Challenges in Information Extraction from Tables in Biomedical Research Publications: a Dataset Analysis

Tatyana Shmanina[1,2], **Lawrence Cavedon**[3], **Ingrid Zukerman**[1,2]
[1]Clayton School of Information Technology, Monash University, Australia
[2]NICTA Victoria Research Laboratory, Melbourne, Australia
[3]School of Computer Science and IT, RMIT University, Australia
[1]firstname.lastname@monash.edu, [3]firstname.lastname@rmit.edu.au

Abstract

We present a study of a dataset of tables from biomedical research publications. Our aim is to identify characteristics of biomedical tables that pose challenges for the task of extracting information from tables, and to determine which parts of research papers typically contain information that is useful for this task. Our results indicate that biomedical tables are hard to interpret without their source papers due to the brevity of the entries in the tables. In many cases, unstructured text segments, such as table titles, footnotes and non-table prose discussing a table, are required to interpret the table's entries.

1 Introduction

Automation of information extraction (*IE*) from biomedical literature has become an important task (Shatkay and Craven, 2012). In particular, biomedical *IE* enables the semi-automation of tasks such as document indexing (Aronson et al., 2004) and database curation, e.g., (Donaldson et al., 2003; Karamanis et al., 2008).

Most research in biomedical *IE* has concentrated on information extraction from prose. However, much important data, such as experimental results and relations between biomedical entities, often appear only in tables (Ansari et al., 2013). This insight was confirmed experimentally for the task of mutation database curation. In particular, Wong *et al.* (2009) showed that for a sample of research articles used to populate the *Mismatch Repair* database (Woods et al., 2007), tables served as a sole source of information about mutations for 59% of the documents. Yepes and Verspoor (2013) reported that a text mining tool applied to full articles and their supplementary material, used to catalogue mutations in the *COSMIC* (Bamford et al., 2004) and *InSiGHT* (Plazzer et al., 2013) databases, could recover only 3-8% of the mutations if only prose was considered. An additional 1% of the mutations was extracted from tables in the papers, with an improvement of mutation coverage to about 50% when supplementary material (mostly tables) was considered.

Information extraction from tables (*Table IE*) comprises various tasks, such as (1) classification of table entries or columns into a set of specific classes (Quercini and Reynaud, 2013; Wong et al., 2009); (2) association of table entries or columns with concepts from a domain vocabulary (Assem et al., 2010; Yosef et al., 2011); and (3) extraction of relations, defined in a vocabulary, between entities in tables – usually done with Task 2 (Hignette et al., 2009; Limaye et al., 2010; Mulwad et al., 2013; Venetis et al., 2011). These tasks are often performed by consulting external knowledge sources. However, despite the intuition that unstructured text accompanying tables often provides helpful information, little use has been made of such text. Examples of such usage are the works of Yosef *et al.* (2011), who performed collective named entity normalisation in Web texts and tables; Hignette *et al.* (2009), who employed table titles to improve relation extraction from Web tables; and Govindaraju *et al.* (2013), who improved performance in extracting a few predefined relations from papers in Economics, Geology and Petrology by processing jointly the text and tables in the papers.

This paper describes the first step of a project that aims to automatically perform Tasks 2 and 3 on biomedical tables. In this step, we manually analyse a dataset of tables from the biomedical literature to identify characteristics of biomedical tables that pose challenges for column annotation, and determine the parts of a research paper that typically contain information which is useful for interpreting tables.

Our results show that tables in biomedical research papers are generally hard to interpret without their source papers due to the brevity of the entries in the tables. Further, in many cases, un-

Tatyana Shmanina, Lawrence Cavedon and Ingrid Zukerman. 2014. Challenges in Information Extraction from Tables in Biomedical Research Publications: a Dataset Analysis. In *Proceedings of Australasian Language Technology Association Workshop*, pages 118−122.

structured text (e.g., table titles, footnotes and non-table prose discussing a table) must be considered to disambiguate table entries.

2 Analysis Design

The dataset used in our analysis comprises a set of biomedical research papers discussing genetic variation. To build the dataset, we randomly sampled five articles from each of the three datasets used in (Wong et al., 2009) and (Yepes and Verspoor, 2013). The resulting sample contains 39 tables, with a total of 280 columns.

