Table Data Extraction for Text Mining in Neuroscience Papers

Yoshinobu Kano[1,*]

[1] Faculty of Informatics, Shizuoka University, Japan
kano@inf.shizuoka.ac.jp

Abstract. In neuroscience papers, one of the key findings described is a mapping between coordinate values and brain functions. When trying to extract such mappings, it is inevitable to extract table data where coordinate values are described. However, table data extraction from scientific literatures is still a difficult issue because there are various different formats for table formatting, in addition to the variation of the original paper formats. We report current status of table data extraction, discussing our future roadmap for text mining in neuroscience papers.

Keywords: text mining, scientific literatures, neuroscience, brain, table.

1 Introduction

While there have been many attempts to process scientific literatures, we cannot apply the same method to different domains due to the difference of text characteristics and required outputs. Furthermore, journal publishers just started providing full text data for text mining in these couple of years. Processing full text data requires quite different issues to be resolved compared to abstract data; abstract data was the main text data to process if full text data is unavailable.

Regarding our target domain, we focus on neuroscience papers to make our goal of outputs clear. In many neuroscience papers, their main findings are relationships between brain function and coordinate locations in brain which are in charge of that specific function. The coordinate values are normally described in tables. Because such relationships are derived from experimental results, there are several attributes of relationships described, such as experimental conditions. Both the tables and these attributes are mostly described in papers' text body, not in abstracts. Therefore, full text is required to extract the coordinate-function relationships.

These coordinate-function relationships are useful for neuroscience researchers. The NeuroImaging Platform held by Riken BSI provides searching GUI with brain atlas[1]. This system is based on the coordinate-function relationships that are manually extracted by researchers from neuroscience papers. Automatic extraction will provide advantages both in efficiency and effectiveness, extracting more detailed information.

[1] https://nimg.neuroinf.jp/modules/nimgsearch/

As a first step of our automatic coordinate-function relationship extraction, we obtained full text data and tried to extract table data from them. Although it seems straightforward for humans to read table data, it is not so easy for machines to read. Original formats of papers could be HTML, XML, or PDF depending on journals. If the table parts are explicitly marked by HTML or XML tags, there would be not much issues to extract table data. However, many papers are provided in PDFs, which basically describe everything in coordinates in a page, e.g. lines and texts, not as meaningful groups like tables. Tables are described in different formats, e.g. surrounding borders may omitted, or sometimes double lined. We need to extract numerical and textual data from these tables to make the coordinate-function relationships.

We explain current issues in obtaining full text data, both in license and technical issues in Section 2. We show our ongoing implementations for table data extractions in Section 3. Text mining and assumed applications are described as future work in Section 4. We conclude our paper in Section 5.

2 Obtaining Full Text Papers

2.1 License Issues

When publishers started to provide scientific literatures in digital formats, only abstract part of the papers were available in most cases. For example, PubMed[2] has been providing a huge amount of abstracts with their citation data. While full text data has been increasingly provided for these years, it was limited to open access journals or open access papers due to their license issues. PubMed Central (PMC)[3] is one of such collections of full text papers.

Recently, text mining has been recognized as important application for scientific papers. After long active discussions in the academic community, publishers just started to provide full text data for text mining even when they require subscriptions. Many publishers say any full text data is available for text mining, for subscribed users. However, it is not accurate in many cases. Academic subscribers normally make contracts via their libraries as institutional subscription. Publishers, e.g. Elsevier and Nature, may have special terms to allow text mining in the institutional contracts. After adding the special terms, the text mining access will be legally possible. Sometimes additional registration is required for individual users.

2.2 API Access and Crawling

Each publisher has an individual API specification. To avoid redundant work, there is

[2] http://www.ncbi.nlm.nih.gov/pubmed
[3] http://www.ncbi.nlm.nih.gov/pmc/?

a common gateway API called CrossRef[4]. We use CrossRef to obtain a list of full papers and obtain full text data. Available full text data could be in XML, HTML, and PDF depending on publication date, journal, and publisher.

We plan to obtain full text data from Elsevier, Wiley, and Springer-Nature. Springer and the Nature Publishing Group were merged in 2015 as Springer-Nature. We use Wiley's journal papers but not use others until the contract issue is resolved.

We implemented a full text data crawler which fetches full text data of specified papers via the CrossRef API. We plan to obtain relevant papers by listing neuroscience journals in the order of impact factors.

3 Table Data Extraction

It is straightforward to extract table data from HTML as they are explicitly described in table tags. However, PDF has difficulty in extracting table data. We use PDFMiner[5] to convert PDF into HTML, then try extracting table data from the converted HTML documents. Unfortunately, converted HTML documents includes a variety of expressions for tables, depending on styles of journals. While there has been efforts to extract table attributes by graphical methods (Harper & Agrawala, 2014) in a generic way, precise manual tuning would be required to capture differences in various table styles.

3.1 Bordered Table with Individual Columns of X, Y, Z

When a converted table has a border surrounding that entire table, we detect headers with “x”, “y”, “z”, then take the cell strings as coordinate values. This extraction fails when the cell values are in multi-lines, concatenated, or empty. Figure 1 (right) shows an example of bordered table with individual columns.

