Understanding Tables in Context Using Standard NLP Toolkits

Vidhya Govindaraju   **Ce Zhang**   **Christopher Ré**
University of Wisconsin-Madison
{vidhya, czhang, chrisre}@cs.wisc.edu

Abstract

Tabular information in text documents contains a wealth of information, and so tables are a natural candidate for information extraction. There are many cues buried in both a table and its surrounding text that allow us to understand the meaning of the data in a table. We study how natural-language tools, such as part-of-speech tagging, dependency paths, and named-entity recognition, can be used to improve the quality of relation extraction from tables. In three domains we show that (1) a model that performs joint probabilistic inference across tabular and natural language features achieves an F1 score that is twice as high as either a pure-table or pure-text system, and (2) using only shallower features or non-joint inference results in lower quality.

1   Introduction

Tabular data is ubiquitous and often contains high-quality, structured relational data. Recent studies found billions of high-quality relations on the web in HTML (Cafarella et al., 2008). In financial applications, a huge amount of data is buried in the tables of corporate filings and earnings reports; in science, millions of journal articles contain billions of scientific facts in tables. Although tables describe precise, structured relations, tables are rarely written in a way that is self-describing, e.g., tables may contain abbreviations or only informal schema information; in turn, the contents of tables are often ambiguously specified, which makes extracting the relations implicit in tabular data difficult.

Tables are, however, not written in isolation. The text surrounding a table in a journal article explains its contents to its intended audience, a human reader. For example, in a simple study, we demonstrate that humans can achieve more than 60% higher recall by jointly reading the text and tables in a journal article than by only looking at the tables. The conclusion of this experiment is not surprising, but it raises a question: *How should a system combine tabular and natural-language features to understand tables in text?*

The literature provides a broad spectrum of answers to this question. Most previous approaches use textual or tabular features separately, e.g., tabular approaches that do not use text features (Dalvi et al., 2012; Wu and Lee, 2006; Pinto et al., 2003) or textual approaches that do not use tabular features (Mintz et al., 2009; Wu and Weld, 2010; Poon and Domingos, 2007). In a prescient study, Liu et al. (2007) proposed to learn the target relation independently from both table and surface textual features, and then combine the result using a linear combination of the predictions.

In a similar spirit, we propose to use both types of features in our approach of relation extraction. Our proposed approach differs from prior approaches in two ways: (1) We use deeper–but standard–NLP features than prior approaches for table extraction. In contrast to the shallow, lexical features that prior approaches have used, we use standard NLP features, such as dependency paths, parts of speech, etc. Our hypothesis is that a deeper understanding of the text in which a table is embedded will lead to higher quality table extraction. (2) Our probabilistic model jointly uses both tabular and textual features. One advantage of a joint approach is that one can predict portions of the complicated predicate that is buried in a table. For example, in a geology journal article, we may read a measure-
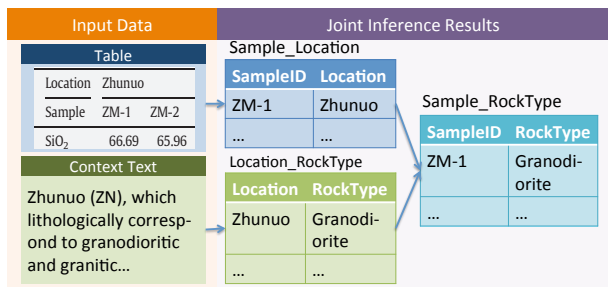
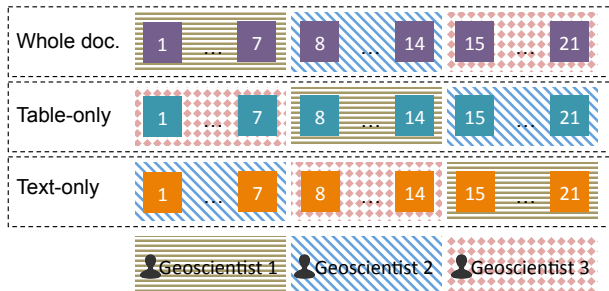Figure 1: An example of joint inference between a table and its context.

Figure 2: Job assignments for the human study.

Figure 3: Human quality to extract SAMPLE-ROCKTYPE relations in PETDB.

