IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITP4263

Information Extraction from Free-Text Business Documents

Witold Abramowicz
Department of Computer Science, The Poznan University of Economics, Poland, W.Abramowicz@kie.ae.poznan.pl

Jakub Piskorski
German Research Center for Artificial Intelligence (DFKI), Germany, piskorsk@dfki.de

ABSTRACT

One of the most difficult aspects of using search technology is the process of getting information in shape for searching. The objective of this paper is an investigation of the applicability of information extraction techniques in real-world business applications dealing with textual data since business relevant data is mainly transmitted through free-text documents. Further, we demonstrate an enormous indexing potential of lightweight linguistic text processing techniques applied in information extraction systems in other closely related fields of information technology which concern processing vast amounts of textual data.

INTRODUCTION

Nowadays, knowledge relevant to business of any kind is mainly transmitted through free-text documents: World-Wide Web, newswire feeds, corporate reports, government document, litigation records etc. One of the most difficult issues concerning applying search technology for retrieving relevant information from textual data collections is the process of converting such data into a shape for searching. IR systems using conventional indexing techniques applied even to homogenous collection of text documents fall far from obtaining optimal recall and precision simultaneously. Since structured data is obviously easier to search, an ever-growing need for effective and intelligent techniques for analyzing free-text documents and building expressive representation of their content in form of structured data can be observed.

Recent trends in information technology such as Information Extraction (IE) provide dramatic improvements in conversion of the overflow of raw textual information into valuable and structured data which could be further used as input for data mining engines for discovering more complex patterns in textual data collections. The task of IE is to identify predefined set of concepts in a specific domains and ignoring other irrelevant information, where domain consists of a corpus of texts together with a clearly specified information need. Due to the specific phenomena and complexity of natural language this is a non-trivial task. However, recent advances in Natural Language Processing (NLP) concerning new robust, efficient, high coverage shallow processing techniques for analyzing free-texts contributed to deploying IE techniques in business information systems.

In this paper we investigate the usability of IE techniques in real-world business applications dealing with vast amount of textual data and demonstrate their enormous potential for general indexing purposes. The rest of this paper is organized as follows. In Section 2 we give an overview of the IE task and designing IE systems. The existing IE systems applied in the financial, insurance and legal domain are presented in section 3. Section 4 demonstrates an enormous potential of shallow text processing – core IE technology to areas strictly related to information extraction. Finally, Section 5 provides some conclusions.

INFORMATION EXTRACTION

Information Extraction Task

The task of *Information extraction* (IE) is identification of instances of a particular pre-specified class of events or relationships and entities in natural language texts, and the extraction of the relevant arguments of the events or relationships [SAIC, 98]. The information to be extracted is pre-specified in user-defined structures called templates (e.g., company information, meetings of important peoples), each consisting of a number of slots, which must be instantiated by an IE system as it processes the text. The slots are usually filled with: some strings from the text, one of a number of pre-defined values or a reference to other already generated template. One way of thinking about an IE system is in terms of database construction since an IE system creates a structured representation of selected information drawn from the analyzed text.

In recent years IE technology has progressed quite rapidly, from small-scale systems applicable within very limited domains to useful systems which can perform information extraction from a very broad range of texts. IE technology is now coming to the market and is of great significance to finance companies, banks, publishers and governments. For instance, financial organization want to know facts about foundations of international joint-ventures happening in a given time span. The process of extracting such information involves locating names of companies and finding linguistic relations between them and other relevant entities (e.g., locations and temporal expressions). However, in this particular scenario an IE system requires some specific domain knowledge (understanding the fact that ventures generally involve at least two partners and result in the formation of a new company) in order to merge partial information into an adequate template structure. Generally, IE systems rely always to some degree on domain knowledge.

