Table parsing notes

Alex Ratner

July 8, 2015

1 Motivating examples

1.1 SEC Filings

Item 2. Properties.

The Company leases its principal executive offices in Deerfield, Illinois. The following table indicates the principal properties of the Company and its subsidiaries as of December 31, 2012:

	Owned	Leased
Production and Warehouse Facilities:		
Canada	6	—
France	8	—
India	1	13
Mexico	3	6
Spain	11	—
U.K.	3	—
U.S. – Kentucky	13	2
U.S. – Maine	3	—
U.S. Virgin Islands	3	—
Total	51	21
Distribution Facilities:		
Germany (EMEA)	—	1
India (APSA)	—	7
Spain (EMEA)	1	—
Total	1	8
Total	52	29

The production and warehouse facilities listed above support the operations of each of our segments. In addition to the leased property located in Deerfield, Illinois, the Company also leases properties located in Australia, India, Mexico, and Spain related to corporate and administrative functions.

Figure 1: Property ownership overview table from Beam Inc's 10-K

Suppose we wish to extract (Production facility location, Production facility quantity) tuples, for owned properties, from SEC filings. Suppose that we define candidate entities for this tuple as any capitalized phrase and number respectively. We would hope to ultimately extract e.g. (U.S.-Kentucky, 13), but not (U.S.-Kentucky, 2), (Germany (EMEA), -), or (Total, 51).

A single line / sentence, corresponding to our atomic processing unit, might look like:

```
<tr><td>\tU.S. - Kentucky</td><td>13</td><td>2</td></tr>
```

The crux of our task is then to enable the use of *global* features i.e. feature function inputs from beyond this line, having to do with the overall structure of the table, such as:

```
<b><u>Owned</u></b>
Production and Warehouse Facilities:
<b>Item 2.\tProperties.
CATEGORY_HEADER=False
SUMMARY_LINE=False
```

We see here that **parsing the table's structure** is a tertiary task to our secondary task of **extracting *attributes*-** e.g. column labels, row labels, etc.- connected to candidate entities in our primary **relation extraction** task.

One hypothesis is that to accomplish this, a simple parsing of the table structure can be used- i.e. in contrast to something like a 2D PCFG like in [6]. For example in [5], a CRF is used to label rows of ASCII text tables with certain table type labels (including 'not a table', crucial in their scenario). Especially with data from a slightly more structured format, e.g. XML or HTML, where columns are more deterministicly resolved, these types of labels might be sufficient information for determining which attributes are related to a given entity / cell.

The rough idea would then be to do this table row categorization task **jointly** with our primary relation extraction task.

1.2 Genomics / Scientific Literature

Table 1 Clinical features of patients with a *GNAO1* variant

	Patient 1	*Patient 2*	*Patient 3*	*Patient 4*
Age, gender	20 months, female	14 months, female	13 years, female	18 years, female
Variants[a]	c.680C>T, p.(Ala227Val)	c.607G>A, p.(Gly203Arg)	c.736G>A, p.(Glu246Lys)	c.625C>T, p.(Arg209Cys)
Diagnosis	Early-onset epileptic encephalopathy	Early-onset epileptic encephalopathy	Movement disorder, intellectual disability with developmental delay	Movement disorder, intellectual disability with developmental delay
Initial symptom	Infantile spasms at 2 months	Tonic-clonic seizures at 7 days	Developmental delay at 4 months	Developmental delay at 7 months
Initial EEG	Hypsarrhythmia	Slow-wave bursts, migrating focal epileptiform discharges	No abnormalities at 12 years	No abnormalities at 4 years
Course of seizures	Complex partial seizures	Complex partial seizures	No seizures	Complex partial seizures at 10 and 11 years
Course of EEG	Changed to multifocal with ictal	Multifocal epileptiform discharges with right hemisphere dominancy	NA	Diffused low activity
Intractable seizures	+	+	—	—
Involuntary movement	Hand stereotypies	Severe chorea	Severe athetosis	Severe chorea
Development				
Head control	—	—	—	3 months to 10 years
Sitting	—	—	—	11 months to 10 years
Meaningful words	—	—	—	5 years to 10 years
MRI	Progressive cerebral atrophy, thin corpus callosum at 10 months	Normal at 20 days; Progressive cerebral atrophy with delayed myelination at 14 months	Normal at 4 and 12 years	Progressive cerebral and cerebellar atrophy, brainstem atrophy, thin corpus callosum

Abbreviations: EEG, electroencephalography; MRI, magnetic resonance imaging; NA, not assessed.
[a]*GNAO1* variants were annotated based on transcript variant 1 (NM_020988.2).

Figure 2: A table in a recent paper, used by our clinical collaborators to identify a critical genetic variant for embryonic screening

In this example we again see challenging issues such as in-table embedded hierarchy, etc. However here we may also have to deal with significant challenges in the *row and column resolution* stages- the above table is taken from a PDF, where it was rotated 90 degrees, has cells which span multiple rows with no clear e.g. line delineation, etc.

