International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 6 3686 - 3689

Information Extraction using Context-free Grammatical Inference from Positive Examples

Ramesh Thakur
International Institute of Professional Studies
Devi Ahilya University
Indore, India
r_thakur@rediffmail.com

***Abstract*—** Information extraction from textual data has various applications, such as semantic search. Learning from positive example have theoretical limitations, for many useful applications (including natural languages), substantial part of practical structure (CFG) can be captured by framework introduced in this paper. Our approach to automate identification of structural information is based on grammatical inference. This paper mainly introduces the Context-free Grammar learning from positive examples. We aim to extract Information from unstructured and semi-structured document using Grammatical Inference.

Keywords- *Knoledge discovery, Grammatical inference, Context-free grammar.*

*****

I. INTRODUCTION

The computer and information systems have gained significant achievements over the last two decades as expressed by their dominance in various business and scientific applications. The management of unstructured (text) data is recognized as one of the major unanswered problem in information technology due to unavailability of suitable tools and techniques to transform it for business intelligence. As estimated [1, 2] more than 85% of all business information exists in the form of unstructured and semi-structured data, it is commonly available in the form of text documents and web pages. The most common document used in the web pages are HTML which are intended to be browsed by human for viewing, and not for the application to process it. The unstructured and semi-structured data is without a well defined schema or data model and do not have a global schema.

Grishman and Sundheim [3] described Information Extraction as "The identification and extraction of instances of a particular class of events or relationships in a natural language text and their transformation into a structured representation (*e.g.* database)." Thus Information Extraction is a process to automate the extraction of structured information such as entities, relationships between entities and attributes describing the entities from unstructured and semi-structured data, which enables much richer query system to process data instead of keyword search alone. With emergence of Natural Language Processing (NLP) techniques, the extraction of structure from unstructured source has become challenging research area in the last two decades.

Extracting information from web pages is usually done by software called wrappers. Approach used in wrapping HTML document is based on manual technique [4, 5, 6]. The problem with manually coded wrappers is that writing them is difficult and time consuming job. As the websites are updated from time-to-time, resulting change in semantics makes maintenance more difficult. As a result the key challenges in semi-structured document information extraction are to develop the technique that allows the automation of extraction process. Our approach to automate identification of structural information is based on grammatical inference.

II. TYPES OF DOCUMENTS

The increase of computer storage and processing power has opened the way for more resource intensive applications, which used to be unreachable. The trend in increase of resources also creates structured corpora from different types of documents. These new applications can already be found in several fields, for example Natural language parsing [7, 8, 9, 10], Evaluation of natural language grammars [11, 12], Machine translation [13, 14, 15] etc. The different types of document which are available in electronic form for which information extraction are required are unstructured, semi-structured, Web documents and structured text.

A. Unstructured Document (Free Text)

Un-structured document (free text) is information that does not have a pre-defined data model (schema). Unstructured information is typically text, but may contain data such as dates, numbers, and facts etc. These documents are prepared using natural language schemas; hence they have irregularities and ambiguities that make it difficult for traditional computer application to process it. The unstructured data has following characteristics.

- data can be of different types,
- not necessarily following any format or sequence,
- does not follow any rules,
- is not predictable.

The Information extraction system for unstructured data generally uses natural language techniques, and extract rules, typically based on patterns involved in syntactic relations between words or classes of words.

B. Semi-structured Documents

Semi-structured data [16] is a form of unstructured that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, It contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data e.g., bibliographic data, Web data, *Electronic data*

interchange (EDI), scientific data etc. Semi-structured data have following characteristics.

- organized in semantic entities,
- similar entities are grouped together,
- entities in same group may not have same attributes,
- order of attributes not necessarily important,
- not all attributes may be required,
- size of same attributes in a group may differ,
- type of same attributes in a group may differ.

Therefore, it is also known as schema-less or self-describing structure. Semi-structured text is ungrammatical and often telegraphic in style, and does not follow any rigid format. Natural Language Processing (NLP) techniques are deployed to design rules for extraction of information form free text will usually not work for semi-structured text. Hence, for semi-structured texts the traditional techniques of Information Extraction can not be used, and simple rules used for structured text will not be sufficient. Some form of structuring is, however present in semi-structured text, and extraction patterns are often based on tokens and delimiters like for instance HTML tags, syntactic and semantic information can only be utilized to a limited extent.

C. *Web Documents*

The Explosive growth of the World-Wide-Web has resulted in huge amount of information source on the Internet. Generally these information are semi-structured (HTML), we can also find structured and unstructured. The information is also dynamic, it contains hyperlinks and may be represented in different forms and is globally shared over multiple sites and platforms. The web is driving force of research on information extraction form semi-structured and unstructured data.