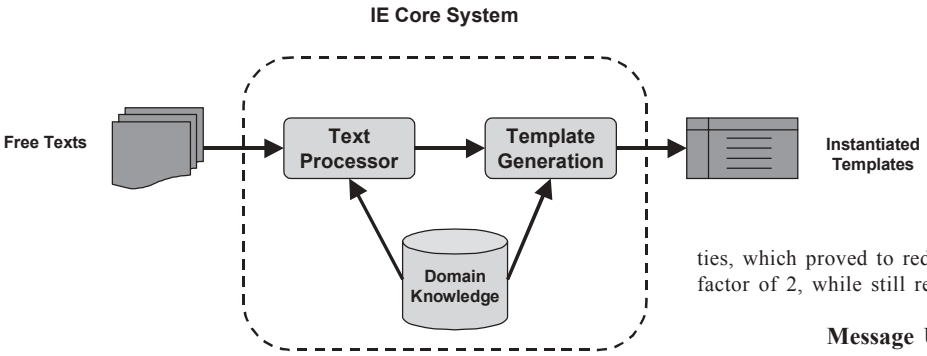
Designing IE Systems

There are two basic approaches to designing IE systems: Knowledge Engineering Approach and Learning Approach [Appelt and Israel, 99]. In the knowledge engineering approach the development of rules for marking and extracting sought-after information is done by a human expert through inspection of the test corpus and his or her own intuition. In the learning approach the rules are learned from an annotated corpora and interaction with the user. Generally, higher performance can be achieved by handcrafted systems, particularly when training data is sparse. However, in particular scenario automatically trained components of IE system might show better performance than their handcrafted counterparts. Approaches to building hybrid systems based on both approaches are currently investigated. IE systems built for different tasks often differ from each other in many ways. Nevertheless, there are core components shared by nearly every IE system, disregarding the underlying design approach.

The coarse-grained architecture of a typical IE system is presented in figure 1. It consists of two main components: text processor and template generation module. The task of the text processor is performing general linguistic analysis in order to extract as much linguistic structure as possible. Due to the problem of ambiguity pervading all levels of natural language processing, this is a non-trivial task. Instead of computing all possible interpretations and grammati-

This paper appears in *Issues and Trends of Information Technology Management in Contemporary Organizations,* the proceedings of the Information Resources Management Association International Conference.

Figure 1: A coarse-grained architecture of an information extraction system



cal relations in natural language text (so called deep text processing - DTP), there is an increased tendency towards applying only partial analysis, co called *shallow text processing*[1] (STP) [Piskorski and Skut, 00] which is considerably less time-consuming[2] and could be seen as a trade-off between pattern matching and fully-fledged linguistic analysis. In shallow text analysis language regularities which cause problems are not handled and instead of computing all possible readings only underspecified structures are computed. The use of STP instead of DTP may be advantageous since it might be sufficient for the extraction and assembly of the relevant information and it requires less knowledge engineering, which means faster development cycle and less development expenses.

The scope of information computed by the text processor may vary depending on the requirements of particular application. Usually, linguistic analysis performed by text processor of an IE system includes following steps:

- Segmentation of text into a sequence of sentences, each of which is a sequence of lexical items representing words together with their lexical attributes
- Recognition of small scale-structures (e.g., abbreviations, core nominal phrases, verb clusters and named entities)
- Parsing, which takes as input a sequence of lexical items and small-scale structures and computes the structure of the sentence, so called parse tree

Depending on the application scenario it might be desirable for the text processor to perform additional tasks, such as: part-of-speech disambiguation, word sense tagging, anaphora resolution or semantic interpretation (e.g., translating parse tree or parse fragments into a semantic structure or logical form). A benefit of the IE-task orientation is that it helps to focus on linguistic phenomena that are most prevalent to particular domain or particular extraction task.

The template generation module merges the linguistic structures computed by the text processor and using domain knowledge (e.g., domain-specific extraction patterns and inference rules) derives domain-specific relations in form of instantiated templates. In practice, the boundary between text processor and template generation component may be blurred.

The input and output of an IE system can be defined precisely, which facilitates the evaluation of different systems and approaches. For the evaluation of IE systems the precision, recall and f-measures were adopted from the IR research community[3].

Information Extraction vs. Information Retrieval

IE systems are obviously more difficult and knowledge intensive to build and they are more computationally intensive than IR systems. Generally, IE systems achieve higher precision than IR systems. However, IE and IR techniques can be seen as complementary and can potentially be combined in various ways. For instance, IR could be embedded within IE for pre-processing a huge document collection to a manageable subset to which IE techniques could be applied. On the other side, IE can be used as subcomponent of an IR system to identify terms for intelligent document indexing (e.g., conceptual indicies). Such combinations clearly represent significant improvement in retrieval of accurate and prompt business information. For example, [Mihalcea and Moldovan, 01] introduced an approach for document indexing using named entities, which proved to reduce the number of retrieved documents by a factor of 2, while still retrieving relevant documents.

