Rule-based Information Extraction is Dead! Long Live Rule-based Information Extraction Systems!

Laura Chiticariu
IBM Research - Almaden
San Jose, CA
chiti@us.ibm.com

Yunyao Li
IBM Research - Almaden
San Jose, CA
yunyaoli@us.ibm.com

Frederick R. Reiss
IBM Research - Almaden
San Jose, CA
frreiss@us.ibm.com

Abstract

The rise of "Big Data" analytics over unstructured text has led to renewed interest in information extraction (IE). We surveyed the landscape of IE technologies and identified a major disconnect between industry and academia: while rule-based IE dominates the commercial world, it is widely regarded as dead-end technology by the academia. We believe the disconnect stems from the way in which the two communities measure the benefits and costs of IE, as well as academia's perception that rule-based IE is devoid of research challenges. We make a case for the importance of rule-based IE to industry practitioners. We then lay out a research agenda in advancing the state-of-the-art in rule-based IE systems which we believe has the potential to bridge the gap between academic research and industry practice.

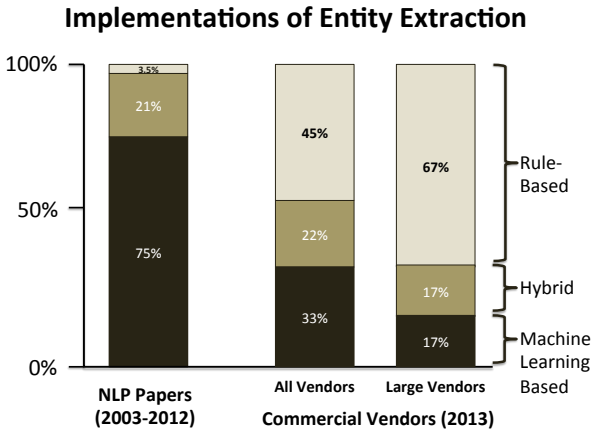

Figure 1: Fraction of NLP conference papers from EMNLP, ACL, and NAACL over 10 years that use machine learning versus rule-based techniques to perform entity extraction over text (left); the same breakdown for commercial entity extraction vendors one year after the end of this 10-year period (right). The rule-based approach, although largely ignored in the research community, dominates the commercial market.

1 Introduction

The recent growth of "Big Data" analytics over large quantities of unstructured text has led to increased interest in information extraction technologies from both academia and industry (Mendel, 2013).

Most recent academic research in this area starts from the assumption that statistical machine learning is the best approach to solving information extraction problems. Figure 1 shows empirical evidence of this trend drawn from a survey of recent published research papers. We examined the EMNLP, ACL, and NAACL conference proceedings from 2003 through 2012 and identified 177 different EMNLP research papers on the topic of entity extraction. We then classified these papers into three categories, based on the techniques used: purely rule-based, purely machine learning-based, or a hybrid of the two. We focus on entity extraction, as it is a classical IE task, and most industrial IE systems offer this feature.

The left side of the graph shows the breakdown of research papers according to this categorization. Only six papers relied solely on rules to perform the extraction tasks described. The remainder relied entirely or substantially on statistical techniques. As shown in Figure 2, these fractions were roughly constant across the 10-year period studied, indicating that attitudes regarding the relative importance of the different techniques have remained constant.

We found that distinguishing "hybrid" systems

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 827–832,
Seattle, Washington, USA, 18-21 October 2013.

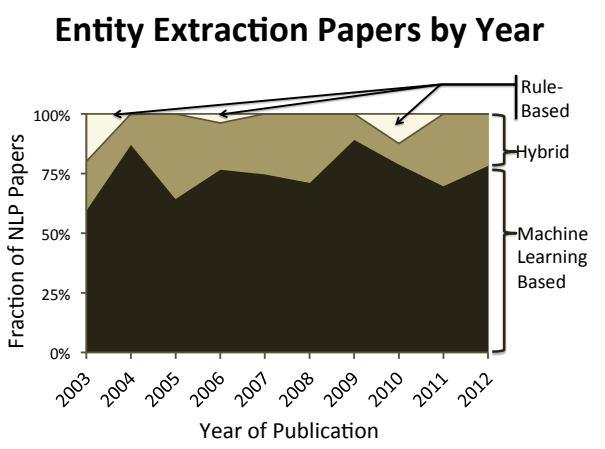

Figure 2: The conference paper data (left-hand bar) from Figure 1, broken down by year of publication. The relative fractions of the three different techniques have not changed significantly over time.

from pure machine learning systems was quite challenging. The papers that use a mixture of rule-based and machine learning techniques were generally written so as to obfuscate the use of rules, emphasizing the machine learning aspect of the work. Authors hid rules behind euphemisms such as "dependency restrictions" (Mausam et al., 2012), "entity type constraints" (Yao et al., 2011), or "seed dictionaries" (Putthividhya and Hu, 2011).