Task	TEXT	TABLE	JOINT
NER	POS tags Stanford NER Regular Expression Dictionary	pdftotable NER of neighbor cells Regular expression Dictionary # columns	Whether a mention in table also appears in the text.
EL	POS tags Bing query results Freebase Stanford Parser	Pdftotable Bing query results Freebase	Subjective mentions in the sentence near a table
RE	Dependency path Term proximity Word sequence	Table headers Table subheaders RE of neighbor rows	Join between relations (See Figure 1 for an example)

Figure 4: List of features we used in TEXT, TABLE, and JOINT approaches. **NER**, **EL**, and **RE** refer to named-entity recognition, entity linking, and relation extraction, respectively.

ment in a table that tells us the type of rock and its weight—but data such as the location where this rock was unearthed and in what geological time interval this rock appeared may not be specified in the table.

We consider tasks in three domains: PETROLOGY, FINANCE, and GEOLOGY. For each domain, we build a system to extract relations from text, tables, or both. We found that a joint inference system that uses non-shallow, but standard NLP features can significantly improve the quality of the extracted relations, and *that this result holds consistently across all three domains*. For example, in our Petrology application to extract a knowledge base, called PETDB[1], by using information extracted from both text and tables, we can achieve twice as high F1 compared to either a pure-table or pure-text system.

2 Motivating Human Study

We describe a simple human study that motivated our approach to jointly combine both tabular features and natural language features to extract relations from tables. The hypoth-
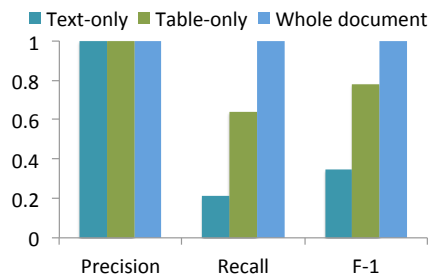
esis that we want to validate is that the text surrounding a table could provide valuable information even for a human reader, and therefore, an ideal machine reading system should also try to capture similar information.

We asked three geoscientists to manually read journal articles and extract relations for the PETROLOGY domain. We report our results for the target relation, SAMPLE-ROCKTYPE, which associates a rock type with a rock sample (see Figure 1 for an example). We randomly sampled 21 journal articles. For each journal article, we produced three variants: (1) the original document; (2) *table-only*, which is the set of tables in the document (without the text); (3) *text-only*, which is the text of the document with the tables removed from the document. Each geoscientist was asked to read and extract the relations from one of the three variants. We then judged the precision and recall of their extraction, as shown in Figure 2.

[1] http://www.earthchem.org/petdb

659

As shown in Figure 3, human readers not surprisingly achieve perfect precision on each of the variants, but lower recall on both the table-only and text-only variants. However, summing the recall of table-only (60%) and text-only (20%) variants together would achieve only 80% recall; this implies that in the best case more than 20% of the extractions require that the human reader read the table *and* its surrounding text *jointly*. Figure 1 shows one representative example.

This motivates our approach, which uses a *joint* inference system to model features from a table and its surrounding text. We also propose to use deep linguistic features instead of shallower features to get as close as possible to the ability of human readers in understanding the surrounding text of a table.

3 Empirical Study & Experiments

We describe our experiments to test the hypothesis that (1) deeper linguistic features can help to extract higher quality relations from tables, and (2) joint inference across tables and text improves extraction quality compared to approaches that use pure-table, pure-text, and non-joint ways of combining these two. We briefly describe some experiments for a dataset that we call GEOLOGY (Zhang et al., 2013). The detailed experimental results in all three domains are in the technical report version of this paper.

3.1 Experimental Setup

We consider the task of constructing a geology knowledge base. Specifically, our goal is to extract a ROCK-TOTALORGANICCARBON relation that maps rock formations (e.g., "Barnett Formation") to their total organic carbon (e.g., "6%"). Such data is important for estimating stored energy and for global climate research.

Dataset. We selected 100 geology journal articles.[2] We asked three geoscientists to annotate these journal articles manually to extract the ROCK-TOTALORGANICCARBON relation (1.5K tuples). We processed each document using Stanford CoreNLP (de Marneffe et al., 2006; Toutanova and Manning, 2000),

PDFtoHTML[3], and pdf2table (Yildiz, 2004). We then extracted features following state-of-the-art practices (see Figure 4).

