Tackling Incompleteness in Information Extraction – A Complementarity Approach

Christina Feilmayr

Johannes Kepler University Linz, Altenberger Straße 69, 4040 Linz, Austria
cfeilmayr@faw.jku.at

Abstract. Incomplete templates (attribute-value-pairs) and loss of structural and/or semantic information in information extraction tasks lead to problems in downstream information processing steps. Methods such as emerging data mining techniques that help to overcome this incompleteness by obtaining new, additional information are consequently needed. This research work integrates data mining and information extraction methods into a single complementary approach in order to benefit from their respective advantages and reduce incompleteness in information extraction. In this context, complementarity is the combination of pieces of information from different sources, resulting in (i) reassessment of contextual information and suggestion generation and (ii) better assessment of plausibility to enable more precise value selection, class assignment, and matching. For these purposes, a recommendation model that determines which methods can attack a specific problem is proposed. In conclusion, the improvements in information extraction domain analysis will be evaluated.

Keywords: Information Extraction, Data Mining, Text Mining, Interestingness Measures, Information Integration.

1 Motivation

Imperfections in information extraction (IE) manifest as negative characteristics of the information retrieved, such as *unreliability, ambiguity, uncertainty, inconsistency, incompleteness* and *imprecision*. Fully correct, complete, and precise IE from unstructured text is not feasible with current IE methods. Inaccurate IE yields noisy and incomplete datasets with many missing values. In the case of the semantic web, this means inaccurate statements predominate, resulting primarily in erroneous annotations and ultimately in inaccurate reasoning on the web.

Incomplete data collection in IE tasks, especially in scenario template (ST) production, has serious consequences in subsequent processing. Information and model quality decreases proportionally with the number of values missing from template slots. Incompleteness generally results in indecisiveness, which means that on the one hand value selection, class assignment, and matching and mapping processes in IE domain analysis are profoundly affected by incomplete knowledge sources (ontologies, lexica), missing values in template attributes (originating from upstream IE tasks, e.g., natural language processing), missing descriptive context information, and missing or incomplete constraints (e.g., quality, background, syntactical, lexical or morphological constraints) and

E. Simperl et al. (Eds.): ESWC 2012, LNCS 7295, pp. 808–812, 2012.

conditions (in terms of recurrent occurring collocations/co-occurrences, concept dependencies). On the other hand, downstream processing is affected by erroneous and incomplete template slots. Incompleteness in IE causes inaccurate inference of semantic classes and inaccurate plausibility assessment.

In this context, this thesis focuses on the incompleteness problem in IE and analyzes the applicability of complementarity, and its impact on improving the IE process.

2 Proposed Approach

While tackling the above-mentioned problems primarily necessitates a reassessment of contextual information, it also requires strategies for predicting missing values and generating suggestions for missing slot values. Consequently, methods that obtain new, additional information –such as text and data mining– are needed. Further, a means of exploiting available context information to establish meaningful constraints, conditions, and thresholds for value selection, class assignment, and the matching and mapping procedures is required. Finally, a procedure must be devised to combine the information obtained, evaluate and estimate its reliability, and incorporate the available contextual information.

2.1 Contribution of the Research Work

This research work responds to these requirements by considering the application of the *principle of complementarity* –an approach known from the field of information integration– to IE, examining the impact of redundancy on the probability of correctness. Complementarity is defined as the combination of pieces of information from different sources, taking their respective levels of reliability into account [1]. These pieces of information are the outputs of several data mining methods and text mining tasks. Therefore it enables the alignment of (intermediate) information extraction results with results from data mining methods (or rather complete text mining tasks). Using complementarity allows several data mining and text mining tasks to be integrated. Hence, the main contribution is threefold: **(i)** A **recommendation model**, which supports the user in selecting appropriate text mining tasks and data mining methods (in accordance with the identified incompleteness types). Integrating several information sources (complementarity) supports **(ii)** efficient **reassessment of contextual information** and **suggestion generation** using prediction. Analysis of different sources according to confidence and interestingness yields **(iii) better identification of material with high information potential** and leads thereby to **better plausibility assessment**. In summary, these three elements contribute to better value selection, class assignment, matching and mapping, and consequently to more robust IE results. This makes IE results become more *precise* (in statistical terms), more *certain* (in terms of confidence values) and more reliable (confirmed by human) evaluation.