In the commercial world, the situation is largely reversed. The right side of Figure 1 shows the result of a parallel survey of commercial entity extraction products from 54 different vendors listed in (Yuen and Koehler-Kruener, 2012). We studied analyst reports and product literature, then classified each product according to the same three categories. Table 1 shows the 41 products considered in the study [1]. We conducted this industry survey in 2013, one year after the ten-year run of NLP papers we studied. One would expect the industrial landscape to reflect the research efforts of the previous 10 years, as mature technology moved from academia to industry. Instead, results of this second survey showed the opposite effect, with rule-based systems comprising the largest fraction of those surveyed. Only 1/3 of the vendors relied entirely on machine learning. Among public companies and private compa-

[1] Other products do not offer entity extraction, or we did not find sufficient evidence to classify the technology.

Table 1: Vendors and products considered in the study.

ai-one	*NathanApp*
Attensity	*Command Center*
Basis Technology	*Rosette*
Clarabridge	*Analyze*
Daedalus	*Stilus NER*
GATE	*Information Extraction*
General Sentiment	
HP	*Autonomy IDOL Eduction*
IBM	*InfoSphere BigInsights Text Analytics*
IBM	*InfoSphere Streams Text Analytics*
IBM	*SPSS Text Analytics for Surveys*
IntraFind	*iFinder NAMER*
IxReveal	*uHarmonize*
Knime	
Language Computer	*Cicero LITE*
Lexanalytics	*Salience*
alias-i	*LingPipe*
Marklogic	*Analytics & Business Intelligence*
MeshLabs	*eZi CORE*
Microsoft	*FAST Search Server*
MotiveQuest	
Nice Systems	*NiceTrack Open Source Intelligence*
OpenAmplify	*Insights*
OpenText	*Content Analytics*
Pingar	
Provalis Research	*WordStat*
Rapid-I	*Text Processing Extension*
Rocket	*AeroText*
salesforce.com	*Radian 6*
SAP	*HANA Text Analysis*
SAS	*Text Analytics*
Serendio	
Smartlogic	*Semaphore Classification and Text Mining Server*
SRA International	*NetOwl Text Analytics*
StatSoft	*STATISTICA Text Miner*
Temis	*Luxid Content Enrichment Platform*
Teradata	*(integration w/ Attensity)*
TextKernel	*Extract!*
Thompson Reuters	*OpenCalais*
Veda	*Semantics Entity Identifier*
ZyLab	*Text Mining&Analytics*

Table 2: Pros and Cons

	Pros	Cons
Rule-based	• Declarative • Easy to comprehend • Easy to maintain • Easy to incorporate domain knowledge • Easy to trace and fix the cause of errors	• Heuristic • Requires tedious manual labor
ML-based	• Trainable • Adaptable • Reduces manual effort	• Requires labeled data • Requires retraining for domain adaptation • Requires ML expertise to use or maintain • Opaque

nies with more than $100 million in revenue, the situation is even more skewed towards rule-based systems, with large vendors such as IBM, SAP, and Microsoft being completely rule-based.

2 Explaining the Disconnect

What is the source of this disconnect between research and industry? There does not appear to be a lack of interaction between the two communities. Indeed, many of the smaller companies we surveyed were founded by NLP researchers, and many of the larger vendors actively publish in the NLP literature. We believe that the disconnect arises from a difference in how the two communities measure the costs and benefits of information extraction.

