Designing a Multi-Dimensional Space for Hybrid Information Extraction

Christina Feilmayr, Klaudija Vojinovic, Birgit Pröll
Johannes Kepler University Linz
Institute of Application Oriented Knowledge Processing (FAW)
Linz, Austria
{cfeilmayr, cvojinovic, bproell}@faw.jku.at

***Abstract*—Information extraction systems are developed for various specific application domains to manage an increasing amount of unstructured data. The majority build either upon the knowledge-based approach, which promises high accuracy but involves labour-intensive coding of extraction rules, or upon the automatically trainable systems approach, which produces highly portable solutions but requires an appropriate learning set. In this paper, we present results of a project that aims to provide a new methodology which combines the knowledge-based and the machine learning approach into a hybrid one in order to compensate for their respective shortcomings and to achieve high IE performance. Firstly, we propose the idea of a multi-dimensional space that guides users in selecting appropriate methods, i.e., different hybrid concepts, depending on the extraction task and the level of available features. Secondly, we provide the concept of one hybrid approach, namely the sequential processing of a knowledge-based approach and a selection of different machine learning methods. Thirdly, we present the evaluation of an implementation of the sequential extraction on a curriculum vitae corpus. Thus, we provide first results for filling the multi-dimensional space for hybrid information extraction.**

Hybrid Information Extraction, (Statistical) Machine Learning, Extraction Methodology

I. INTRODUCTION

Developing a knowledge-based (KB) or an automatically trainable (machine learned, ML) information extraction (IE) system is time- and labor-intensive. The challenge in the iterative engineering process is that extraction rules must be (i) sufficiently generic to extract the full extent of available information and (ii) sufficiently specific to extract relevant information according to a given specification [2]. Consequently, building a scalable IE system manually is not feasible. (Statistical) machine learning is a promising approach to overcoming these problems. ML is becoming increasingly popular for exploring the linguistic richness of corpora and for promoting adaptivity. Hence, the method of choice is currently the machine learning approach, where a generic process builds a classifier by learning rules from a predefined training set. There are two key issues with ML: (i) a sufficiently large amount of training data must be available, and (ii) an appropriate set of features must be chosen when training the system. Careful selection of both the features and the document set from which to extract them are thus essential to the learning procedure. Consequently, the automatically trained systems seem much more appealing because they require less system expertise in the knowledge domain for customization [2]. Hence, the question arises which of the two approaches performs better. In fact, as previously mentioned, either has its advantages and disadvantages.

A possible solution is to combine both approaches, using the advantages of automatically trained systems to counteract the disadvantages of the knowledge-based approach. The ML approach can use the (prior) results of the KB approach as a training corpus to build a classifier.

The overall aim of this research work is to develop methods and processes that enable a more precise IE, and consequently to conceive a new IE methodology that guides a user in selecting appropriate (hybrid) IE methods according to several (natural language) characteristics. The methodology is based on a multi-dimensional space, where the data points in the multi-dimensional space are the results of several ML or hybrid IE methods enriched with meta data, such as the features and parameters used.

This paper provides three main contributions: Firstly, the idea of a multi-dimensional space that guides users in selecting appropriate methods is proposed. Secondly, the concept of one hybrid approach, namely the sequential processing of KB approach and a selection of different ML methods is provided. Thirdly, an implementation of the sequential extraction is evaluated using a curriculum vitae corpus. In order to validate the methodology, first a corpus of curriculum vitae in the field of eRecruitment, and second the Reuters Corpus[1] were used. In addition, our industrial partner Join Vision GmbH[2] is about to implement a test framework for the designed scenarios and their respective hybrid methods.

The remainder of this paper is structured as follows: Section 2 introduces the methodology of the hybrid IE, the multi-dimensional space, and the first application scenario (the sequential combination of KB and ML IE). Section 3 presents the applied evaluation approach and the research methodology. Section 4 discusses the preliminary experimental results and a first version of the multi-dimensional space. Section 5 summarizes related work and initiatives to design hybrid IE methods.