There is different view about Web Pages. Some researchers define all Web Pages as semi-structured information. As they all contain the structuring information concerning display style i.e. HTML tags. These tags are the instruction to browser for presentation. However [17] give a better categorization of types of web pages: A Web Page that provides itemized information is structured, if each attribute in a tuple can correctly be extracted based on some uniform syntactic clues, such as delimiters or the orders of attributes. Semi-structured Web Pages, however may contains tuples with missing attributes, attributes with multiple values, variants attribute permutations, and exceptions. A Web Page is unstructured if linguistic knowledge is required to extract the attributes correctly. Since the semantic of Web Page are restricted to each Web page or a class of Web pages, hence it is not fully structured information. We consider Web Pages are semi-structured information.

D. *Structured Text*

Structured text is defined as textual information in a database or file following a uniform predefined and strict format having following characteristics.

- have the same defined format,
- have a predefined length,
- are all present,
- and follow the same order.

Such information can easily be correctly extracted using the format description. Usually a simple technique is sufficient for extracting information form text provided that the format is known otherwise the format must be learned.

III. INFORMATION EXTRACTION

Due to the growth of the internet more and more text becomes available online, which resulted in need for a system that extract information automatically form text data. An Information Extraction (IE) system can serve as a front end of high precision information retrieval or text routing, as a first step in knowledge discovery systems that look for trends in massive amounts of text data, or as input to an intelligent agent whose actions depend on understanding the content of text-based information. The IE system must work for semi-structured and unstructured data such as tabular information to free text (news, stories etc.). A Key element of such system is a set of text extraction rules that identify relevant information to be extracted.

A. *Information Extraction and Information Retrieval*

Information Extraction (IE) is different from the more mature technology of Information Retrieval (IR). Rather than to extract information the objective of IR is to select a relevant subset of documents from a large collection based on user query. Manning and Raghavan [18] described Information Retrieval (IR) as follows: “Information retrieval (IR) is to find the material (usually documents) of an unstructured nature (usually text) that satisfies information need from within large collections (usually stored on computers)”. In Contrast, goal of Information extraction (IE) is to extract relevant information from the documents. Hence the two techniques are complementary, and used in combination they can provide more powerful tools for text processing [19].

Not only the IE and IR differ in aims, they also usually differ in the technique. The IE has emerged from research on rule-based system in computational linguistic and natural language processing, while information theory, probability theory and statistics have influenced the IR [19].

IV. APPROACHES TO INFORMATION EXTRACTION (IE)

There are two main approaches to the design of IE system. The first is Knowledge engineering approach, and the second is automatic training approach [20]. In Knowledge engineering approach grammar expressing rules of the system are constructed by hand coded using knowledge of application domain. The skills and knowledge base play an important role for the system.

For the automatic training approach there is no need for system expertise when customizing the system for a new domain. The system needs to be trained on a set of training documents. Once a training corpus has been annotated, a training algorithm runs *i.e.* training the system for analyzing novel texts. This approach is faster than the knowledge engineering approach, but requires a sufficient volume of tanning data.

V. EVALUATION OF GRAMMAR INFERENCE METHODS

The evaluation of Information Extraction using grammatical inference problem has different approaches. Generally, the evaluation of grammar inference algorithm is carried out by giving input to the algorithm a set of unstructured data and evaluating its output (grammar rules). Three principal evaluation strategies usually applied for evaluating grammar inference algorithm [21].

- Looks-Good-to-me,

- Compare Against Treebank,
- Rebuilding Known Grammars.

A. *Looks-Good-to-me*

In Looks-Good-to-me strategy grammar inference algorithm is applied to a piece of unstructured text and the resulting grammar is qualitatively evaluated on the base of the linguistic intuitions of the evaluator, that highlights the grammatical structures which look "good". This kind of evaluation is mainly conducted by experts who have specific knowledge of the syntax of the language.

B. *Compare Against Treebank*

In *Compare Against Treebank* evaluation strategy consists of applying the grammar inference algorithm to a set of plain unstructured sentences that are extracted from an annotated treebank, which is selected as a "gold standard". The structured sentences generated by the algorithm are then compared against the original structured sentences from the Treebank and the *recall* and *precision* are computed.

C. *Rebuilding Known Grammars*

The Rebuilding Known Grammars approach is another evaluation strategy. This method, starting from a pre-defined (simple) grammar, generates a set of example sentences, which are given as input to the grammar inference algorithm and the resulting grammar is compared manually to the original grammar. If the inferred grammar is similar or equal to the original grammar then the learning system is considered good.