2 Task components & related work

1. **Row parsing** (from ASCII with CRFs [5])
2. **Column parsing** (as global opt. problem [2])
3. **Table relevance classification**
4. **Table structure parsing** (with 2D PCFGS [6]; for NL query execution [4])
5. **Relation extraction** (jointly with surrounding text [3]; using global corpus statistics [1])

3 Notes

- **Expansion of table relevance classification task:** Prior work has approached the task of classifying whether a section of text- either demarcated or not specifically as a

table- is of relevance for extraction. Most commonly this has been to make a table vs. non-table distinction (e.g. in ASCII text) or a content vs. non-content decision (e.g. HTML tables used for formating vs. content). In our case however we may want to use global features to do inference over whether the table is related at all to our extraction schema

- **Simplification of table parsing grammar:** If our task is primarily focused on the end goal of relation extraction rather than on e.g. global table extraction & collation, then we may be best served by a simple 'grammar' at least to start. For example, a 2D grammar like the one proposed in [6], which is able to formally distinguish between *flat*, *hierarchical*, and *multidimensional* tables, may be more than is needed for the type of data illustrated above. Proper row and column parsing (done jointly perhaps) and row tagging of the type done with CRFs in [5]- e.g. identifying contiguous spanning rows, header, subheader & etc rows...- may be sufficient.

- **Joint inference:** Over the sections defined in Section 2 above. We could start with XML and HTML tables only and thus restrict consideration to sections 3-5. The high-level intuition, besides the obvious advantages of doing inference jointly (if tractable), is the rough idea of side-stepping work that is orthogonal to our end gola of relation extraction

- **Inference which allows hard constraints + global moves:** If we do inference jointly, then using something closer to MC-SAT- or a modified version which considers a specific set of 'moves' e.g. a defined MH proposal function specific to table relation extraction- might be much better than any simpler Gibbs (simple, block or tree) scheme

- **Visual features:** Could try these if/when we focus on PDF, OCR data...

4 Problem statement in more detail

Goal: To build a TRUC.

4.1 Relation extraction

We consider the problem setup of *relation extraction* or *knowledge base creation*. In this task we are given a dataset, a schema of *relations* we wish to extract, and either weights over this schema or labeled data to learn these weights from.

More concretely, let $\mathcal{D}$ be a data corpus, and assume we have a *candidate extraction* function C. Let us call $x \in C(\mathcal{D})$ a *candidate*. A *relation* is then a function $R \in \mathcal{R}$ which operates over one or more candidates and returns a value. For simplicity we will consider the binary case $R : C(\mathcal{D})^n \mapsto \{0, 1\}$, where generally we will consider $n \in \{1, 2\}$, i.e. unary and binary relations, and where the binary output can be interpreted as whether or not a set of candidates constitutes a relation R.

4.2 Extraction of relations from tables

When dealing with text datasets, some of the candidates x will occur within semi-structured fomrats- we consider *tables*. We wish to handle these occurences specially, e.g. define features unique to these candidates and the relations over them. We consider only the features unique to these candidates (i.e. not lexical, global context, type, etc. features), in other words *table structure features*. There are several challenges to extracting such structure and features:

4.2.1 Table identification

Correctly identifying relevant tables (versus normal text, tables used to define e.g. HTML page layout, etc.) is its own challenge which has been approached with both heuristic and machine learning strategies in previous works [**?**].

4.2.2 Table extraction

Given a block of data identified as a table, extracting a grid of table *cells* may be trivial- e.g. given HTML or XML formats- or may be highly non-trivial; some examples of the latter include:

- Table in image e.g. OCR
- PDF format
- Raw text format
- List format ([2])

4.2.3 Table structure parsing

We wish to parse the structure of a table, at least precisely enough / specifically in order to extract features over sets of candidates within the table, which can be used for relation extraction.

4.3 Joint inference

We propose to do the table extraction, structure parsing and relation extraction **jointly**.

References

[1] MJ Cafarella, A Halevy, DZ Wang, and E Wu. Webtables: exploring the power of tables on the web. *Proceedings of the VLDB ...*, 2008.

[2] X Chu, Y He, K Chakrabarti, and K Ganjam. Tegra: Table extraction by global record alignment. *Proceedings of the 2015 ACM ...*, 2015.

[3] V Govindaraju, C Zhang, and C Ré. Understanding tables in context using standard nlp toolkits. *ACL (2)*, 2013.

[4] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *ACL 2015*, 2015.

[5] D Pinto, A McCallum, X Wei, and WB Croft. Table extraction using conditional random fields. *Proceedings of the 26th annual ...*, 2003.

[6] Dekai Wu and Ken Lee. A grammatical approach to understanding textual tables using two-dimensional scfgs. page 905–912, 2006.