Approaches. To validate our hypothesis, we implement four systems, each of which has access to different types of data:

(1) TABLE. This approach follows Pinto et al. (2003) and Dalvi et al. (2012) and only uses the tables in a document.

(2) TEXT. This approach only has access to the text in a document and contains all the features mentioned in Wu and Weld (2010) and Mintz et al. (2009).

The features used in (1) and (2) are shown in Figure 4. In both TABLE and TEXT, we use a conditional random field (Lafferty et al., 2001) model for the ROCK-TOTALORGANICCARBON relation.

(3) MERGE. Using TABLE and TEXT, we extract all facts and their associated probability. Following Duin (2002), we combine these two probabilities using a linear combination. MERGE is a baseline approach that uses information from both tables and text.

(4) JOINT. We build a joint approach that uses information from both tables and text. This approach is a large factor graph in which we embed the CRFs developed in TABLE and TEXT. Additionally, we allow JOINT to predict projections of each relation, as shown in Figure 4. Recall that a key advantage of a joint approach is that we do not need to predict all arguments of the relation (if such a prediction is unwarranted from the data). The inference is done by Gibbs sampling using our inference engine ELEMENTARY (Zhang and Ré, 2013). We describe the JOINT system in more detail in the technical report version of this paper.

3.2 End-to-End Quality

We were able to validate that JOINT achieves higher quality than the other three approaches we considered. Figure 5 shows the P/R curve of different approaches on three domains. We analyzed the domain GEOLOGY.

JOINT dominates all other approaches. At a recall of 10%, JOINT achieves 3x higher precision than all other approaches. In our error analysis, we saw that tables in geology articles often contain ambiguous words; for example,

---

[2]We choose a set of documents that (1) are in English, and (2) contain at least one table.

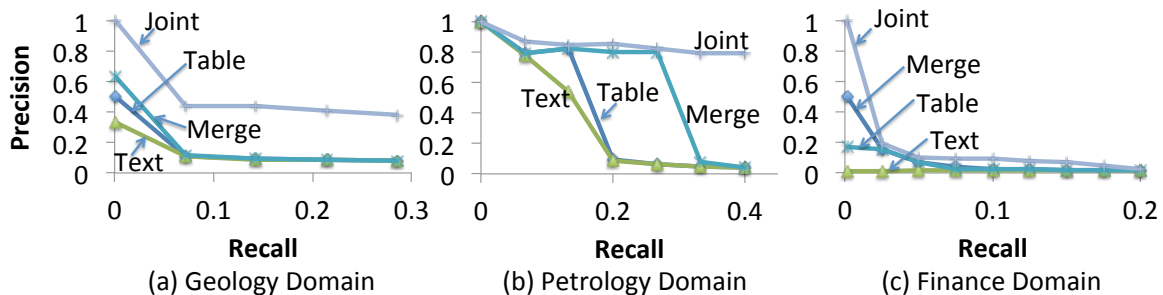[3]http://pdftohtml.sourceforge.net/

Figure 5: End-to-end extraction quality on PETROLOGY, FINANCE, and GEODEEPDIVE. The recall is limited by the quality of state-of-the-art table recognition software on PDFs.

the word "Barnett" in a table may refer to either a location or a rock formation. By using features extracted from text, JOINT achieves higher precision. For recall in the range of 0–10%, MERGE outperforms both TEXT and TABLE, with 3%–90% improvement in precision.

In GEOLOGY, MERGE has precision that is similar to TEXT and TABLE for the higher recall range (>10%). In this domain, we found that relations that appeared in the text often repeated relations described in the table. In other domains, such as PETROLOGY, where the relations in text and tables have lower degrees of overlap, MERGE significantly improves over TEXT and TABLE (Figure 5(b)).

We conducted a statistical significance test to check whether the improvement of JOINT over the three other approaches is statistically significant. For each of the three probability thresholds, $t \in \{.99, .90, .50\}$, we created the set of predictions that JOINT assigns probability greater than t. Figure 6 shows the results of the statistical significance test in which the null hypothesis is that the F1 scores of two approaches are the same. With $p = 0.01$, JOINT has statistically significant improvement of F1 score over all three other approaches with each probability threshold.