We manually analysed the dataset to collect statistics regarding typical data types in the tables (Section 3.1). Columns in the tables were annotated with *Semantic Types* (*STs*) from the *Unified Medical Language System* (*UMLS*), which has 133 *STs* in total. To assign a label to a column in a table, the annotator first located a specific *UMLS* concept corresponding to a fine-grained type of the entities listed in the column (e.g., "[C0009221] Codon (nucleotide sequence)" for Columns 3-7 in Figure 1), after which the *ST* corresponding to the selected concept was assigned to the column (e.g., "Nucleotide Sequence [T086]"). Individual data entries were not annotated due to insufficient coverage of specific values (e.g., mutations) in *UMLS*, and the predominantly numerical nature of the data (Section 3.1).

On the basis of our annotation, we gathered statistics regarding issues that may influence the performance of an automatic Table IE system, e.g., the consistency of the data types in tables (Section 3.2), and the sources of information that are useful for concept annotation (Section 3.3). It is worth noting that the annotator (first author of the paper) had little background in biomedical science at the time of annotation, and employed external sources such as NCBI databases[1] and Wikipedia to assist with the annotation. This lack of biomedical background may have affected the accuracy of the disambiguation of biomedical entities. However, we posit that the obtained results provide more relevant insights into the use of non-table components in automatic Table IE than those obtained from expert annotation.

3 Results

3.1 Content of the Data Entries

We analysed our dataset to determine which data types are typically contained in biomedical tables. It was previously noted that, in general, table entries contain very little text, which often does not provide enough context for entity disambiguation (Limaye et al., 2010). Unlike the interpretation of noun phrases, interpreting numerical data is the biggest challenge for Table IE, because numbers are highly ambiguous (in principle they could be assigned most of the *UMLS STs*). Another significant challenge in both general and biomedical *IE* is the use of abbreviations.

In light of the above, our analysis shows that biomedical tables are very difficult to interpret:

- 42% of the columns in our sample contain numbers, and 3% contain numerical expressions (e.g., $45/290$ and 45 ± 6), both representing information such as statistical data, percentages, times, lengths, patient IDs and DNA sequences (e.g., codons 175, 176 and 179 in Column 3 in Figure 1).
- 32% of the columns comprise abbreviated entries (e.g., *MSI*, *N* and *A* in Figure 1) and symbolic representations (e.g., $\pm$ for *heterozygote*).
- 7% of the columns contain free text.
- Only 12% of the columns comprise biomedical terms as entries.
- The remaining 4% of the columns contain a mixture of abbreviations, free text, and numerical expressions.

Our study shows that numerical and abbreviated entries can be interpreted correctly if they are appropriately expanded using mentions from table titles, footnotes and prose. For example, in the table in Figure 1, the abbreviations *MSI*, *N* and *A* can be expanded using the table footnote; and codon mentions in Columns 3-7 can be expanded using the prose describing the table (highlighted).[2]

3.2 Quality of the Column Headers

We analysed our dataset to determine whether it is possible to identify types of biomedical table entries based only on the content of column headers. To do so, we first identified the number of cases where column headers were sufficient for column type identification during the manual table annotation phase (Section 2). We determined that although 97% of the columns in our sample have headers, in many cases they are too ambiguous to be used as the only evidence for the column type.

[1] http://www.ncbi.nlm.nih.gov/

[2] It was impossible to determine that Columns 3-7 in Figure 1 referred to codons without the prose.