[1] http://about.reuters.com/researchandstandards/corpus/index.asp, last visited: May 8, 2012

[2] http://www.joinvision.com/, last visited: May 8, 2012

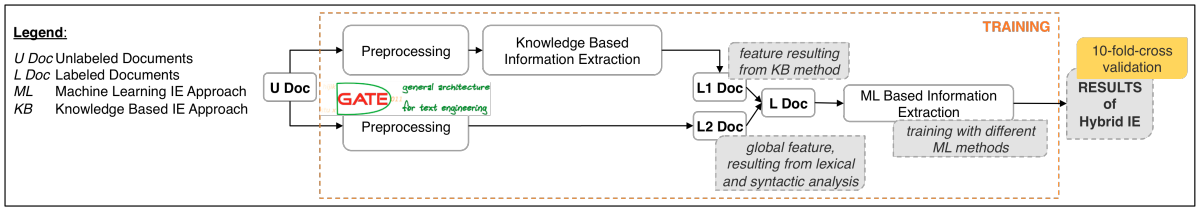

Figure 1. Concept of sequential extraction

Finally, Section 6 concludes the paper with various aspects of future work.

II. State-of-the-Art Hybrid Information Extraction

In the literature the sequential processing has been termed and discussed as *hybrid information extraction* [7][11] or *multi-strategy approach* [3][12][16]. The combination of various ML methods is also named hybrid IE [11][13][15].

The most promising approaches to designing a hybrid method are statistical IE methods. Fresko et al. [8] described a framework termed MERGE (Maximum Entropy Rule Guided Extraction), which is a hybrid named entity recognition system that combines ML techniques. To obtain a TEG model, Feldman et al. [7] trained three different classes of trainable parameters: the probabilities of rules of nonterminals, the probabilities of different expansions of n-grams, and the probabilities of terms in word classes. Finally, the algorithm interpolates between the three models. The idea of selecting different learners for different learning task is one possible approach to designing a new hybrid IE method. De-Sitter and Daelemans [6] presented an approach consisting of two classification-based ML loops. In [13], a hybrid ML method for IE on semi-structured texts that combines conventional text classification techniques and HMM was proposed. Larkey and Croft [10] applied – individually and in combination – kNN, relevance feedback, and Bayesian independence classifiers. Zhang [17] proposed different ML strategies and introduced a random-subspace-based algorithm. Zhang recommended using supervised and weakly supervised methods (e.g., active learning and bootstrapping) in an integrated method.

Further conceptions depend on the final evaluation and, consequently, on the conclusions drawn. Nevertheless, all of the approaches mentioned are good starting points for designing a new hybrid method that builds upon KB methods.

III. Methodology of our Hybrid Information Extraction Approach

The research approach is subdivided into three phases. The *first phase* is the conception phase that focuses primarily on:

Design of hybrid concepts. The two approaches, the KB and the ML one, can be interfaced in different ways that result in different IE scenarios. The following scenarios are considered: (i) sequential extraction that uses the extraction results of the KB approach in the ML part, (ii) automatic rule base extension that takes the KB results into account to learn new rules or update existing ones and results in a more accurate rule base, (iii) knowledge base extension, which aims to extend the existing KB (in the form of gazetteer lists, thesauri, ontologies, etc.).

Definition of the multi-dimensional space, which results from the preprocessed tasks.

The *second phase* is the implementation phase of the test framework. The test system implements the hybrid concepts for a concrete test scenario. In this case, curriculum vitae (e.g., for a job application) are used.

In the *third phase*, the results of the preliminary filled multi-dimensional space (phase 1) and the results of the test system (phase 2) are evaluated and verified.

A. Idea of the Multi-Dimensional Space

The proposed multi-dimensional space has three axes: the first indicates the information extraction task [5] (*Named Entity (NE)* recognition, *Template Element (TE)* construction, *Template Relation (TR)* construction, *Scenario Template (ST)* production), the second indicates the hybrid concept (*Sequential Extraction (SE)*, *Rule Base (RB) Extension*, *Knowledge Base (KB) Extension*), and the third indicates the granularity of the used features that emerge from the preprocessing knowledge-based approach. This multi-dimensional space can be extended to further IE aspects, such as text structure, number of available training and test examples, and available context/structure information. The proposed data space results in a set of quintuples [`h, fl, t, m, x`] that represent the data points. Here, `h` denotes the selected hybrid concept, `fl` the feature level (described in detail in the next section), `t` the IE task, `m` the ML method (e.g., SVM, k-NN, CRF), and `x` is the f-measure resulting from k-fold-cross-validation. Examples of such quintuples are:

```
[Sequential Extraction, Level2, TE, SVM, 0.87]
[Sequential Extraction, Level2, TE, k-NN, 0.64]
[Sequential Extraction, Level2, TE, CRF, 0.91]
```

The overall motivation for designing such a multi-dimensional space is to support IE system designers in selecting a method (ML method(s) and hybrid concept) that is appropriate for the IE task. Conversely, also an appropriate IE task can be selected for a given method.