Message Understanding Conferences

The rapid development of the field of IE has been essentially influenced by the Message Understanding Conferences (MUC). These conferences were conducted under the auspices of several United States government agencies with the intention to coordinate multiple research groups and government agencies seeking to improve IE and IR technologies [Grishman and Sundheim, 96]. The MUC conferences defined several generic types of IE tasks. They were intended to be prototypes of IE tasks that arise in real-world applications and they illustrate the main functional capabilities of current IE systems. The IE tasks defined in MUC competitions [4] focused on extracting information from newswire articles (e.g., concerning terrorist events, international joint venture foundations and management succession). The generic IE tasks for MUC-7 (1998) were defined as follows:

- Named Entity Recognition (NE) requires the identification and classification of named entities such as organizations, persons, locations, product names and temporal expressions
- Template Element Task (TE) requires the filling of small-scale templates for specified classes of entities in the texts, such as organizations, persons, certain artifacts with slots such as name variants, title, description as supplied in the text.
- Template Relation Task (TR) requires filling a two-slot template representing a binary relation with pointers to template elements standing in the relation, which were previously identified in the TE task (e.g., an employee relation between a person and a company).
- Co-reference Resolution (CO) requires the identification of expressions in the text that refer to the same object, set or activity (eg., variant forms of name expressions, definite noun phrases and their antecedents).
- Scenario Template (ST) requires filling a template structure with extracted information involving several relations or events of interest, for instance, identification of partners, products, profits and capitalization of joint ventures.

State-of-the-art results for IE tasks for English reported in MUC-7 are presented in figure 2.

IE SYSTEMS IN THE BUSINESS DOMAIN

Early IE Systems

The earliest IE systems were deployed as commercial product already in the late eighties. One of the first attempts to apply IE in the financial field using templates was the ATRANS system [Lytinen and Gershman, 86], based on simple language processing techniques and script-frames approach for extracting information from telex messages regarding money transfers between banks. JASPER [Andersen et. al., 92] is an IE system that extracts information from reports on

Figure 2: State-of-the-art results reported in MUC-7

MEASURE\TASK	NE	CO	RE	TR	ST
RECALL	92	56	86	67	42
PRECISION	95	69	87	86	65

corporate earnings from small sentences fragments using robust NLP methods. SCISOR [Jacobs et. al., 90] is an integrated system incorporating IE for extraction of facts related to the company and financial information. These early IE systems had a major shortcoming, namely they were not easily adaptable to new scenarios. On the other side, they demonstrated that relatively simple NLP techniques are sufficient for solving IE tasks narrow in scope and utility.

LOLITA

The LOLITA System [Costantino et. al., 97], developed at the University of Durham, was the first general purpose IE system with fine-grained classification of predefined templates relevant to the financial domain. Further, it provides a user-friendly interface for defining new templates. LOLITA is based on deep natural language understanding and uses semantic networks. Different applications were built around its core. Among others, LOLITA was used for extracting information from financial news articles which represent an extremely wide domain, including different kind of news (e.g., financial, economical, political, etc.). The templates have been defined according to the "financial activities" approach and can be used by the financial operators to support their decision making process and to analyze the effect of news on price behavior. A financial activity is one potentially able to influence the decisions of the players in the market (brokers, investors, analysts etc.). The system uses three main groups of templates for financial activities: company related activities – related to the life of the company, company restructuring activities - related to changes in the productive structure of companies and general macroeconomics activities, including general macroeconomics news that can affect the prices of the shares quoted in the stock exchange.

In the "takeover template" task, as defined in MUC-6, the system achieved precision of 63% and recall of 43%. However, since the system is based on DTP techniques, the performance in terms of speed can be, in particular situations, penalized in comparison to systems based on STP methods. The output of LOLITA was fed to the financial expert system [Costantino, 99] to process an incoming stream of news from on-line news providers, companies and other structured numerical market data to produce investment suggestions.