Table 2 summarizes the pros and cons of machine learning (ML) and rule-based IE technologies (Atzmueller and Kluegl, 2008; Grimes, 2011; Leung et al., 2011; Feldman and Rosenfeld, 2006; Guo et al., 2006; Krishnan et al., 2005; Yakushiji et al., 2006; Kluegl et al., 2009). On the surface, both academia and commercial vendors acknowledge essentially the same pros and cons for the two approaches. However, the two communities weight the pros and cons significantly differently, leading to the drastic disconnect in Figure 1.

Evaluating the benefits of IE. Academic papers evaluate IE performance in terms of precision and recall over standard labeled data sets. This simple, clean, and objective measure is useful for judging competitions, but the reality of the business world is much more fluid and less well-defined.

In a business context, definitions of even basic entities like "product" and "revenue" vary widely from one company to another. Within any of these ill-defined categories, some entities are more important to get right than others. For example, in electronic legal discovery, correctly identifying names of executives is much more important than finding other types of person names.

In real-world applications, the output of extraction is often the input to a larger process, and it is the quality of the larger process that drives business value. This quality may derive from an aspect of extracted output that is only loosely correlated with overall precision and recall. For example, does extracted sentiment, when broken down and aggregated by product, produce an unbiased estimate of average sentiment polarity for each product?

To be useful in a business context, IE must function well with metrics that are ill-defined and subject to change. ML-based IE models, which require a careful up-front definition of the IE task, are poor fit for these metrics. The commercial world greatly values rule-based IE for its interpretability, which makes IE programs easier to adopt, understand, debug, and maintain in the face of changing requirements (Kluegl et al., 2009; Atzmueller and Kluegl, 2008). Furthermore, rule-based IE programs are valued for allowing one to easily incorporate domain knowledge, which is essential for targeting specific business problems (Grimes, 2011). As an example, an application may pose simple requirements to its entity recognition component to output only full person names, and not include salutation. With a rule-based system, such a requirement translates to removing a few rules. On the other hand, a ML-based approach requires a complete retrain.

Evaluating the costs of IE. In a business setting, the most significant costs of using information extraction are the *labor cost* of developing or adapting extractors for a particular business problem, and the *hardware cost* of compute resources required by the system.

NLP researchers generally have a well-developed sense of the labor cost of writing extraction rules, viewing this task as a "*tedious and time-consuming process*" that "*is not really practical*" (Yakushiji et al., 2006). These criticisms are valid, and, as we

point out in the next section, they motivate a research effort to build better languages and tools.

But there is a strong tendency in the NLP literature to ignore the complex and time-consuming tasks inherent in solving an extraction problem using machine learning. These tasks include: defining the business problem to be solved in strict mathematical terms; understanding the tradeoffs between different types of models in the context of the NLP task definition; performing feature engineering based on a solid working understanding of the chosen model; and gathering extensive labeled data — far more than is needed to measure precision and recall — often through clever automation.

All these steps are time-consuming; even highly-qualified workers with postgraduate degrees routinely fail to execute them effectively. Not surprisingly, in industry, ML-based systems are often deemed risky to adopt and difficult to understand and maintain, largely due to model opaqueness (Fry, 2011; Wagstaff, 2012; Malioutov and Varshney, 2013). The infeasibility of gathering labeled data in many real-world scenarios further increases the risk of committing to a ML-based solution.

A measure of the system's scalability and runtime efficiency, hardware costs are a function of two metrics: throughput and memory footprint. These figures, while extremely important for commercial vendors, are typically not reported in NLP literature. Nevertheless, our experience in practice suggests that ML-based approaches are much slower, and require more memory compared to rule-based approaches, whose throughput can be in the order of MB/second/core for complex extraction tasks like NER (Chiticariu et al., 2010).

The other explanation. Finally, we believe that the most notable reason behind the academic community's steering away from rule-based IE systems is the (false) perception of lack of research problems. The general attitude is one of *"What's the research in rule-based IE? Just go ahead and write the rules."* as indicated by anecdotal evidence and only implicitly stated in the literature, where any usage of rules is significantly underplayed as explained earlier. In the next section, we strive to debunk this perception.

3 Bridging the Gap

As NLP researchers who also work regularly with business customers, we have become increasingly worried about the gap in perception between information extraction research and industry. The recent growth of Big Data analytics has turned IE into big business (Mendel, 2013). If current trends continue, the business world will move ahead with unprincipled, ad-hoc solutions to customers' business problems, while researchers pursue ever more complex and impractical statistical approaches that become increasingly irrelevant. Eventually, the gap between research and practice will become insurmountable, an outcome in neither community's best interest.