B. Concepts of Hybrid Information Extraction

The focus of this research paper is on the first application scenario, that is, the sequential processing of the KB and the ML approach, hereafter referred to as sequential extraction.

Sequential extraction, shown in Figure 1, is based on the IE component(s) of the KB approach. On the one hand, it uses the results provided by the manually encoded rules, and on the other hand it combines them with lexical and syntactic features (*global features*) that result from an additional pre-

processing phase. The results of the preprocessing phase and of the rule-based IE component are used as the feature set, and accordingly the annotated documents are the training set for several learners. The main goal of this approach is to reduce the test-and-debug cycles when implementing an IE system. The accuracy of the ML part increases with the amount of training data. Consequently, the impact of the ML part should be directly proportional to the increase in the training set.

IV. Evaluation Approach

The domain we selected for preparing the multi-dimensional space is the extraction of personal data from curriculum vitae (e.g., name, address, and job titles). The main reasons for our choice were our industrial project partner, the availability of good data sources, and that it poses a range of problems that must be addressed. A corpus of 180 curriculum vitae was collected; most of them are in German, some are in English. The documents differ in structure, length, and detail. The IE tasks range from extracting structural types that are used for recognizing named entities to learning relations between named entities and/or templates. This paper covers the following IE tasks:

- Learning structural types (section indicator)
- Learning named entities (name, job title)
- Learning template elements (address)

A. Preprocessing and Annotation

The data was preprocessed using (i) a rule-based IE system provided by our industrial partner and (ii) the GATE[3] system [4], which enables tokenization, orthography, part-of-speech tagging, and chunk recognition text features. Furthermore, GATE provides the lexical and syntactic features that are used for ML. The features selected are a fairly basic set in terms of linguistic/syntactic processing. A manually coded rule base that uses a comprehensive domain taxonomy handles the annotation process.

There is a data imbalance within different classes (positive and negative examples in the classes) and between the classes (different total number of examples in each class) in the training and test corpus. The fact that there are significantly fewer training instances of one class compared to another poses a great problem to ML. The misclassifications (primarily *false negatives* that decrease the recall), resulting from ML with imbalanced data sets must thus be considered in the evaluation phase. Furthermore, when trying to learn a set of classification rules for all classes, it may happen that the smaller classes are largely ignored. One solution to this problem is to learn the classification rules that predict only one small class (recognition-based learning). In the literature, there are several examples that report success [14].

The last preprocessing task is to produce the feature files that can be used independently from the machine learning API. Three different machine learning APIs are used:

- **GATE** framework and its *Batch Learner* processing resource.
- **MALLET API**[4], which includes many tools for sequence tagging (e.g., for named entity recognition in text).
- **RapidMiner API**[5], which also includes many algorithms for classification, clustering and rule induction, mining algorithms.

B. Feature Level

Learning at three different feature levels should reveal the influence that using features of annotations resulting from a preprocessed KB approach has on the ML results. It must be pointed out that the higher the feature level (and consequently the higher the influence of the rule-based part) is, the more likely are incorrect training examples (if there is no human annotator who corrects the results of the knowledge-based part).

Feature Level 1: Lexical and syntactic features (global features). Given all the available training documents, the textual patterns necessary to extract the annotated information are automatically learned. The features used are structured as lexical, shallow syntactical and deep syntactical (chunk tags) features, which follows closely [17].

Feature Level 2: Global features extended to features of the knowledge-based approach. The features available from the global feature set are extended to extraction results of the KB approach and are used as contextual data of the information sought after. The annotations and their features are classified as

- structural features, derived from annotations, e.g., paragraphs, phrases, different section indicators, font information, and dated and undated blocks.
- semantic discourse features that refer to features whose values are computed by using smaller text fragments. Semantic discourse features, such as duration, date, and location information are derived from annotations.
- semantic features that refer to features that result from information extraction results, e.g., birthday, nationality, email, skills, phone number.