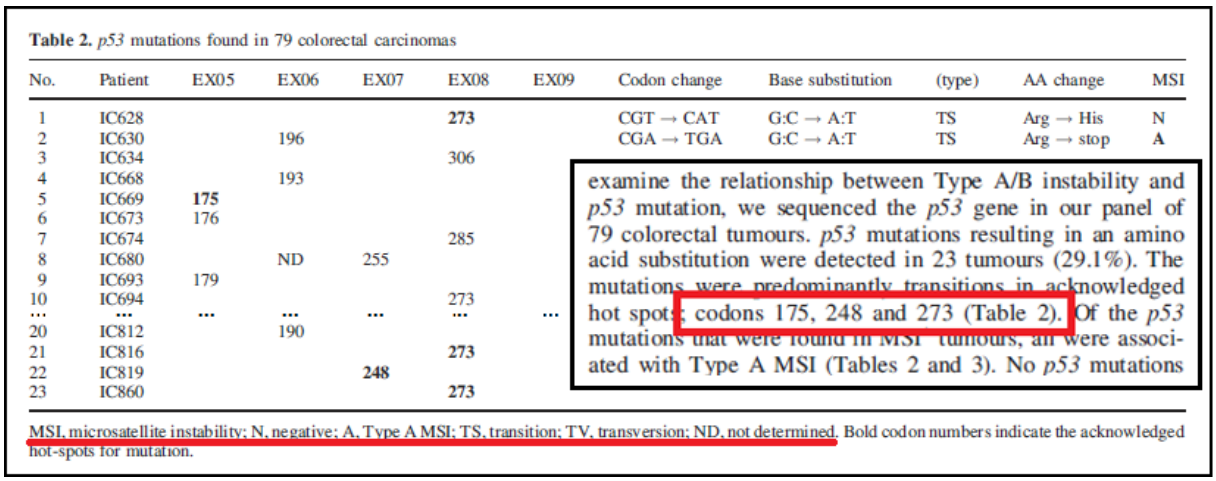

Table 2. *p53* mutations found in 79 colorectal carcinomas

No.	Patient	EX05	EX06	EX07	EX08	EX09	Codon change	Base substitution	(type)	AA change	MSI
1	IC628				**273**		CGT → CAT	G:C → A:T	TS	Arg → His	N
2	IC630		196				CGA → TGA	G:C → A:T	TS	Arg → stop	A
3	IC634				306						
4	IC668		193								
5	IC669	**175**									
6	IC673	176									
7	IC674				285						
8	IC680		ND	255							
9	IC693	179									
10	IC694				273						
...	...	...	...	...	...	...					
20	IC812		190								
21	IC816				**273**						
22	IC819			**248**							
23	IC860				**273**						

MSI, microsatellite instability; N, negative; A, Type A MSI; TS, transition; TV, transversion; ND, not determined. Bold codon numbers indicate the acknowledged hot-spots for mutation.

examine the relationship between Type A/B instability and *p53* mutation, we sequenced the *p53* gene in our panel of 79 colorectal tumours. *p53* mutations resulting in an amino acid substitution were detected in 23 tumours (29.1%). The mutations were predominantly transitions in acknowledged hot spots; codons 175, 248 and 273 (Table 2). Of the *p53* mutations that were found in MSI tumours, all were associated with Type A MSI (Tables 2 and 3). No *p53* mutations

Figure 1: An example of a biomedical table and prose discussing the table. Source: (Oda et al., 2005)

In fact, only 34% of the columns in our sample could be annotated without referring to parts of the documents other than the column entries and their headers. In 57% of the cases, additional information was required to confirm the type of a column (e.g., Columns 3-7 in Figure 1), and in 9% of the cases, headers were not helpful in column type identification (Table 1). This finding agrees with observations in the Web domain, e.g., (Limaye et al., 2010; Venetis et al., 2011).

We then compared the labels (*STs*) assigned to table columns to the *STs* of the entities in the corresponding headers. The comparison showed that in only 53% of the cases a header was labeled with the same *ST* as the entries in the column. For instance, Columns 3-7 in Figure 1 contain entities of the class "Codon" (*ST* "Nucleotide Sequence [T086]"), while the headers, which designate exons, have the *ST* "Nucleic Acid, Nucleotide, or Nucleotide [T114]" or "Biologically Active Substance [T123]". We therefore conclude that, in general, headers in isolation are insufficient, and often misleading, for column type identification.

3.3 Sufficiency and Criticality of Information Sources for Column Annotation

We analysed the dataset to determine the contribution of different sources of information in a table and its source article to the identification of the types of biomedical table entries. To this effect, we found it useful to consider the following information sources for each column: (1) the content of the data entries in the column, (2) the header of the column, (3) the headers of other columns, (4) the title of the table, (5) table footnotes, and (6) prose describing the content of the table (referred to as "prose" for simplicity). We distinguish between two aspects of these sources: *sufficiency* and *criticality*.

- The ***sufficiency*** categories are: (1) *Sufficient*, if the source on its own was enough to identify the column label; (2) *Insufficient*, if the source allowed the formulation of a hypothesis about the column label, but required information from other sources to confirm the hypothesis; and (3) *Non-indicative*, if the source did not contribute to the column labelling.
- The ***criticality*** categories are: (1) *Critical*, if disregarding the source is very likely to lead to an annotation error; (2) *Probably Critical*, if disregarding the source may lead to an annotation error; and (3) *Non-critical*, if the source could be disregarded without causing an error.