The academic NLP community needs to stop treating rule-based IE as a dead-end technology. As discussed in Section 2, the domination of rule-based IE systems in the industry is well-justified. Even in their current form, with ad-hoc solutions built on techniques from the early 1980's, rule-based systems serve the industry needs better than the latest ML techniques. Nonetheless, there is an enormous untapped opportunity for researchers to make the rule-based approach more principled, effective, and efficient. In the remainder of this section, we lay out a research agenda centered around capturing this opportunity. Specifically, taking a systemic approach to rule-based IE, one can identify a set of research problems by separating rule development and deployment. In particular, we believe research should focus on: (a) data models and rule language, (b) systems research in rule evaluation and (c) machine learning research for learning problems in this richer target language.

Define standard IE rule language and data model. If research on rule-based IE is to move forward in a principled way, the community needs a standard way to express rules. We believe that the NLP community can replicate the success of the SQL language in connecting data management research and practice. SQL has been successful largely due to: (1) *expressivity*: the language provides all primitives required for performing basic manipulation of structured data, (2) *extensibility*: the language can be extended with new features without fundamental changes to the language, (3) *declarativity*: the language allows the specification of com-

putation logic without describing its control flow, thus allowing developers to code *what* the program should accomplish, rather than *how* to accomplish it.

An earlier attempt in late 1980's to formalize a rule language resulted in the Common Pattern Specification Language (CPSL) (Appelt and Onyshkevych, 1998). While CPSL did not succeed due to multiple drawbacks, including expressivity limitations, performance limitations, and its lack of support for core operations such as part of speech (Chiticariu et al., 2010), CPSL did gain some traction, e.g., it powers the JAPE language of the GATE open-source NLP system (Cunningham et al., 2011). Meanwhile, a number of declarative IE languages developed in the database community, including AQL (Chiticariu et al., 2010; Li et al., 2011), xLog (Shen et al., 2007), and SQL extensions (Wang et al., 2010; Jain et al., 2009), have shown that formalisms of rule-based IE systems are possible, as exemplified by (Fagin et al., 2013). However, they largely remain unknown in the NLP community.

We believe now is the right time to establish a standard IE rule language, drawing from existing proposals and experience over the past 30 years. Towards this goal, IE researchers need to answer the following questions: What is the right data model to capture text, annotations over text, and their properties? Can we establish a standard declarative extensible rule language for processing data in this model with a clear set of constructs that is sufficiently expressive to solve most IE tasks encountered so far?

Systems research based on standard IE rule language. Standard IE data model and language enables the development of systems implementing the standard. One may again wonder, *"Where is the research in that?"* As in the database community, initial research should focus on systemic issues such as data representation and speeding up rule evaluation via automatic performance optimization. Once baseline systems are established, system-related research would naturally diverge in several directions, such as extending the language with new primitives (and corresponding optimizations), and exploring modern hardware.

ML research based on standard IE rule language. A standard rule language and corresponding execution engine enables researchers to use the standard language as the expressivity of the output model, and define learning problems for this target language, including learning basic primitives such as regular expressions and dictionaries, or complete rule sets. (One need not worry about choosing the language, nor runtime efficiency.) With an expressive rule language, a major challenge is to prevent the system from generating arbitrarily complex rule sets, which would be difficult to understand or maintain. Some interesting research directions include devising proper measures for rule complexity, constraining the search space such that the learnt rules closely resemble those written by humans, active learning techniques to cope with scarcity of labeled data, and visualization tools to assist rule developers in exploring and choosing between different automatically generated rules. Finally, it is conceivable that some problems will not fit in the target language, and therefore will need alternative solutions. However, the community would have shown – objectively – that the problem is not learnable with the available set of constructs, thus motivating follow-on research on extending the standard with new primitives, if possible, or developing novel hybrid IE solutions by leveraging the standard IE rule language together with ML technology.

4 Conclusion

While rule-based IE dominates the commercial world, it is widely considered obsolete by the academia. We made a case for the importance of rule-based approaches to industry practitioners. Drawing inspiration from the success of SQL and the database community, we proposed directions for addressing the disconnect. Specifically, we call for the standardization of an IE rule language and outline an ambitious research agenda for NLP researchers who wish to tackle research problems of wide interest and value in the industry.