Precision, which measures the number of correctly learned constituents as a percentage of the number of all learned constituents. The higher the precision, the better the algorithm is at ensuring that what has been learned is correct.

$$\text{Precision} = \frac{\sum \text{Correctly Learned Constituentes}}{\sum \text{Learned Constituentes}}$$

Recall, which measures the number of correctly learned constituents as a percentage of the total number of correct constituents. The higher the recall, the better the algorithm is at not missing correct constituents.

$$Recall = \frac{\sum Correctly\ Learned\ \text{Constituentes}}{\sum possible\ correct\ \text{Constituentes}}$$

When comparing the performance of different systems, both precision and recall must be considered. However, as it is not straightforward to compare the two parameters at the same time, various combination methods have been proposed. One such measure is *F-Score*, which combines precision, *P* and recall *R*, in a single measurement as follows:

$$F = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Using the *F*-score, the relative performance of systems reporting different values for recall and precision, can easily be compared.

VI. Information Extraction using Grammatical Inference

The learning of the syntax of the language is usually referred to as grammatical inference or grammar induction. The product of this process is a grammar, a formalism that captures the syntax of a language. The objective of the grammatical inference to infer a formal language, such as context-free grammar, which describes the given sample set. These grammar rules will be used to create structural descriptions of the unstructured and semi-structured documents. In automated grammar learning, the task is to infer grammar rules from given information about the target language. Information Extraction from textual data has various applications, such as semantic search [22]. If the sentences confirm to a language described by a known grammar, several techniques exist to generate the syntactic structure of these sentences. Parsing [23] is one of such technique that rely on knowledge of grammar.

A. *Grammar Learning*

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience [24]. When we use this general definition we can say that the task is to learn a grammar, the performance measure could be a metric that calculates the difference between the grammar found and the target grammar (i.e., the grammar to be learned) and the experience could be the linguistic input in one or another form (e.g. unstructured or semi-structured text).

We can define a collection of grammars that are possible to learn. We call this collection the *hypothesis space*. One of these grammars is the grammar that the learning algorithm is supposed to learn the *target grammar*. Then, we could say that an algorithm for grammatical inference typically should identify an *hypothesis grammar* g_h from the hypothesis space *G* as the *target grammar* g_t. Note that $g_t, g_h \in G$. The process of induction is shown in the Fig. 1.

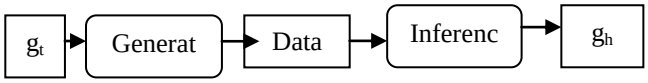

Figure 1. :Induction process for the target grammar [25].

B. *Identification in the Limit*

In automated grammar learning, the task is to infer grammar rules from given information about the target language. The sentences (or strings of alphabet) are given as examples for such learning. If the example belongs to the target language, it is called as a positive example. Otherwise, it is called as a negative example. A language that can be inferred by looking at a finite number of positive examples only said to be identifiable in the limit [26]. In the seminal identification in the limit model by Gold [27], an infinite sequence of examples of the unknown language is given to the learning algorithm, and the algorithm tries to successfully learn the grammar or the language after some finite time. Gold has presented one main result: It is not possible to identify the target language from only positive examples for any super finite class of languages, i.e., the class of languages that contains all finite languages and, at least, one infinite language [28]. This condition holds for all basic languages, such as regular languages. Hence, the general learning tasks of learning from positive examples alone have proven to be impossible. Since the learning of such a general class of languages is not possible, many researchers have considered further restrictions on the class of languages.

Although positive example have theoretical limitations, for many useful applications (including natural languages),

substantial part of practical structure can be captured by framework introduced in this thesis. The expressive power context-free grammar, for modeling, frequent linguistic phenomena of natural language and lower parsing complexity. We mainly restrict to context-free Grammar inference from semi-structured and unstructured data.

REFERENCES

[1] Blumberg, R., & Atre, S. "The problem with unstructured data." *DM REVIEW, 13,* pp 42-49 2003..

[2] James, S., Mark, D. Roger, F., Melliyal, A., Jean, I., & Xavier, L. "Managing Unstructured Data with Oracle Database 11g". *An Oracle White Paper* , pp. 1-9 Feb 2009.

[3] B. M. Sundheim, "Overview of the third message understanding evaluation and conference," *In Proceedings of the Third Message Understanding Conference (MUC-3),* pp. 3–16, San Diego, CA, 1991.

[4] Atzeni, P., Mecca, G. "Cut and Paste." *In Proceedings of the 16th ACM SIGMOD International Symposium on Principles of Database Systems (PODS'97) (Tucson, AZ)*. ACM, New York, pp 144–153 1997.