Criticality was assigned to each information source in an incremental manner depending on the sufficiency of the source: if some "cheap" sources of information were sufficient for column type identification, more "expensive" sources were not considered to be critical. The cost of a source was based on the complexity of the methods required to locate and process this source, increasing in the following order: column header, other headers, table title, table footnotes and prose.

To illustrate these ideas, consider Column 3 (concept "Codon") in Figure 1. The other headers, table title and footnotes were classified as *Non-indicative*, and hence *Non-critical*, since they do not contain any explicit information regarding the column type ("codon" is mentioned in the footnote in a sentence about formatting, which is not considered at present). The header and prose were classified as *Insufficient*, because each merely suggests the column class, and *Critical*, because both

Information source	S	IS	NI
—	1%	18%	81%
Column header	34%	57%	9%
Other column headers	0%	22%	78%
Table title	3%	36%	61%
Table footnotes	12%	34%	54%
Prose	13%	62%	25%

Table 1: Percentages of cases where sources of information were characterised as *Sufficient* (*S*), *Insufficient* (*IS*) and *Non-indicative* (*NI*) if considered in addition to the content of the column.

Information source	C	PC	NC
Column content	19%	0%	81%
Column header	87%	4%	9%
Other column headers	8%	10%	82%
Table title	15%	16%	69%
Table footnotes	27%	10%	63%
Prose	28%	20%	52%

Table 2: Percentages of cases where sources of information were characterised as *Critical* (*C*), *Probably Critical* (*PC*) and *Non-critical* (*NC*).

were required to label the column. When annotating Column 8 ("Codon change", *ST* "Genetic Function [T045]"), the title was classified as *Probably Critical*, because there was no direct correspondence with any *UMLS* concept – the mapping was performed intuitively, and the title confirmed the chosen hypothesis.

The results of our analysis are summarised in Tables 1 and 2, which respectively show statistics regarding the sufficiency and criticality of various sources of information. The results in Table 1 indicate that none of the information sources were sufficient for each table column in our dataset when taken in isolation. However, it was possible to label every column when all the sources were considered jointly. It is worth noting that the combination of the information sources that enabled labelling all the columns of a single table varied from table to table.

As seen in Table 2, each type of unstructured text associated with tables (i.e., table titles, footnotes and prose) was characterised as critical or probably critical in a substantial number of cases. In addition, we observed that in 59.3% of the cases, a table title or prose segments were characterised as critical or probably critical; and in 70.9% of the cases a table title, footnotes or prose were critical or probably critical.

Table footnotes represent an important source of information for abbreviation expansion: 97% of the tables in our sample have footnotes in the form of unstructured text, and about 62% of the footnotes introduce at least some of the abbreviations in the tables. Further, about 72% of the footnotes contain remarks associated with column headers or data entries. No other uses of footnotes were identified.

The prose that was required to interpret the tables during annotation was found in referencing paragraphs (i.e., containing descriptors such as "(Table 4)") in 70% of the cases; in 22% of the cases the prose was found elsewhere in the sections containing referencing paragraphs; and in 8% of the cases it was found elsewhere in the source document.

Our analysis shows that table titles, footnotes and prose tend to be complementary and, in general, none of them can be disregarded during annotation (Tables 1 and 2). For example, although all the tables in our sample have titles, on average only 40% of the columns in each table are represented in the titles — column "representatives" are either not mentioned in the titles, or their entity types in the titles differ from the types of the columns.

We therefore conclude that all unstructured text associated with biomedical tables (i.e., table titles, footnotes and prose) is vital for interpreting them.

4 Conclusion

In this paper, we presented an analysis of a dataset of tables from biomedical research papers performed from the perspective of information extraction from tables. Our results show that tables in biomedical research papers are characterised by an abundance of numerical and abbreviated data, for which existing approaches to Table IE do not perform well. Further, we ascertained that in many cases, unstructured text (e.g., table titles, footnotes and non-table prose discussing a table) must be considered in order to disambiguate table entries, and determine the types of table columns.

We conclude that considering unstructured text related to tables – in particular, combining existing techniques for the interpretation of stand-alone tables with IE from unstructured text – will improve the performance of Table IE. In the near future, we propose to develop techniques for locating table descriptions in the full text of source articles, and incorporating text processing techniques into approaches to Table IE.

Acknowledgments

We would like to thank the anonymous reviewers for their very detailed and insightful comments.

NICTA is funded by the Australian Government through the Department of Communications and by the Australian Research Council through the ICT Centre of Excellence Program.

