ExplainIE -Explaining Information Extraction systems

(Research-in-Progress)

Wojciech Barczynski, Falk Brauer, Adrian Mocan
SAP Research CEC Dresden, {firstname}.{surname}@sap.com

Abstract: Business Intelligence (BI) over unstructured text is under intense scrutiny both in the industry and research. Recent work in this field includes automatic integrating of unstructured text into business analytics, model recognition, and probabilistic databases to handle uncertainty of Information Extraction (IE). However, still an open issue is how to handle IE quality, which is a part of ETL like process for the BI. Precision of IE is still too low for BI and, according to *Sunita Sarawagi* in recent survey on IE, we are still far from a comprehensive quality model for IE. Currently the BI user has neither methodology nor tools, which would help him to discover if the result is an unexpected fact or an error in IE. In this work we present preliminary results on developing methodology and tool (*ExplainIE*), which helps users to debug unexpected results. *ExplainIE* presents results within BI tool and auxiliary view on low level detail (e.g., entity graph). We consider two kinds of users: BI and IE expert.

Key Words: Data Quality, Information Quality, Information Extraction, Business Intelligence, Unstructured text

Research Problem and Research Question

This work addresses a problem of information quality (IQ) in Information Extraction (IE), which is a part of ETL like process for BI over unstructured text. According to *Feldman* [1] accuracy of IE is 90-98% only for simple entities and it is much lower for facts and events (see [2]) - 50-60%. Furthermore it decreases when quality of input data is low. *Sarawagi*, in comprehensive survey on IE [2], remarks that it is nearly impossible to guarantee IE accuracy in real-life deployment. Moreover she provides insight on how difficult is to define IQ model (metrics), that capture the variety of IE techniques. Our experience with IE technology follows all these observations. Therefore we develop a methodology (and a tool *ExplainIE*) to explain unexpected results. Explain means to present IQ information, which are needed to assess quality of BI over unstructured text. Note that this work continues our research on unstructured data in enterprises [3]. The importance of research problem is motivated by growing importance of BI over unstructured data. Simply blogs, social networks, etc. became an integral part of our culture, hence they contain information about, e.g., us as customers. Unfortunately currently such an analysis still needs high manual effort, due to low legibility of IE. Second, BI over unstructured data should be performed by BI analytics not IE and NLP expert. Summing up, we formulate the research question as: How to explain to the BI user a complex IE process, so she or he is able to assess the correctness and quality of BI result?

State of Art

In IE area the functionality, which we propose, has not been targeted yet. There are works on assessment of rule based system quality, e.g., [4] (using principle of maximum entropy), but explain functionality for end users was not investigated. Problems like a quality of data have been already target in database community, e.g., [5] (quality driven integration). A prominent work in this community, which is close to ours, is *Uldbs* presented by *Benjelloun at el.* in [6], which tackle together data lineage and quality in one system. We deal with much bigger complexity, because we don't have homogeneous and well defined environment like databases. Even using a IE algebra (e.g., [7],[8]) we still need to handle variability of operators. Moreover we focus on user perspective on data quality. An explain functionality is under intense investigation in Expert System (ES) community. Recent work in this field was presented in [9]. *Glass et al.* propose Proof Markup Language to handle explanation in ES. The explanation can cover also IE, but that work doesn't focus on IE (treat IE as black boxes). Furthermore the domain is different, in question answer system, you don't need have massive data volume. There are also works in Semantic Web, e.g., [10] (extension to SPARQL), but they are conceptually similar to work done in ES.

Research Approach and Contribution

The aim of our work is to build *ExplainIE* system, which allows user to drill into meta information about an usage of IE. The system should assist in performing BI over unstructured text and detect what could be