3.3 Shallow vs. Linguistic Features

We validate the hypothesis that using linguistic features, e.g., part-of-speech tags (Toutanova and Manning, 2000), named-entity tags (Finkel et al., 2005), and dependency trees (de Marneffe et al., 2006), helps improve the quality of our approach, called JOINT. There are different ways to use shallow and linguistic features; we select

Approaches \ Prob.	.99	.90	.50
TEXT	+	+	+
TABLE	+	+	+
MERGE	+	+	+

Figure 6: Approximate randomization test from Chinchor (1992) of F1 score with $p = 0.01$ on the impact of joint inference compared with pure-table or pure-text approaches for different probability thresholds. A + sign indicates that the F1 score of joint approach increased significantly.

Type	Features
Shallow	Regular Expressions (Dalvi et al., 2012) Term proximity (Matsuo et al., 2003) Dictionary and Freebase (Mintz et al., 2009)
Linguistic	POS tags (Wu et al., 2010) Stanford NER tags (Mintz et al., 2009) Dependence trees (Mintz et al., 2009)

Figure 7: Types of Features.

state-of-the-art approaches from the literature (see Figure 7).

We created the following variants of JOINT. JOINT$^{(-PARSE)}$ removes features generated by the dependency parser and syntax parser. Similarly, JOINT$^{(-NER)}$ (JOINT$^{(-POS)}$) removes all features related to NER (resp. POS). JOINT$^{(-POS)}$ also removes NER and parser features because the latter two are dependent on POS features.

Figure 8 shows the P/R curve for all these variants on GEOLOGY, and Figure 9 shows the results of statistical significance test. For probability threshold .90, JOINT outperforms JOINT$^{(-POS)}$ significantly. The difference between JOINT, JOINT$^{(-PARSE)}$,

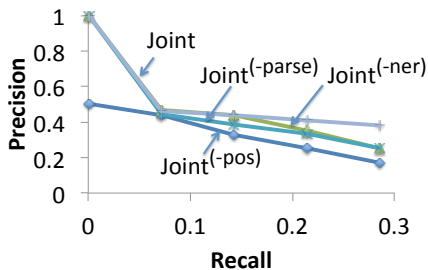Figure 8: Lesion study of different features for GEOLOGY.

Features \ Prob.	.90	.50
JOINT$^{(-\text{PARSE})}$ → JOINT	0	+
JOINT$^{(-\text{NER})}$ → JOINT	0	+
JOINT$^{(-\text{POS})}$ → JOINT	+	+

Figure 9: Approximate randomization test of F1 score with $p = 0.01$ on the impact of linguistic features. For $x \rightarrow y$, a + indicates that the F1 score of y is significantly higher than x. 0 indicates that the F1 score does not change significantly.

and JOINT$^{(-\text{NER})}$ is not significant because there are "easy-to-extract" facts in the high-probability range. For probability threshold .50, JOINT outperforms all three other variants significantly.

4 Related Work

The intuition that context features might help table-related tasks has existed for decades. For example, Hurst and Nasukawa (2000) mentioned (as future work) that context features could be used to further improve their relation extraction approaches from tables. Lin et al. (2010) use bag-of-words features and hyperlinks to recommend new columns for web tables. Liu et al. (2007) extract features, including font size and title, from PDF documents in which a table appears to help the table ranking task. They find that these features only contribute less than 2% to precision. In contrast, in our approach linguistic features are quite useful. The above approaches use context features that can be extracted without POS tagging or linguistic parsing. One aspect of our work is to demonstrate that traditional NLP tools can enhance the quality of table extraction.

Extracting information from tables has been discussed by different communities in the last decade, including NLP (Wu and Lee, 2006; Tengli et al., 2004; Chen et al., 2000), artificial intelligence (Fang et al., 2012; Pivk, 2006), information retrieval (Wei et al., 2006; Pinto et al., 2003), database (Cafarella et al., 2008), and the web (Dalvi et al., 2012). This body of work considers only features derived from tables and does not examine richer NLP features as we do.