MITA

IE technology has been recently successfully used in the insurance domain. MITA (Metalife's Intelligent Text Analyzer) was developed in order to improve the insurance underwriting process [Glasgow et. al., 98]. The Metalife's life insurance applications contain free-form textual fields (an average of 2.3 textual fields per application) like for instance: physician reason field - describing a reason a proposed insured last visited a personal physician, family history field – describing insured's family medical history and major treatments and exams field which describes any major medical event within the last five years. In order to identify any concepts from such textual fields that might have underwriting significance, the system applies STP techniques and returns a categorization of these concepts for risk assessment by subsequent domain-specific analyzers.

The MITA system has been tested in production environment and 89% of the information in the textual field was successfully analyzed. Further, a blind testing was undertaken to determine whether the output of MITA is sufficient to make underwriting decisions equivalent to those produced by an underwriter with access to full text. Results showed that only up to 7% of the extractions resulted in different underwriting conclusions.

History Assistant

[Jackson et. al., 98] presents History Assistant - an information extraction and retrieval system for the juridical domain. It extracts rulings from electronically imported court opinions and retrieves relevant prior cases and cases affected from a citator database, and links them to the current case. The role of a citator database enriched with such linking information is to track historical relations among cases. On-line citators are of great interest to the legal profession because they provide a way of testing whether a case is still good law. History Assistant is based on DTP (e.g., uses context-free grammars for computing all possible parses of the sentence). In the prior case retrieval task it achieved a recall of 93.3%.

Trends

The most recent approaches to IE concentrated on constructing general purpose, highly modular, robust, efficient and domain adaptive IE systems. FASTUS [Hobbs et. al., 97] is a very fast and robust general purpose IE system which deploys lightweight linguistic techniques. It was built in the Artificial Intelligence Center of SRI International. It is conceptually very simple, since it works essentially as a set of cascaded nondeterministic finite-state transducers. It was one of the best scoring systems in the MUC Conferences and was used by commercial client for discovering ontology underlying complex Congressional bills, for ensuring the consistency of laws with the regulations that implement them.

[Humphreys et. al., 98] describe LaSIE-II, a highly flexible and modular IE system, which was an attempt to find a pragmatic middle way in the shallow vs. deep analysis debate which characterized the last several MUCs. The result is an eclectic mixture of techniques ranging from finite-state recognition of domain-specific lexical patterns to using restricted context-free grammars for partial parsing. Its highly modular architecture enabled one to take deeper insight into the strengths and weaknesses of the particular subcomponents and their interaction.

Similarly to LaSIE-II, the two top requirements on the design of the IE2 system [Aone et. al., 99], developed at SRA International Inc., were modularity and flexibility. SGML was used to spell out system interface requirements between the sub-modules, which allow an easy replacement of any sub-module in the workflow. The IE2 system achieved the highest score in TE task (recall: 86%, precision 87%), TR task (recall: 67%, precision: 86%) and ST task (recall: 42%, precision: 65%) in MUC-7 competition. REES presented in [Aone and Santacruz, 00] was the first attempt to constructing large-scale event and relation extraction system based on STP methods. It can extract more than 100 types of relations and events related to the area of business, finance and politics, which represents much wider coverage than is typical of IE systems. For 26 types of events related to finance it achieved an F-measure of 70%.

BEYOND INFORMATION EXTRACTION

The last decade has witnessed great advances and interest in the area of information extraction using simple shallow processing methods. In the very recent period, new trends in information processing from texts based on lightweight linguistic analysis closely related to IE have emerged.

Textual Question Answering

Textual Question Answering (Q/A) aims at identifying the answer of a question in large collections of on-line documents, where the questions are formulated in natural language and the answers are presented in form of highlighted pieces of text containing the desired information. The current Q/A approaches integrate existing IE and IR technologies. Knowledge extracted from documents may be modeled as a set of entities extracted from the text and relations between them and further used for concept-oriented indexing. [Srihari and Li, 99] presented Textract - a Q/A system, based on relatively simple IE techniques using NLP methods. This system extracts open-ended domain independent general-event templates expressing the information like WHO did WHAT (to WHOM) WHEN and WHERE (in predicate-argument structure). Such information may refer to argument structures centering around the verb notions and associated information of location and time. The results are stored in a database and used as a basis for question answering, summarization and intelligent browsing. Textract, and other similar systems based on light-

weight NLP techniques, [Harabagiu et. al., 00] achieved surprising quality in the competition of answering fact-based questions in TREC (Text Retrieval Conference) [Voorhess, 99].