Acknowledgments

We would like to thank our colleagues, Howard Ho, Rajasekar Krishnamurthy, and Shivakumar Vaithyanathan, as well as the anonymous reviewers for their thoughtful and constructive comments.

References

Douglas E. Appelt and Boyan Onyshkevych. 1998. The Common Pattern Specification Language. In *Proceedings of a workshop held at Baltimore, Maryland: October 13-15, 1998*, TIPSTER '98, pages 23–30.

Martin Atzmueller and Peter Kluegl. 2008. Rule-Based Information Extraction for Structured Data Acquisition using TextMarker. In *LWA*.

Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Sriram Raghavan, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. SystemT: An Algebraic Approach to Declarative Information Extraction. In *ACL*.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. 2011. Text Processing with GATE (Version 6), Chapter 8: JAPE: Regular Expressions over Annotations.

Ronald Fagin, Benny Kimelfeld, Frederick Reiss, and Stijn Vansummeren. 2013. Spanners: a formal framework for information extraction. In *PODS*.

Ronen Feldman and Benjamin Rosenfeld. 2006. Boosting Unsupervised Relation Extraction by Using NER. In *EMNLP*, pages 473–481.

C. Fry. 2011. Closing the Gap between Analytics and Action. *INFORMS Analytics Mag.*, 4(6):405.

Seth Grimes. 2011. Text/Content Analytics 2011: User Perspectives on Solutions. `http://www.medallia.com/resources/item/text-analytics-market-study/`.

Hong Lei Guo, Li Zhang, and Zhong Su. 2006. Empirical study on the performance stability of named entity recognition model across domains. In *EMNLP*, pages 509–516.

Alpa Jain, Panagiotis Ipeirotis, and Luis Gravano. 2009. Building query optimizers for information extraction: the sqout project. *SIGMOD Record*, 37(4):28–34.

Peter Kluegl, Martin Atzmueller, and Frank Puppe. 2009. TextMarker: A Tool for Rule-Based Information Extraction. In *UIMA@GSCL Workshop*, pages 233–240.

Vijay Krishnan, Sujatha Das, and Soumen Chakrabarti. 2005. Enhanced answer type inference from questions using sequential models. In *HLT*, pages 315–322.

Cane Wing-ki Leung, Jing Jiang, Kian Ming A. Chai, Hai Leong Chieu, and Loo-Nin Teow. 2011. Unsupervised Information Extraction with Distributional Prior Knowledge. In *EMNLP*, pages 814–824.

Yunyao Li, Frederick Reiss, and Laura Chiticariu. 2011. Systemt: A declarative information extraction system. In *ACL*.

Dmitry M. Malioutov and Kush R. Varshney. 2013. Exact rule learning via boolean compressed sensing. In *ICML*.

Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. Open Language Learning for Information Extraction. In *EMNLP-CoNLL*, pages 523–534.

Thomas Mendel. 2013. Business Intelligence and Big Data Trends 2013. `http://www.hfsresearch.com/Business-Intelligence-and-Big-Data-Trends-2013` (accessed March 28th, 2013).

Duangmanee Putthividhya and Junling Hu. 2011. Bootstrapped Named Entity Recognition for Product Attribute Extraction. In *EMNLP*, pages 1557–1567.

Warren Shen, AnHai Doan, Jeffrey F. Naughton, and Raghu Ramakrishnan. 2007. Declarative Information Extraction Using Datalog with Embedded Extraction Predicates. In *VLDB*, pages 1033–1044.

Kiri Wagstaff. 2012. Machine learning that matters. In *ICML*.

Daisy Zhe Wang, Eirinaios Michelakis, Michael J. Franklin, Minos N. Garofalakis, and Joseph M. Hellerstein. 2010. Probabilistic Declarative Information Extraction. In *ICDE*.

Akane Yakushiji, Yusuke Miyao, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2006. Automatic construction of predicate-argument structure patterns for biomedical information extraction. In *EMNLP*, pages 284–292.

Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. 2011. Structured Relation Discovery using Generative Models. In *EMNLP*, pages 1456–1466.

Daniel Yuen and Hanns Koehler-Kruener. 2012. Who's Who in Text Analytics, September.