[5] Hammer, J., Garcia-molina, H., Cho, J., Aranha, R., and Crespo, A. "Extracting Semi-structured Information from the Web". *In Proceedings of the Workshop on the Management of Semistructured Data (in conjunction with ACM SIGMOD 1997). ACM, New York* (1997).

[6] Sahuguet, A., and Azavant, F. "Web Ecology: Recycling HTML pages as XML Documents using W4F". *In Proceedings of the 2nd Workshop on the Web and Databases (WebDB'99) (in conjunction with SIGMOD'99). ACM, New York* 1999.

[7] Allen, J. "Natural Language Understanding. Benjamin/Cummings", *Redwood City:CA, USA, 2nd edition.* 1995.

[8] Bod, R. "Enriching Linguistics with Statistics: Performance Models of Natural Language". *PhD thesis, University of Amsterdam, Amsterdam, the Netherlands,* 1995.

[9] Charniak, E. "Statistical parsing with a context-free grammar and word statistics". *In Proceedings of the Fourteenth National Conference on Artificial Intelligence,* pages 598–603. American Association for Artificial Intelligence (AAAI), 1997.

[10] Jurafsky, D. and Martin, J. H. "Speech and Language Processing". *Prentice Hall, Englewood Cliffs: NJ, USA,* (2000).

[11] Black, E., Abney, S., Flickinger, D., Gdaniec, C., Grishman, R., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., Klavans, J., Liberman, M., Marcus, M., Roukos, S., Santorini, B., and Strzalkowski, T. "A procedure for quantitatively comparing the syntactic coverage of English grammars". *In Proceedings of a Workshop Speech and Natural Language,* pp 306–311, (1991).

[12] Sampson, G. "A proposal for improving the measurement of parse accuracy". *International Journal of Corpus Linguistics,* 5(1) pp 53–68, 2000.

[13] Poutsma, A. "Data-Oriented Translation-using the Data-Oriented Parsing framework for machine translation". *Master's thesis, University of Amsterdam, Amsterdam, the Netherlands,* 2000.

[14] Sadler, V. and Vendelmans, R. "Pilot implementation of a bilingual knowledge bank". *In Proceedings of the 13th International Conference on Computational Linguistics (COLING); Helsinki, Finland,* pp 449–451, 1990.

[15] Way, A. "A hybrid architecture for robust MT using LFG-DOP", *Journal of Experimental and Theoretical Artificial Intelligence, Special Issue on Memory-Based Language Processing* 11(4),. 1999.

[16] Peter Buneman, "Tutorial on semi-structured data", *from Symposium on Principles of Database Systems,* 1997.

[17] C-H Hsu and M-T Dung "Generating Finite-State Transducers for semi-structured Data Extraction From the Web". *Information System,* Vol 20. No. 8, pp 521-538, 1998.

[18] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang Ning Tan "Web usage mining: Discovery and Applications of usage patterns from web data", *SIGKDD Explorations,* 1(2) pp 12-33, 2000.

[19] R. Gaizauskas, Y Wilks. "Information Extraction: Beyond Document Retrival", *Computational Linguistics and Chinese Language Processing,* vol. 3, no. 2, pp 17-60, 1998.

[20] D. E. Appelt, D.J. Israel. "Introduction to information extraction technology ", *Tutorial for IJCAI-99, Stockhoam,* 1999.

[21] D'Ulizia, Arianna, Fernando Ferri, and Patrizia Grifoni. "A survey of grammatical inference methods for natural language learning." *Artificial Intelligence Review* 36, No. 1 pp 1-27, 2011.

[22] P. Palaga, L. Nguyen, U. Leser, and J. Hakenberg, "High-performance information extraction with AliBaba ," *In Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, EDBT '09 ACM New York* pp 1140–1143, 2009.

[23] Allen J., "Natural Language Understanding," *The Benjamin/Cummings Publishing Company, Inc., Redwood City, CA, USA. Second Edition,* 1995.

[24] Mitchell, T. M. "Machine learning" New-York, USA: MIT Press, McGraw-Hill pp 2–3, 1997.

[25] Menno M. van Zaanen "Bootstrapping Structure into Language Alignment-Based Learning*", Phd thesis, The University of Leeds School of Computing,* 2001.

[26] E. M. GOLD, "Language identification in the limit," *Inform Control.* vol.10, no.5, pp 447–474,1967.

[27] E M. Gold, "Complexity of automaton identification from given data," *Inform. Control,* vol. 37, pp 302–320, 1978.

[28] Agrawal and R. Srikant. "Mining Sequential Patterns." *In Proceedings of the International Conference on Data Engineering (ICDE), Taipei, Taiwan,* 1995..