Feature Level 3: Enriched feature space with features resulting from final annotations. As a result of the KB extraction, the final annotations (e.g., name, address, job title) come with the following features that extend the feature space resulting from levels 1 and 2:

- class, unit, and type (similar to semantic and semantic discourse feature).
- annotation-specific Boolean-valued features, e.g., *partOfEmail, personalInfoSection, firstWordInParagraph, OnlyWordInParagraph, language*.

V. Preliminary Experimental Results

Performance is reported using the standard IE measures precision, recall, and f1-measure.

Table 1 shows the four different IE tasks and their highest precision, recall, and f1-measures based on 10-fold cross-validation at the three different feature levels. Different ML

[3] http://www.gate.ac.uk/, last visited: May 8, 2012

[4] http://mallet.cs.umass.edu/, last visited: May 8, 2012

[5] http://rapid-i.com/, last visited: May 8, 2012

TABLE I. PERFORMANCE AT LEVELS 1 TO 3 ON DIFFERENT IE TASKS

IE TASK		PAUM			SVM			kNN			CRF		
		P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
SECTION IND-ICATOR	Results of KB	Precision (P) = 0.91 Recall (R) = 0.86 F1-measure (F) = 0.84											
	Level 1	0.81	0.65	0.72	0.76	0.70	0.72	0.32	0.27	0.29	**0.78**	**0.62**	**0.69**
	Level 2	0.91	0.90	0.91	0.93	0.87	0.90	0.75	0.43	0.54	**0.99**	**0.99**	**0.99**
	Level 3	0.99	0.92	0.95	0.99	0.91	0.95	0.74	0.36	0.48	**1**	**0.99**	**0.99**
PERSONS' NAME	Results of KB	Precision (P) = 0.86 Recall (R) = 0.82 F1-measure (F) = 0.84											
	Level 1	0.55	0.59	0.57	0.56	0.59	0.57	0.39	0.53	0.44	0.68	0.71	0.68
	Level 2	0.94	0.78	0.85	0.96	0.80	0.87	0.82	0.64	0.71	**0.98**	**1**	**0.99**
	Level 3	0.98	0.80	0.88	100	0.81	0.89	0.98	0.82	0.89	**1**	**1**	**1**
JOB TITLE	Results of KB	Precision (P) = 0.93 Recall (R) = 0.94 F1-measure (F) = 0.93											
	Level 1	0.52	0.42	0.46	0.58	0.43	0.49	0.24	0.14	0.17	**0.64**	**0.66**	**0.65**
	Level 2	0.56	0.44	0.49	0.56	0.46	0.50	0.17	0.08	0.11	**0.69**	**0.69**	**0.69**
	Level 3	0.86	0.84	0.85	0.84	0.80	0.82	0.74	0.44	0.55	**0.99**	**1**	**0.99**
ADDRESS	Results of KB	Precision (P) = 0.86 Recall (R) = 0.75 F1-measure (F) = 0.79											
	Level 1	0.57	0.50	0.51	0.50	0.51	0.50	0.54	0.43	0.47	**0.64**	**0.59**	**0.61**
	Level 2	0.72	0.68	0.69	0.72	0.70	0.70	0.57	0.48	0.50	**0.76**	**0.82**	**0.79**
	Level 3	0.95	0.95	0.95	0.98	0.98	0.98	0.67	0.54	0.59	**1**	**0.99**	**0.99**

techniques were used for testing: Perceptron with Uneven Margins (PAUM), Support Vector Machine (SVM), k-Nearest-Neighbor (kNN), Conditional Random Fields (CRF). Tests with the methods C4.5, NaïveBayes, and Hidden Markov Models (HMM) were not investigated further because these methods performed badly with several feature sets. Especially Naïve Bayes and HMM suffer from too many *false positives,* resulting in a high recall. C4.5 treated the minority class as noise and therefore to disregard it. Therefore, these methods are not listed in the table below.