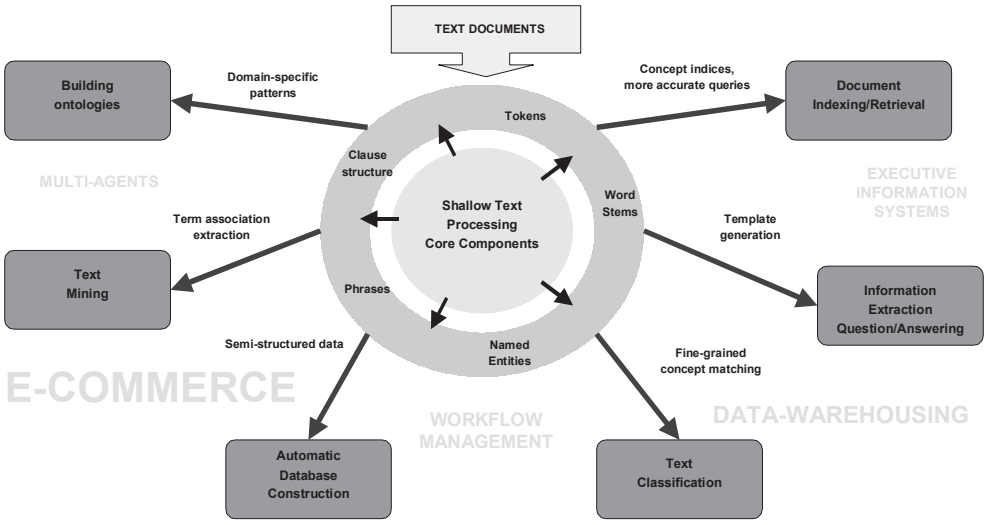
Text Classification

The task of *Text Classification* (TC) is assigning one or more predefined categories from a closed set of such categories to each document in a collection. Traditional approaches in the area of TC use word-based techniques for fulfilling this task. [Riloff and Lorenzen, 98] presented AutoSlog-TS, an unsupervised system that generates domain specific extraction patterns, which was used for automatic construction of high-precision text categorization system. Autoslog-TS retrieves extraction patterns (with single slot) representing local linguistic expressions that are slightly more sophisticated than keywords. Such patterns are not simply extracting adjacent words since extracting information depends on identifying local syntactic constructs (verb and its arguments). AutoSlog-TS takes as input only a collection of pre-classified texts associated with a given domain and uses simple STP techniques and simple statistical methods for automatic generation of extraction patterns for text classification. This new approach to integrating STP techniques in TC proved to outperform classification using word-based approaches. Further, similar unsupervised approaches [Yangarber et. al., 00], using light linguistic analysis were presented for acquisition of lexico-syntactic patterns (syntactic normalization: transformation of clauses into common predicate-argument structure), and extracting scenario-specific terms and relations between them [Finkelstein-Landau and Morin, 99], which shows an enormous potential of shallow processing techniques in the field of text mining.

Text Mining

Text mining (TM) combines the disciplines of data mining, information extraction, information retrieval, text categorization, probabilistic modeling, linear algebra, machine learning, and computational linguistics to discover valid, implicit, previously unknown, and comprehensible knowledge from unstructured textual data. Obviously, there is an overlap between text mining and information extraction, but in text mining the knowledge to be extracted is not necessarily known in advance. [Rajman, 97] presents two examples of information that can be automatically extracted from text collections using simple shallow processing methods: probabilistic associations of keywords and prototypical document instances. Association extraction from the keyword sets allows to satisfy information needs expressed by queries like "find all associations between a set of companies including Siemens and Microsoft and any person". Prototypical document instances may be used as representative of classes of repetitive document structures in the collection of texts and constitute good candidates for a partial synthesis of the information content hidden in a textual base. Text mining contributes to the discovery of information for business and also to the future of information services by mining large collections of text [Abramowicz and Zurada, 01]. It will become a central technology to many businesses branches, since companies and enterprises "don't know what they don't know" [Tkach, 99].

