RULE-BASED INFORMATION EXTRACTION: ADVANTAGES, LIMITATIONS, AND PERSPECTIVES

Bernhard Waltl / Georg Bonczek / Florian Matthes

Research Associate, Technical University of Munich, Department of Informatics,
Software Engineering for Business Information Systems
Boltzmannstraße 3, 85748 Garching bei München, DE
b.waltl@tum.de; http://wwwmatthes.in.tum.de

Student Assistant, Technical University of Munich, Chair for Software Engineering for Business Information Systems
Boltzmannstraße 3, 85748 Garching bei München, DE
georg.bonczek@tum.de; http://www.matthes.in.tum.de

Professor, Technical University of Munich, Department of Informatics, Software Engineering for Business Information Systems
Boltzmannstraße 3, 85748 Garching bei München, DE
matthes@tum.de; https://wwwmatthes.in.tum.de

Keywords: ***Rule-based Information Extraction, Apache UIMA, Apache Ruta, JAPE Grammar***

Abstract: *With the clear advantages and capabilities of neural networks and machine learning in general, rule-based approaches and their usage and potential for information extraction (IE) are hardly addressed by academic research. This is counter-intuitive since recent studies have shown that a majority of large vendors of information extraction software, including IBM, SAP, and Microsoft, are using this technology predominantly within their offered products and services. This paper summarizes the advantages and limitations of rule-based IE and discusses its role for legal information retrieval. In addition, several different perspectives for academic research and industrial usage are illustrated.*

1. Introduction

As more and more textual data is digitally available, the need for performant and maintainable algorithms and technologies for text mining and natural language processing grows. Most approaches on information extraction (IE) can be grouped in either category of rule-based, i.e. knowledge-based, or machine learning (ML) based technologies. A large part of information extraction systems in research are nowadays based on statistical methods using models generated by ML algorithms, leaving rule-based methodologies out of the focus of modern research on IE. However, systems based on ML are still underrepresented in industry (Chiticariu et al. 2013). Although this is not fully in-line with current trends in academic research there is strong evidence that rule-based information extraction offers huge research potential in both application and development of tools. In addition, there are many research questions, which are highly interesting for the information extraction practice, but are not addressed by academic nor industrial research.

These main research areas for the analysis of rule-based information extraction cover five different areas:

1. Expressiveness of rule and domain specific languages
2. Induction and creation of rules
3. Testing and evaluation of rules (including performance)
4. Maintenance of rules in an industrial and productive environment
5. Frameworks and tools for the management of rules and their application on large document corpora

The main focus of this short paper is to discuss the importance of domain specific languages for rule-based pattern annotation on textual data. Thereby, the focus is set on current shortcomings of pattern languages and rule-based IE in general and proposes possible research questions that address these deficiencies.

The main annotation frameworks for modern IE systems are the Apache UIMA and Gate (Wilcock 2017). Both frameworks feature a comprehensive rule language, the UIMA Ruta[1] (rule-based text annotation) and JAPE[2] respectively, which are also used in academia. Multiple other languages have emerged, adding valuable features and novel concepts addressing the shortcomings of already existent languages. Until now, little systemic research has been conducted to prepare the ecosystem of rules for the new age of machine learning methods. Showing the versatile powers of rules by creating more prominent information extraction systems using rule-based approaches, bundling research effort into a common platform and creating reusable and system independent tools, could revitalize the research interest in the promising field of rule-based information extraction.

2. Advantages of Rule-based IE

Asides from standard evaluation metrics of IE systems such as precision and recall, the techniques can also be judged by means of implementation. While statistical machine learning is widely used as a black-box technology regarding traceability transparency of decisions, rule-based approaches follow a mostly declarative approach leading to highly transparent and expressive models. Other advantages of such rule languages include readability, and maintainability and foremost the possibility of directly transferring domain knowledge into rules. The potential of directly including the knowledge of a domain expert into the IE process rather than choosing the most promising training data, hyper-parameters or weights, thus indirectly incorporating the domain knowledge, is a major advantage compared to other approaches to IE.