2.2 The Complementarity Approach

Different types of incompleteness require different approaches to attacking the problems involved. Problems of incompleteness are subdivided into *schema-level-problems* regarding template design and *instance-level-problems* that refer to

incompleteness in the attribute-value-pair level. Consequently, incompleteness in IE is classified into missing (i) attribute values, (ii) structural information (class labels, relationships between and within templates), and (iii) semantic information (conditions, constraints, and contextual information).

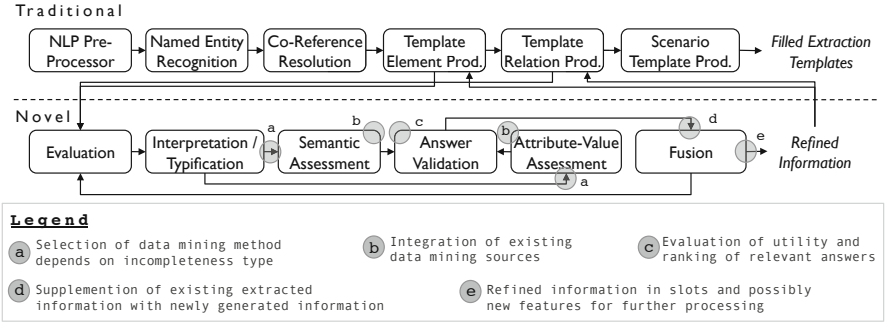

Fig. 1. General Approach of Complementarity

After the type of incompleteness has been identified, the appropriate data mining methods must be selected. As already mentioned, the thrust of the new IE approach is to integrate text and data mining into IE. Possible integration approaches are: (i) identification of collocations and co-occurrences, which can be used to resolve contradictions, perform word sense disambiguation, and validate semantic relations; (ii) constraint-based mining, which can, for example, learn different kinds of constraints (e.g., data type constraints); (iii) identification of frequent item sets (associations), which can also be used as constraints for ST and template merging; (iv) prediction provides additional evidence for missing slot values. (i)-(iii) are approaches, which are applied to assess semantics, and (iv) to assess attribute-values. New information in the form of constraints, conditions, or even suggestions for slots is generated and used to improve IE results. The results are evaluated for a selected set of features (using standard measures of the respective data mining method, e.g., accuracy, and/or selected interestingness measures [4]) that might impact the construction of answers in the presence of incompleteness. These features are combined in a flexible *utility function* (adapted from [7]) that expresses the overall value of information to a user. Determining utility means that first the utility is calculated for each answer, and second, if fusion is performed, the utility of the new value must be calculated in order to determine whether it is more appropriate than the available alternatives. Consequently, the utility value allows us to (i) define a meaningful ranking of candidates for filling incomplete templates and (ii) discover the best fusion.

A possible single point of failure is the automatic determination of the incompleteness type. Thus, the processable incompleteness types must be selected with care. Another possible limitation occurs if measures of interestingness are insufficiently meaningful: This renders the determined utility value useless and leads (depending

on the ranking and filtering methods used) to too much additional information being selected for fusion and even more contradictory, uncertain, and (semantically) imprecise information being produced.

3 Background

In the early PhD phase, the literature review focused mainly on the general idea of integrating data mining into IE. In [3], the author discussed the role of IE in text mining applications and summarized initial research work on the integration of data mining into IE. These first initiatives have been successful, but they discuss relatively simple problems. Most importantly, the projects [6], [8], and [10] demonstrate that the information extracted by such an integrated approach is of high quality (in terms of correctness, completeness, and level of interest).

To the best of the author's knowledge, there is no other research activity ongoing that deals with the integration of data mining into IE to overcome the specific problem of incompleteness. There are some well-established approaches based on complementarity in the knowledge fusion community. A general overview of knowledge fusion is given in [1]. Nikolov [9] outlined a knowledge fusion system that makes decisions depending on the type of problem and the amount of domain information available. Zeng et al. [11] implemented a classifier to acquire context knowledge about data sources and built an aggregation system capable of explaining incomplete data. Ciravegna et al. [2] proposed an approach based on a combination of information extraction, information integration, and machine learning techniques. There, methodologies of information integration are used to corroborate the newly acquired information, for instance, using evidence from multiple different sources. How to exploit redundancy in terms of IE and question answering/answer validation are described in [2] and [5], respectively.