References

S. Ansari, R. E. Mercer, and P. Rogan. 2013. Automated phenotype-genotype table understanding. In *Contemporary Challenges and Solutions in Applied Artificial Intelligence*, pages 47–52. Springer.

A. R. Aronson, J. G. Mork, C. W. Gay, S. M. Humphrey, and W. J. Rogers. 2004. The NLM Indexing Initiative's Medical Text Indexer. *Medinfo*, 11(Pt 1):268–72.

M. Van Assem, H. Rijgersberg, M. Wigham, and J. Top. 2010. Converting and annotating quantitative data tables. In *The Semantic Web–ISWC 2010*, pages 16–31. Springer.

S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, A. Flanagan, J. Teague, P. A. Futreal, M. R. Stratton, and R. Wooster. 2004. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British Journal of Cancer*, 91(2):355–358.

I. Donaldson, J. Martin, B. de Bruijn, C. Wolting, V. Lay, B. Tuekam, S. Zhang, B. Baskin, G. D. Bader, K. Michalickova, T. Pawson, and C. WV. Hogue. 2003. PreBIND and Textomy – mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, 4(1):11.

V. Govindaraju, C. Zhang, and C. Ré. 2013. Understanding tables in context using standard NLP toolkits. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 658–664.

G. Hignette, P. Buche, J. Dibie-Barthélemy, and O. Haemmerlé. 2009. Fuzzy annotation of Web data tables driven by a domain ontology. In *The Semantic Web: Research and Applications*, pages 638–653. Springer.

N. Karamanis, R. Seal, I. Lewin, P. McQuilton, A. Vlachos, C. Gasperin, R. Drysdale, and T. Briscoe. 2008. Natural Language Processing in aid of FlyBase curators. *BMC Bioinformatics*, 9(1):193.

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

V. Mulwad, T. Finin, and A. Joshi. 2012. A domain independent framework for extracting linked semantic data from tables. In *Search Computing*, pages 16–33. Springer.

V. Mulwad, T. Finin, and A. Joshi. 2013. Semantic message passing for generating linked data from tables. In *The Semantic Web – ISWC 2013*, pages 363–378. Springer.

S. Oda, Y. Maehara, Y. Ikeda, E. Oki, A. Egashira, Y. Okamura, I. Takahashi, Y. Kakeji, Y. Sumiyoshi, K. Miyashita, Y. Yamada, Y. Zhao, H. Hattori, K. Taguchi, T. Ikeuchi, T. Tsuzuki, M. Sekiguchi, P. Karran, and M. A. Yoshida. 2005. Two modes of microsatellite instability in human cancer: differential connection of defective DNA mismatch repair to dinucleotide repeat instability. *Nucleic Acids Research*, 33(5):1628–1636.

J. P. Plazzer, R. H. Sijmons, M. O. Woods, P. Peltomäki, B. Thompson, J. T. Den Dunnen, and F. Macrae. 2013. The InSiGHT database: utilizing 100 years of insights into Lynch Syndrome. *Familial Cancer*, 12(2):175–180.

G. Quercini and C. Reynaud. 2013. Entity discovery and annotation in tables. In *Proceedings of the 16th International Conference on Extending Database Technology*, EDBT '13, pages 693–704, New York, NY, USA. ACM.

H. Shatkay and M. Craven. 2012. *Mining the biomedical literature*. MIT Press.

P. Venetis, A. Halevy, J. Madhavan, M. Paşca, W. Shen, F. Wu, G. Miao, and C. Wu. 2011. Recovering semantics of tables on the Web. *Proceedings of the VLDB Endowment*, 4(9):528–538.

W. Wong, D. Martinez, and L. Cavedon. 2009. Extraction of named entities from tables in gene mutation literature. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 46–54. Association for Computational Linguistics.

M. O. Woods, P. Williams, A. Careen, L. Edwards, S. Bartlett, J. R. McLaughlin, and H. B. Younghusband. 2007. A new variant database for mismatch repair genes associated with Lynch Syndrome. *Human Mutation*, 28(7):669–673.

A. Jimeno Yepes and K. Verspoor. 2013. Towards automatic large-scale curation of genomic variation: improving coverage based on supplementary material. In *BioLINK SIG 2013*, pages 39–43, Berlin, Germany, July.

M. A. Yosef, J. Hoffart, I. Bordino, M. Spaniol, and G. 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.