However, rule-based and ML approaches must not be used exclusively but can complement each other very well. As supervised learning always requires (large) amount of training data, one of the first steps is to sample and create training data sets. In domains, where pre-labelled samples are scarce or non-existent, this step is a mostly time-consuming task. To extract a great number of high quality training samples, rule languages can be used. Based on their declarative nature, the knowledge engineer is free to control the resulting precision and recall by either specifying broader rules that capture more samples but also some false examples or by specifying narrower rules that result in a smaller training data set of high quality in terms of precision. This methodology is a form of bootstrapping modern machine learning technologies and the most recent project focus on exactly this challenge. For example, the project «Snorkel» run by Stanford University[3] is a framework that allows «creating, modelling, and managing training data, currently focused on accelerating the development of structured or «dark» data extraction applications for domains in which large labelled training sets are not available or easy to obtain.» The main idea thereby is called «data programming», whereas functions, i.e. rules, are used to create data sets, which are subsequently used to train powerful and flexible machine learning approaches.

Even if rule-based IE systems require the manual implementation of rules, the manual labour implied directly translates into the quality of the rules, while ML techniques require very deep theoretical knowledge and experimentation to maximize their efficiency. This is also due to the rule's syntactic and semantic similarity to regular expressions, which makes them easier to learn and thus better to maintain by teams with different skillsets.

Using an appropriate interface, rules can be written by anyone regardless, of technical background or familiarity in rule languages or general-purpose languages. An example for such an interface is the VINERy IDE that allows for an easy visual drag and drop creation of rules (Li et al. 2015). These kinds of interfaces and tools designed to support end-users increase the usability of rules and address the disadvantage also formulated by Chiticariu et al. 2013: «the implementation of rules is tedious manual labour». This emphasizes that the

[1] https://uima.apache.org/ruta.html (all websites last accessed in January 2018).

[2] https://gate.ac.uk/releases/gate-8.4.1-build5753-ALL/doc/tao/splitch8.html#x12-2080008.

[3] https://github.com/HazyResearch/snorkel.

limitations of rule languages and rule-based information extraction systems can be mitigated by proper tool support, preferably one that is tailored to the specific applications and domains.

3. Limitations and Shortcomings

Although there are many advantages of rule-based IE, current state-of-the-art rule languages such as UIMA Ruta and JAPE also have several drawbacks. The declarative nature offers huge potential for directly applying domain knowledge to information extraction tasks. The tooling and syntax of these languages still prohibits the widespread use of these domain specific languages outside of the NLP community. Domain experts, i.e. end-users that have little technical background, find it hard to accommodate themselves with the syntax and quirks of the current mainstream rule languages. Consequently, the need for a joint development with domain experts, software engineers and so-called legal data scientists emerges.

Other drawbacks of using manually crafted rules include the amount of labour implied and the interoperability of rule languages. While the infrastructure for loading and processing data for similar tasks can be generalized in machine learning systems, only requiring a common input and output format, non-trivial rules, e.g., including more complex annotation information, can only be used if the environment is sufficiently uniform.

The efforts invested into a rule system usually lead to better precision and recall immediately. However, dealing with high linguistic variety, e.g. the classification of natural language sentences, requires a disproportionate amount of manual implementation effort to capture the whole variety of linguistic subtleties. Consequently, if the targeted patterns have many but only minor deviations, the number of rules required grows at least linearly, which is unacceptable most of the time as the code base grows into a large set of mostly duplicated rules. Rules tend to fully reconstruct the expressiveness of natural language, as more general, i.e. abstract, patterns cannot be formalized. But having these general principles of language and codifying them into rules is the main and overarching idea of rules and rule languages.

In a more general form, this is the most limiting feature of rules. Due to their declarative nature, rules inherently generalize not that well to minor variations in the input data. The largely inevitable noises and language variations pose hurdles to the application of rules. This means that in such cases when input data differences only in nuances, statistical machine learning models have the advantage of generalizing better, resulting in higher recall. As discussed above, rules will only capture the occurrences they explicitly cover. Diagnostic scores are a successfully applied method for augmenting more heuristic behaviour (Atzmueller et al. 2007; Baumeister et al. 2006), they add another dimension to the rules and are only practicable in specific use-cases.