the most probable reason of an error. It should also be personalize by, e.g., hiding unnecessary details from BI user and expose them to IE expert. We organize our research in three work streams: IQ problems, IQ model (dimensions), and IQ methodology (after IQ meta framework presented by *Ge et al.* in IQ [14]). First we investigate IQ problems of IE. We focus not only on metrics (precision, confidence) but also on how to leverage IE lineage to capture complex problems, such as a usage of wrong IE operator or domain knowledge. We build our solution on the top of algebraic framework (described in [7], likewise [8]), therefore we can link IQ problems to the elements of IE framework model to create one comprehensive model. Furthermore we propose to divide IQ problems in two groups – line of processing (LoP) and line of explaining (LoE) (inspired by [12]). LoP can be seen as lineage information for IE. LoE goes beyond lineage taking into account semantic dependences between elements of IE framework as well as domain knowledge. It is supported by simple reasoning mechanism, which traverse semantic around IE to detect failures. Moreover we describe IE operations in terms of: causes (why it occurs), context (what could be an influence), and consequence (what are consequences). This categorization we derived from existing work in human science [13]. These are three elements, which are important to explain why something has happened. We see this complex model as our contribution in first work stream.

Second we build quality model (quality dimensions) from the user perspective, which takes into account two kinds of users: BI and IE expert. It is motivated by the fact that each of them expects different granularity of explain information. Having quality problems and models we can create a mapping between them. The model is built based on existing literature on IQ in decision making and reviews with BI analysts. Therefore here the most interesting part is a mapping between IQ problems and IQ model.

The final part is IQ methodology, which combine outcome from two first work streams. IQ methodology includes: how to present explain information to the user , how to detect problematic situations, and how user interacts with a tool. Based on our preliminary work, we support three kinds of presentation: inside OLAP cube, breadcrumb (similar to website breadcrumb but about IE), and auxiliary view for displaying low level details. Here our contribution is detection mechanism, which works also on aggregations. Moreover IE breadcrumb generation and creation of explain cubes are challenging as well.

The general architecture of *ExplainIE* consists of three layers. Basic component is lineage mechanism, which handles LoP. On the top we place *LoE component*, which provides advance explain. It takes as an input: LoP, semantic knowledge about IE (e.g., IE plan), and rules for detecting suspicious IE behavior. Processing of LoE is expensive, therefore we foreseen *filtering component*, which will reduce amount of *LoP* provided to *LoE component*. Here we would like to reuse data mining techniques to preselect data.

Conclusions

In this work we presented preliminary work, which target hard problem of IE quality. We propose explain functionality for IE, which assists BI user or IE expert in assessing quality of BI over unstructured text.

References

[1] R. Feldman. Information Extraction: Theory and Practice. *Tutorial at ICML 2006*, 2006

[2] S. Sarawagi. Information Extraction. *Published in Foundation and Trends in Databases,* 2008.

[3] F. Brauer, W. Barczynski, G. Hackenbroich, M. Schramm., A. Mocan. RankIE: Document Retrieval on Tanked Entity Graph. *In proc. VLDB 2009*, 2009

[4] E. Michelakis, R. Krishnamurthy, P. J. Haas, and S. Vaithyanathan. Uncertainty management in rule-based information extraction systems. *In Proceedings of the 35th SIGMOD 2009*, 2009

[5] F. Naumann. Quality-driven query answering for integrated information systems. Springer-Verlagx Inc., 2002.

[6]O. Benjelloun, A. Sarma, A. Halevy, J. Widom. Uldbs: databases with uncertainty and linage. *In proc. of VLDB06*

[7]W. Barczynski, F. Brauer, A. Loeser, A. Mocan. Algebraic IE on Enterprise Data. *In proc. IK&IR 2009*

[8] F. Reiss, S. Raghavan, R. Krishnamurthy, H. Zhu, S. Vaithyanathan. An algebraic approach to rule-based information extraction. *In proceeding of ICDE 2008*, 2008

[9] A. Glass, D. L. McGuiness, P. P. da Silve, M. Wolverton. Trustable task processing systems, *In Kunstliche Intelligenz, Special Issue on Explanation, pages 12–18. Heft*, January 2008.

[10] B. Schueler, S. Sizov, S. Staab, D. T. Tran. Querying for meta knowledge. *In proceedings of 17th WWW*, 2008

[11] M. Ge, M. Helfert. A review of information quality research–develop a research agenda. *In proc. of ICIQ*, 2007

[12] M. Wick, R. Thompson, and W. B. Reconstructive expert system explanation. Artif. Intell., 54(1-2):33–70, 1992

[13] J Woodward. Scientific explanation. *The Stanford Encyclopedia of Philosophy*, February 2008