3.2 Bordered Table with a Single Column of X, Y, Z

When a converted table has a border surrounding that entire table, we detect headers with “x, y, z”, then take the cell strings as coordinate values. The header name is not always “x, y, z”, but has a variety of different expressions. Therefore, we cannot rely on the header name. We use the value pattern to determine whether the cell is for the coordinate values or not. The cell value is assumed to be concatenated with comma, slash, or a single space. Figure 1 (left) shows an example of bordered table with individual columns.

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

[5] http://www.unixuser.org/~euske/python/pdfminer/

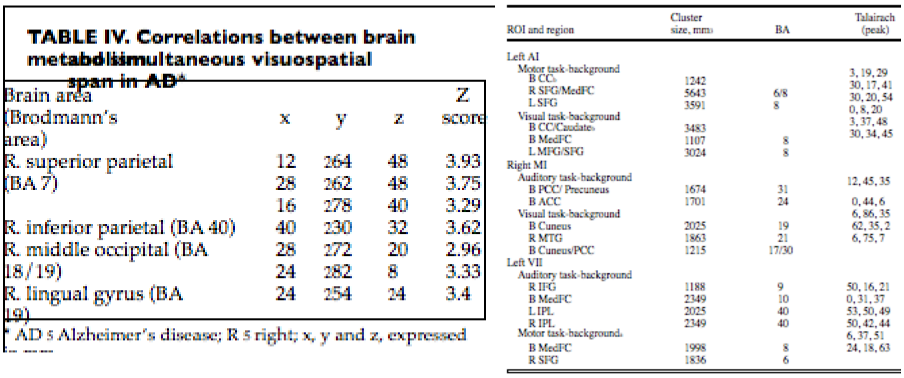

TABLE IV. Correlations between brain metabolism and simultaneous visuospatial span in AD*

Brain area (Brodmann's area)	x	y	z	Z score
R. superior parietal (BA 7)	12	264	48	3.93
	28	262	48	3.75
	16	278	40	3.29
R. inferior parietal (BA 40)	40	230	32	3.62
R. middle occipital (BA 18/19)	28	272	20	2.96
	24	282	8	3.33
R. lingual gyrus (BA 19)	24	254	24	3.4

* AD 5 Alzheimer's disease; R 5 right; x, y and z, expressed

ROI and region	Cluster size, mm	BA	Talairach (peak)
Left AI			
Motor task-background			3, 19, 29
B CC	1242		30, 17, 41
R SFG/MedFC	5643	6/8	30, 20, 54
L SFG	3591	8	0, 8, 20
Visual task-background			3, 37, 48
B CC/Caudate	3483		30, 34, 45
B MedFC	1107	8	
L MFG/SFG	3024	8	
Right MI			
Auditory task-background			12, 45, 35
B PCC/ Precuneus	1674	31	
B ACC	1701	24	0, 44, 6
Visual task-background			6, 86, 35
B Cuneus	2025	19	62, 35, 2
R MTG	1863	21	6, 75, 7
B Cuneus/PCC	1215	17/30	
Left VII			
Auditory task-background			
R IFG	1188	9	50, 16, 21
B MedFC	2349	10	0, 31, 37
L IPL	2025	40	53, 50, 49
R IPL	2349	40	50, 42, 44
Motor task-background			6, 37, 51
B MedFC	1998	8	24, 18, 63
R SFG	1836	6	

Figure 1. Examples of a bordered table (left) taken from (Collette, et al., 1997) and horizontal lines table (right) taken from (Tian et al., 2007).

3.3 Table with Horizontal Lines

When a converted table is represented with a pair of horizontal lines over and under the table, we detect its table caption as "Table X", then take the cell strings as coordinate values. A first row between horizontal lines is regarded as a header row. Figure 1 (right) is an example of such cases, also an example of the single column pattern.

3.4 Extraction Results

We selected 1240 PDF samples from Wiley journal papers. Using the methods above, 598 tables were successfully extracted, which are manually verified.

4 Conclusion and Future Work

As our first step to perform text mining for neuroscience papers, we searched current APIs to obtain full text papers, implemented a crawler system and table data extraction system. This table data extraction system aims to extract coordinate values in brain. The next step would be to obtain attributes related to the table values, which forms coordinate-function relationships. Covering more journals would also be needed.

Acknowledgements

This work was partially supported by JST CREST.

References

Collette, F., Salmon, E., Van der Linden, M., Degueldre, C., & Franck, G. (1997). Functional anatomy of verbal and visuospatial span tasks in Alzheimer's disease. *Human Brain Mapping*, *5*(2), 110–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10096415

Harper, J., & Agrawala, M. (2014). Deconstructing and restyling D3 visualizations. *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology - UIST '14*, 253–262.

Tian, L., Jiang, T., Liang, M., Li, X., He, Y., Wang, K., … Jiang, T. (2007). Stabilities of negative correlations between blood oxygen level-dependent signals associated with sensory and motor cortices. *Human Brain Mapping*, *28*(7), 681–90. http://doi.org/10.1002/hbm.20300