4. Addressing the Shortcomings of Rule-based IE: A Brief Research Perspective

A first step to a revitalized research interest into rule-based information extraction would be the definition of a modern standardized rule language. This can be either accomplished by a formally well-defined pattern language or specification of a virtual machine fostering the embedding of different IE frameworks such as Apache UIMA and GATE. Such an open and common platform could spark and bundle new research interest in this platform. Possible research can address issues such as the representation and efficient indexing of annotations, automatic optimization of rules and platform independent tooling. Apart from the technical concepts and implementations, a centralized community surrounding an open standard increases the amount of documentation, tools, language patterns and training material, in contrast to the current rule languages where documentation is scarce.

To reduce the manual labour introduced, semi-automatic rule induction can be used. Based on a training set containing samples for annotation types, rules are induced, using machine learning algorithms. The generated rules can then be directly applied or used as a starting point and then be refined. While the need for a training set is introduced to rule-based development, its quality is not required to be very high if the rules are only used as a base line for implementation. The usage of these induction algorithms has yet to be evaluated for modern rule-based IE applications. While the TEXTRULER framework (Kluegl et al. 2009) implements multiple algorithms, it is only accessible through the UIMA Ruta workbench, an Eclipse plugin that provides IDE functionality for the UIMA Ruta language. With a comprehensive review of the available algorithms and methodologies for inducing rules and a platform independent framework implementing them, the use of this approach in rule-based development can be advanced to further reduce the already low manual labour implied in the development of rules.

As mentioned above, one of the main obstacle of widespread usage of machine learning in domains such as the legal one is the absence of comprehensive training data. Instead of bootstrapping the development of rule-based IE systems with machine learning, highly precise rules can be used to generate training data.

5. Conclusion

As outlined in this paper, the advantages and disadvantages of rule-based information extraction can – to a large degree – be mitigated by better tool-support to maintain the lifecycle of rules or by adding machine learning models to the system to retain the ease of developing rule-based systems with the efficiency and capabilities of machine learning methods.

Regardless of whether rule-based information extraction is used in combination with machine learning or not, they can be used in a variety of scenarios, especially in domains where labelled training data is scarce or the compilation of a data set is expensive, such as the legal domain. In addition to their usage in bootstrapping machine learning approaches, they can be used to implement transparent, understandable and precise tools for information extraction systems. Using them as a primer for more complex and flexible systems based on recent advantages in the field of machine learning, they might be a solution for the creation of a very large training data set.

In either way, rules have proven themselves over the last decades to be a viable and rewarding technology in text mining in general and will continue to be directly or indirectly relevant in many research areas. However, if the research interest continues to stagnate in creating more efficient, robust and widely adopted technologies, the community surrounding information extraction on legal documents would leave behind an auspicious instrument that could establish ways for widespread usage of machine learning based systems.

6. References

Atzmueller, M./Klügl, P./Baumeister, J./Puppe, F. (2007), October. Rapid knowledge capture using subgroup discovery with incremental refinement. In Proceedings of the 4th international conference on Knowledge capture. ACM, pp. 31–38.

Baumeister, J./Atzmueller, M./Kluegl, P./Puppe, F. (2006), Conservative and Creative Strategies for the Refinement of Scoring Rules. In FLAIRS Conference, pp. 408–413.

Chiticariu, L./Li, Y./Reiss, F.R. (2013), October. Rule-based information extraction is dead! long live rule-based information extraction systems!. In EMNLP, No. October, pp. 827–832.

Kluegl, P./Atzmueller, M./Hermann, T./Puppe, F. (2009), A Framework for Semi-Automatic Development of Rule-based Information Extraction Applications. In LWA, pp. KDML–56.

Li, Y./Kim, E./Touchette, M.A./Venkatachalam, R./Wang, H. (2015), Vinery: A visual ide for information extraction. Proceedings of the VLDB Endowment, 8(12), pp. 1948–1951.

Wilcock G. (2017), The Evolution of Text Annotation Frameworks. In: Ide N./Pustejovsky J. (eds.), Handbook of Linguistic Annotation. Springer, Dordrecht.