While joint inference is popular, it is not clear when a joint inference system outperforms a more traditional NLP pipeline. Recent studies have reached a variety of conclusions: in some, joint inference helps extraction quality (McCallum, 2009; Poon and Domingos, 2007; Singh et al., 2009); and in some, joint inference hurts extraction quality (Poon and Domingos, 2007; Eisner, 2009). Our intuition is that joint inference is helpful in this application because our joint inference approach combines non-redundant signals (textual versus tabular).

5 Conclusion

To improve the quality of extractions of tabular data, we use standard NLP techniques to more deeply understand the text in which a table is embedded. We validate that deeper NLP features combined with a joint probabilistic model has a statistically significant impact on quality, i.e., recall and precision. Our ongoing work is to apply these ideas to a much larger corpus from each of the three domains.

6 Acknowledgments

References

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

Hsin-Hsi Chen, Shih-Chung Tsai, and Jin-He Tsai. 2000. Mining tables from large scale HTML texts. In *Proceedings of the 18th Conference on Computational Linguistics*, COLING '00.

Nancy Chinchor. 1992. The statistical significance of the MUC-4 results. In *Proceedings of the 4th Conference on Message Understanding*, MUC4 '92.

Bhavana Bharat Dalvi, William Cohen, and Jamie Callan. 2012. WebSets: Extracting sets of entities from the web using unsupervised information extraction. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, WSDM '12.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

Robert Duin. 2002. The combining classifier: to train or not to train? In *16th International Conference on Pattern Recognition*.

Jason Eisner. 2009. Joint models with missing data for semi-supervised learning. In *NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing*.

Jing Fang, Prasenjit Mitra, Zhi Tang, and C. Lee Giles. 2012. Table header detection and classification. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, AAAI '12.

Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05.

Matthew Hurst and Tetsuya Nasukawa. 2000. Layout and language: Integrating spatial and linguistic knowledge for layout understanding tasks. In *Proceedings of the 18th Conference on Computational Linguistics*, COLING '00.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Cindy Xide Lin, Bo Zhao, Tim Weninger, Jiawei Han, and Bing Liu. 2010. Entity relation discovery from web tables and links. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10.

Ying Liu, Kun Bai, Prasenjit Mitra, and C. Lee Giles. 2007. TableSeer: Automatic table metadata extraction and searching in digital libraries. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '07.

Andrew McCallum. 2009. Joint inference for natural language processing. In *Proceedings of the 13th Conference on Computational Natural Language Learning*, CoNLL '09.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, ACL '09.

David Pinto, Andrew McCallum, Xing Wei, and W. Bruce Croft. 2003. Table extraction using conditional random fields. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03.

Aleksander Pivk. 2006. Automatic ontology generation from web tabular structures. *AI Communication*, 19(1).

Hoifung Poon and Pedro Domingos. 2007. Joint inference in information extraction. In *Proceedings of the 22nd National Conference on Artificial intelligence*, AAAI'07.

Sameer Singh, Karl Schultz, and Andrew McCallum. 2009. Bi-directional joint inference for entity resolution and segmentation using imperatively-defined factor graphs. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, ECML PKDD '09.

Ashwin Tengli, Yiming Yang, and Nian Li Ma. 2004. Learning table extraction from examples. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04.

Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing*, EMNLP '00.

Xing Wei, Bruce Croft, and Andrew McCallum. 2006. Table extraction for answer retrieval. *Information Retrieval*, 9(5).

Dekai Wu and Ken Wing Kuen Lee. 2006. A grammatical approach to understanding textual tables using two-dimensional scfgs. In *Proceedings of the COLING/ACL*, COLING-ACL '06.

Fei Wu and Daniel S. Weld. 2010. Open information extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10.

Burcu Yildiz. 2004. Information extraction – utilizing table patterns. Master's thesis, Institutfür Softwaretechnik und Interaktive Systeme.

Ce Zhang and Christopher Ré. 2013. Towards high-throughput Gibbs sampling at scale: A study across storage managers. SIGMOD '13.

Ce Zhang, Vidhya Govindaraju, Jackson Borchardt, Tim Foltz, Christopher Ré, and Shanan Peters. 2013. GeoDeepDive: Statistical inference using familiar data-processing languages. SIGMOD '13.