepes, Lawrence Cavedon, Tara McIntosh, Asha Herten-Crabb, Zoë Thomas, and John-Paul Plazzer. 2013. Annotating the biomedical literature for the human variome. *Database: The Journal of Biological Databases and Curation*, 2013.

Jingjing Wang, Haixun Wang, Zhongyuan Wang, and Kenny Q. Zhu. 2012. Understanding tables on the Web. In *Proceedings of the 31st ER International Conference on Conceptual Modeling (ER 2012)*, pages 141–155, Florence, Italy. Springer-Verlag.

Wern Wong, David Martinez, and Lawrence Cavedon. 2009. Extraction of named entities from tables in gene mutation literature. In *Proceedings of the BioNLP 2009 Workshop (BioNLP '09)*, pages 46–54, Boulder, Colorado, USA. Association for Computational Linguistics.

Mohamed Yakout, Kris Ganjam, Kaushik Chakrabarti, and Surajit Chaudhuri. 2012. InfoGather: Entity augmentation and attribute discovery by holistic matching with web tables. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (SIGMOD '12)*, pages 97–108, Scottsdale, Arizona, USA. ACM.

Xiaoxin Yin, Wenzhao Tan, and Chao Liu. 2011. FACTO: a fact lookup engine based on web tables. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*, pages 507–516, Hyderabad, India. ACM.

Mohamed Amir Yosef, Johannes Hoffart, Ilaria Bordino, Marc Spaniol, and Gerhard Weikum. 2011. AIDA: An online tool for accurate disambiguation of named entities in text and tables. *Proceedings of the VLDB Endowment*, 4(12):1450–1453.