The methods were tested with various parameters; the best settings for PAUM were the default ones (`p=50, n=5, optB=0.0`, p denoting the positive margin, n the negative margin, and optB the modification of the bias term), for a linear SVM the selected parameters were `c=0.7` and `τ=0.4` (c determines the cost associated with allowing training errors, τ sets the value of the uneven margins), kNN was used with `k=1`, and CRF was trained with different trainers and parameters. MALLET provides different trainers for CRF that enable different training and test settings. The CRF models were trained in multiple settings that differ in the optimization of the parameters, the smoothing of the training data, and the topology (e.g., using L-BFGS algorithm for optimizing the parameters, applying the conjugate gradient method).

As can be seen in Table 1, CRF using different feature levels performs best for all IE tasks (by ~5-10% compared to other ML-techniques). Using CRF with L-BFGS results in the most robust results. Using CRF has a positive side effect because it determines the boundaries of the extracted chunks more precisely than the extraction rules. This is why the lenient performance measures were used to evaluate the documents. PAUM performs worst; kNN and SVM provide acceptable results, but especially the kNN results have a tendency to favor precision over recall. This is due to the imbalanced training data, which leads to many *false negatives*. One of the most outstanding results is that the sequential processing of KB and ML (especially when using features of levels 2 and 3) performs better than KB alone. But since CRF is the best method for all IE tasks and performs best with different feature levels, a dependency on the three aspects defined in section 1 – IE task, granularity level of KB features, and ML method – could not be proven. This in turn means that the decision criteria defined for selecting an appropriate ML and/or hybrid method must be extended or rather reworked[6].

The 10-fold cross-validation experiments can be summarized as follows: (i) using a hybrid IE method in a sequential manner results in more consistent performance, especially when the influence of the KB approach is strong (feature levels 2 and 3), (ii) all methods except CRF suffer from a higher precision (resulting in too many *false negatives*) than recall. The higher the level, the more balanced are recall and precision.

In the evaluation phase, different context window sizes for features were tested. These tests also reveal some regularities relating to recall and precision. For example, using structural features in a unigram resulted in better performance than using them in a context window. In contrast, semantic and semantic discourse features produced better results when used in a context window that ranges from 3 to 7. In general, all features of level 2 that use a token window size ranging from 3 to 9 (structural, semantic, and semantic discourse features) had a positive influence on recall. Building unigrams with these features optimizes precision. Features that provide information about distances between annotations (e.g., number of tokens between the street term and the house number) or their ordering (e.g., name, *first name-last name* vs. *last name-first name*) positively influenced the precision.

[6] This fact is also reflected in the second domain (using the Reuters corpus).

A consolidated view indicates that the first evaluation phase of the hybrid method has turned out satisfactory. In addition to the positive results, the evaluation phase also shows that unbalanced data (class distributions) has an undesirably strong effect on the performance. It leads either to a trivial classifier that completely ignores the minority class or to overfitting to the training examples. Furthermore, the small training corpus makes it difficult to learn a robust classification model. A further aspect that should be considered in future research is the available amount of context information (e.g., in contrast to the news documents of the Reuters corpus, a person's name in a curriculum vitae is at the top of the document and has sparse context information).

VI. Conclusion & Future Work

The work reported in this paper aims to provide a methodology for combining the existing IE approaches into a hybrid one. The first part of the research work concentrated on one application scenario, the sequential processing of the KB and the ML approach, which uses the results of the KB IE components to subsequently train a classifier. Preliminary evaluation results demonstrate that, compared to the results of the knowledge-based-only approach, the hybrid IE method makes possible a considerable increase in performance.

The knowledge gained from this first evaluation phase can be summarized to the following facts:

Rework of multi-dimensional space. The preliminarily defined decision criteria for selecting an appropriate ML and/or hybrid IE method must be extended because a dependency on the IE task, the granularity level of KB features and the ML method could not be proven. Currently, the multi-dimensional space tends to obtain a more powerful decision tree that considers more factors, which influences the selection of an appropriate hybrid method (e.g., degree of imbalance, one-class-learning vs. multi-class-learning, object-classification task, like token or sequence labeling, corpus size, document structure, comprehensiveness of context information).

Consideration of unbalanced training and test sets. There are various methods for solving the class imbalance problem, including resizing the training data sets (over- and undersampling the majority class [9]), adjusting misclassification costs, and recognition-based learning (learning from the minority class). For our purposes, over- and undersampling are of interest.