Figure 3: Application potential of shallow text processing



CONCLUSIONS

We have learned that IE technology based on shallow linguistic analysis has been successfully used in various business applications dealing with processing huge collections of free-text documents. The diagram in figure 3 reflects an enormous application potential of STP in various fields of information technology discussed in this paper. STP can be considered as an automated generalized indexing procedure. The degree and amount of structured data a STP component is able to extract plays crucial role for subsequent high-level processing of extracted data. In this way, STP offers distinct possibilities for increased productivity in workflow management [Abramowicz and Szymanski, 02A], e-commerce and data warehousing [Abramowicz et. al., 02B].

The question of developing text processing technology base that applies to many problems is still being major challenge of the current research. In particular, future research in this area will focus on multilinguality, cross document event tracking, automated learning methods to acquire background knowledge, portability, greater ease of use and stronger integration of semantics.

ENDNOTES

[1] There is no standardized definition of the term shallow text processing. Shallow text processing can be characterized as a process of computing text analysis which is less complete than the output of deep text processing systems. It is usually restricted to identifying non-recursive structures or structures with limited amount of structural recursion, which can be identified with high degree of certainty.

2 Most of the STP systems follow the finite-state approach, which guarantees time and space efficiency.

3 The recall of an IE system is the ratio between the number of correctly filled slots and the total number of slots expected to be filled. Analogously, the precision is the ratio between the number of correctly filled slots and the number of all slots filled by the system.

4 Altogether 7 MUC competitions took place (1987 – 1998), where the participants were given the same training data for adaptation of their systems to a given scenario. Analogously, the evaluation was performed using same annotated test data.

REFERENCES

[Abramowicz and Zurada, 01] W. Abramowicz, J. Zurada, *Knowledge Discovery for Business Information Systems*. Kluwer Academic Publishers, Boston, 2001, 431 pp.

[Abramowicz and Szymanski, 02A] W. Abramowicz, J. Szymanski, *Workflow technology supporting information filtering from the Internet*. In Proceedings of IRMA 2002, Seattle, USA, 2002.

[Abramowicz et. al., 02B] W. Abramowicz, M. Kowalkiewicz, P. Zawadzki, *Enhancing searching and navigation of knowledge repositories using Skill Maps tech-*

nology. In Proceedings of IRMA 2002, Seattle, USA, 2002.

[Abramowicz et. al. 02] W. Abramowicz, P. Kalczynski and K. Wecel, *Filtering the Web to Feed Data Warehouses*. Springer, London, 2002.

[Andersen et. al., 92] P.M. Andersen, P.J. Hayes, A.K. Heuttner, L.M. Schmandt, I.B. Nirenburg, S.P. Weinstein, *Automatic extraction of facts from press releases to generate news stories*. In Proceedings of the 3rd Conference on Applied Natural Language Processing, Trento, Italy, 1992, pages 170-177

[Aone et. al., 99] C. Aone, L. Halverson, T. Hampton, M. Ramos-Santacruz, T. Hampton, SRA: *Description of the IE2 System used for MUC-7*. Morgan Kaufmann, 1999

[Aone and Ramos-Santacruz, 00] C. Aone, M. Ramos-Santacruz, RESS: *A Large-Scale Relation and Event Extraction System*. In the Proceedings of ANLP 2000, Seattle, USA, 2000

[Appelt and Israel, 99] D. Appelt and D. Israel, *An Introduction to Information Extraction Technology*. A Tutorial prepared for IJCAI Conference, 1999

[Chinchor, 98] – N. A. Chinchor, *Overview of MUC7 /MET-2*. In Proceedings of the Seventh Message Understanding Conference (MUC7), 1998.

[Costatino et. al., 97] M. Costantino, R.G. Morgan, R. J. Collingham, R. Garigliano, *Natural Language Processing and Information Extraction: Qualitative Analysis of Financial News Articles*. In Proceedings of the Conference on Computational Intelligence for Financial Engineering (CIFEr '97), New York City, March 23-25, 1997,

[Costantino, 99] M. Costantino*, IE-Expert: Integrating Natural Language Processing and Expert System Techniques For Real-Time Equity Derivatives Trading*. In the Journal of Computational Intelligence in Finance, Vol.7, No.2, pp.34-52, March 1999.