4 Planned Research Work

The methodology comprises the following steps:

Requirements Analysis (completed). First, the requirements of IE domain analysis (regarding the incompleteness problem) had to be identified by means of an in-depth problem analysis that yielded a comprehensive summary of problematic IE issues, their intra- and interdependencies, and their impact on IE accuracy.

Classification of Data Mining Methods (in progress). Based on the requirements and problem analysis, several incompleteness types are identified and the available data mining methods classified accordingly.

Conceptual Design and Method Selection. For each incompleteness type, suitable IE and data mining methods and techniques must be selected. Based on the knowledge gained, a conceptual design for the interface between IE and data mining will be developed and a detailed conceptual design of complementarity created. Features that might impact the construction of answers in the presence of incompleteness must be identified.

Evaluation of Test Scenarios. The research work will conclude with a three-part analysis demonstrating the improvements in IE domain analysis: (i) the first part is a non-optimized information extraction process that provides the baseline; (ii) the second part integrates a gold standard for a specific problem in order to highlight the seriousness of the incompleteness problem; (iii) the third part is an information extraction process using complementarity in order to overcome the incompleteness problem. In comparison to (i) and (ii) the third part of evaluation should highlight the improvements in reducing incompleteness. Moreover, an expert evaluation is planned, which evaluate the several outcomes of complementarity modules.

Acknowledgement. This work is supported by an Austrian research grant (FIT-IT Semantic Systems Dissertation Fellowship Project) from BMVIT (project nr. 829601).

References

1. Bloch, I., Hunter, A., et al.: Fusion: General Concepts and Characteristics. International Journal of Intelligent Systems, Special Issue: Data and Knowledge Fusion 16(10), 1107–1134 (2001)
2. Ciravegna, F., Chapman, S., Dingli, A., Wilks, Y.: Learning to Harvest Information for the Semantic Web. In: Bussler, C.J., Davies, J., Fensel, D., Studer, R. (eds.) ESWS 2004. LNCS, vol. 3053, pp. 312–326. Springer, Heidelberg (2004)
3. Feilmayr, C.: Text Mining Supported Information Extraction - An Extended Methodology for Developing Information Extraction Systems. In: Proceedings of 22nd International Workshop on Database and Expert Systems Applications (DEXA 2011), pp. 217–221 (2011)
4. Geng, L., Hamilton, H.J.: Interestingness Measures for Data Mining: A Survey. ACM Computing Surveys 38(3), Article 9 (2006)
5. Magnini, B., Negri, M., Prevete, R., Tanev, H.: Is It the Right Answer? Exploiting Web Redundancy for Answer Validation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 425–432 (2002)
6. McCallum, A., Jensen, D.: A Note on the Unification of Information Extraction and Data Mining using Conditional-Probability, Relational Models. In: Proceedings of the IJCAI Workshop on Learning Statistical Models from Relational Data (2003)
7. Motro, A., Anokhin, P., Acar, A.C.: Utility-based Resolution of Data Inconsistencies. In: Proceedings of the International Workshop on Information Quality in Information Systems, pp. 35–43 (2004)
8. Nahm, U.Y., Mooney, R.J.: Using Soft-Matching Mined Rules to Improve Information Extraction. In: Proceedings of the AAAI 2004 Workshop on Adaptive Text Extraction and Mining, pp. 27–32 (2004)
9. Nikolov, A.: Fusing Automatically Extracted Annotations for the Semantic Web, PhD Thesis, Knowledge Media Institute, The Open University (2009)
10. Wong, T.-L., Lam, W.: An Unsupervised Method for Joint Information Extraction and Feature Mining Across Different Web Sites. Data & Knowledge Engineering 68(1), 107–125 (2009)
11. Zeng, H., Fikes, R.: Explaining Data Incompleteness in Knowledge Aggregation, Technical Report, Knowledge Systems, AI Laboratory, KSL-05-04 (2005)