Number of training documents. To obtain more positive examples, semi-supervised machine learning methods can be used [1].

These aspects provide the basis for further improving our proposed hybrid IE method.

Acknowledgment

This work is supported by a grant of the Austrian Research Promotion Agency *FFG* (*BRIDGE* 678422).

References

[1] S. Abney, "Semisupervised Learning for Computational Linguistics", 1. Computational linguistics, Study and teaching (Higher) I. Title. II. Series, Chapman & Hall/CRC, 2007.

[2] D.E. Appelt and D.J. Israel, "Introduction to Information Extraction", AI Communications, Vol.12, No.3, pp.161-172, IOS Press Amsterdam, The Netherlands, 1999.

[3] F. Ciravegna, A. Dingli, J. Iria, and Y. Wilks, "Multi-Strategy Definition of Annotation Services in Melita", Human Language Technologies Workshop at Second International Semantic Web Conference, Sanibel Island, Florida, USA, 2003.

[4] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications". Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, 2002.

[5] H. Cunningham, "Automatic Information Extraction". In: Encyclopedia of Language & Linguistics, ed: K. Brown, Vol. 5, pp. 665-677, Elsevier, 2006.

[6] A. De Sitter, and W. Daelemans, "Information Extraction via Double Classification", In Proceedings of the ECML/PKDD 2003 Workshop on Adaptive Text Extraction and Mining (ATEM 2003), pp. 66-73, Cavtat-Dubronik, Croatia, 2003.

[7] R. Feldman, B. Rozenfeld, and M. Fresko, "TEG: a Hybrid Approach to Information Extraction". Journal of Knowledge and Information Systems Vol.9 (1), pp.1-18, Springer-Verlag New York, USA, 2006.

[8] M. Fresko, B. Rozenfeld, and R. Feldman, "A Hybrid Approach to NER by Integrating Manual Rules into MEMM", AI and Math 2006, Ft. Lauderdale, Florida, USA, 2006.

[9] H. He, E.A. Garcia, "Learning from Imbalanced Data", IEEE Transactions on Knowledge and Data Engineering, Vol.21, Nr.9, pp.1263-1284, IEEE Computer Society, 2009.

[10] L.S. Larkey, and W.B. Croft, "Combining Classifiers in Text Categorization", In Proceeding of SIGIR '96 Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.289-297, ACM New York, USA, 1996.

[11] G. Neumann, "A Hybrid Machine Learning Approach for Information Extraction from Free Texts". In: M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, W. Gaul (eds.): From Data and Information Analysis to Knowledge Engineering, Studies in Classification, Data Analysis, and Knowledge Organization, pp.390-397, Springer Berlin/Heidelberg 2006.

[12] H. Reeve, and H. Han, "Survey of Semantic Annotation Systems", Symposium on Applied Computing, Proceedings of the 2005 ACM Symposium on Applied computing, pp. 1634-1638, Santa Fe, New Mexico, 2005.

[13] E.F.A. Silva, F.A. Barros, and R.B.C. Prudêncio, "A Hybrid Machine Learning Approach for Information Extraction". In Proceedings of the 6th International Conference on Hybrid Intelligent Systems, Auckland, New Zealand, 2006.

[14] G.M. Weiss, "Mining with Rarity: A Unifying Framework", In ACM SIGKDD Explorations Newsletter - Special Issue on Learning from Imbalanced Datasets, Vol.6, Issue 1, pp.7-19, ACM New York, USA, 2004.

[15] J. Xiao, D. Zhu, and L. Zou, "A Hybrid Approach for Web Information Extraction", In Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, pp.1560-1563, Kunming, China, 2008.

[16] P. Yuan, G. Wang, and Q. Zhang, H. Jin, "SASL: A Semantic Annotation System for Literature", In Web Information Systems and Mining, Lecture Notes in Computer Science, pp.. 158-166, Springer Berlin/Heidelberg, 2009.

[17] Z. Zhang, "Mining Relational Data from Text: From Strictly Supervised to Weakly Supervised Learning", In Information Systems Volume 33, Issue 3, pp. 300-314, Elsevier Science, UK, 2008.