[Finkelstein-Landau and Morin, 99] M. Finkelstein-Landau, E. Morin, *Extracting Semantic Relationships between Terms: Supervised vs. Unsupervised Methods*. In proceedings of International Workshop on Ontological Engineering on the Global Information Infrastructure, Dagstuhl Castle, Germany, May 1999, pages 71-80.

[Glasgow et. al., 98] B. Glasgow, A. Mandell, D. Binney, L. Ghemri, D. Fisher, *MITA : An Information-Extraction Approach to the Analysis of Free-Form Text in Life Insurance Applications*. AI magazine, 19(1) :59—71. 1998

[Grishman and Sundheim, 96] R. Grishman and B. Sundheim, *Message Understanding Conference — 6: A Brief History*. In Proceedings of the 16th International Conference on Computational Linguistics (COLING), pages 466—471, Kopenhagen, Denmark, 1996.

[Harabagiu et. al., 00] S. Harabagiu, M. Pasca, S. Maiorano, *Experiments with open-domain textual question answering*. In Proceedings of the COLING-2000. Association for Computational Linguistics, Morgan Kaufmann, 2000.

[Hobbs et. al., 97] J. Hobbs, D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, M. Tyson, *FASTUS - A cascaded Finite-State Transducer for Extracting Information from Natural Language Text*. Chapter 13 in [Roche and Schabes, 97], 1997.

[Humphreys et. al., 98] K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, Y. Wilks, University of Sheffield: *Description of the LaSIE-II System as used for MUC-7*. In Proceedings of the Seventh Message Understanding Conference (MUC-7), 1998

[Jackson et. al., 98] P. Jackson, K. Al-Kofahi, C. Kreilick and B. Grom, *Information extraction from case law and retrieval of prior cases by partial parsing and query generation*. In Proceedings of the ACM 7th International Conference on Information and Knowledge Management, pages 60-67, Washington United States, 1998

[Jacobs et. al., 90] P. Jacobs and L. Rau, *SCISOR: extracting information from online news*. In Communications of the ACM, 33, 11, 1990, pages 88-97

[Lytinen and Gershman, 86] S. Lytinen and A. Gershman, *ATRANS: Automatic Processing of Money Transfer Messages*. In Proceedings of the 5th National Conference of the American Association for Artificial Intelligence, IEEE Computer Society Press, pages93-99, 1993.

[Mihalcea and Moldovan, 01] R. Mihalcea and D. Moldovan, *Document Indexing Using Named Entities*. In Studies in Informatics and Control Journal, Vol. 10, Number 1, March 2001.

[Piskorski and Skut, 00], J. Piskorski, W. Skut, *Intelligent Information Extraction*. In the Proceedings of Business Information Systems'2000, Poznan, Poland, 2000

[Rajman, 97] M. Rajman, *Text Mining, knowledge extraction from unstructured textual data*. In Proceedings of EUROSTAT Conference, Frankfurt, Germany, 1997

[Riloff and Lorenzen, 98] E. Riloff, J. Lorenzen, Extraction-based text categorization: *Generating domain-specific role relationships automatically*. In Strzalkowski, T., ed., Natural Language Information Retrieval, Kluwer Academic Publishers, 1998

[SAIC, 98] SAIC, editor, *Seventh Message Understanding Conference (MUC-7),* http://www.muc.saic.com, 1998

[Srihari and Li, 99] R. Srihari and W. Li, *Information extraction supported question answering*. In Proceedings of the Eighth Text Retrieval Conference (TREC-8), 1999

[Tkach, 99] D. Tkach, *The pillars of knowledge management*. In Knowledge Management, 2(3), page 47

[Voorhess and Tice, 99] E. Voorhess and D. Tice, *The TREC-8 Question Answering Track Evaluation*. National Institute of Standards and Technology, Gaithersbugh, 1999

[Yangarber et. al., 00] R. Yangarber, R. Grishman, P. Tapanainen and S. Huttunen, *Unsupervised Discovery of Scenario-Level Patterns for Information Extraction*. In Proceedings of Conference on Applied Natural Language Processing ANLP-NAACL 2000, May 2000, Seattle, WA