INFORMATION EXTRACTION METHODS AND EXTRACTION TECHNIQUES IN THE CHEMICAL DOCUMENT'S CONTENTS: SURVEY

Muawia Abdelmagid[1], Ali Ahmed[2] and Mubarak Himmat[3]
[1]Deanship of Scientific Research, University of Dammam, Dammam, KSA
[2]Faculty of Engineering, Karary University, Khartoum, Sudan
[3]Faculty of Computing, Universiti Teknologi Malaysia, Skudai, Malaysia
E-mail: muawiasadig@yahoo.com

ABSTRACT

The volume of electronic documents has rapidly increased and the scientific literature has increased too. These huge documents contain considerable information, but it has to be retrieved and managed in a constructive and useful way. Information Extraction (IE) is the field of extracting useful information using different methods and approaches by means of Natural Language Processing (NLP). Researchers still continue to try to identify proper methods to extract information from texts, such as opinions on the internet, medical data, clinical reports, medical reports, notes, papers, patents, etc. Recently a new trend to expand working in IE is taking place by enriching the extraction process to include the extraction of information from images and videos. In this paper, the classification of IE tasks is discussed, as well as the proposed methods and techniques of IE from chemical documents. A more focused approach is then taken into consideration regarding biomedical language processing and ontology. Finally, the paper discusses some of the challenges that the IE field is facing.

Keywords: information extraction, information retrieval, extraction methods, chemical extraction.

INTRODUCTION

In past decades, IE system development has grown rapidly; gaining attention from a lot of researchers. IE dates back to the 1950s when [1] suggested a system that used statistical information to provide an identification of the content-bearing portion of document texts. This system became the starting point of much research and drew the attention of computer science researchers. IE systems have undergone a continuous and interesting development and recently, many extended works have been suggested and developed. The core idea behind IE systems is the aim to retrieve specific and desired information from documents of natural language text. These extraction processes are conducted automatically using computer methods, and this process has developed rapidly within the timeframe because of the development of NLP tools and techniques. IE systems have been developed to extract information from different types of text; structured text, semi-structured and free text [2]. Recently, this method has been extended to include the extraction of information from images and videos. Each type of document previously mentioned has several steps and rules for extraction; here we will discuss the differences between the three types of text documents.

Free text: This type of document contains unstructured text such as news, stories, etc. The extraction process in this kind of text document is difficult because it contains variant information with a week relation.

Semi-structured text: The text that is presented and formatted in a high quality manner in a specific domain; for instance, information about the economy, education, medicine and so forth. A lot of work has been carried out on semi-structured text. [2] has proposed an automatic IE method. Their proposed system was very effective and valuable in many semi-structured types of sources. Furthermore, in 1997, [3] proposed an automatic system that automatically generated wrappers from a variety of internet sources.

Structured text: This type of document is highly structured, organised, and well formatted. A key example is databases.

This paper discusses some important aspects of IE concepts, together with the methods and techniques that are applied in IE. The rest of this paper is organised as follows: the first section focuses on IE classification tasks, then Name Entity Recognition and Approach techniques are discussed. Subsequently, the paper provides a review on the latest works that have been undertaken to extract information from different chemical documents and other sources. Then, we discuss some types of methods and techniques for IE from chemical and biomedical fields. Following that, we introduce the main ideas of ontology-based IE and some information about biomedical language processing. Then, some challenges facing the IE field are outlined and finally the paper is concluded.

CLASSIFICATION OF IE TASKS

The process of IE is generally categorised into three processes. Firstly, the system extracts individual “facts” from the text document through local text analysis. The next process integrates the facts to produce a larger fact or infer a new fact. The last process takes place after the fact integration when the pertinent facts are translated to an output format. Encompassing all of these processes, IE is defined as “the task of automatically identifying, collecting and normalising relevant information from natural language texts and producing a set of target knowledge structures as output"[4]. Recently, a lot of work on statistical methods and learning methods has been applied to the IE field, but generally, we can classify IE into several tasks. The first task is Name Entity Recognition (NER), which is defined as "the process of finding mentions of specified things in running text (person, location, organisation)" [5]. The second task is noun phrase co-reference resolution which is "the process of checking whether two expressions in the natural language refer to the same entity and resolving anaphoric references by pronouns and definite noun phrases". The third task is cross-document co-reference resolution which is used when the same name of an entity, person, organisation, location is discussed in more than one text source. The fourth task is semantics role recognition which is a set of roles that range from the narrow meaning "specific" to the wide meaning "general". The recent work that is proposed by [6] adds a good enhancement on rule-based IE. The fifth task is entity relation recognition which is the task of identifying and detecting the relation between entities and the relation that is possibly typed with a semantic role. The last task is timex recognition and resolution; in this task all temporal expressions like events, absolute, event anchored expressions are detected and recognised.

NAME ENTITY RECOGNITION AND APPROACHES TECHNIQUES

The Named Entity Recognition (NER) term is proposed and construed in the large (MUC-6) Message Understanding Conference that discussed the IE area and put forward the IE research methods. NER is an important process and IE method and involves the processing of extracted information from both structured and unstructured documents. It is used to identify all expressions referring to a specific entity or object in the documents or texts, like names of locations, places, countries, people, and so forth. NER is considered to be the basic task of IE and is one of the essential processes of the Natural Language Processing (NLP) system. NER is divided into two tasks; the first task is identifying the names in text documents, and the second task is classifying these names into dictionaries or groups of predefined classes of interest, such as country names, company names, a person's name, and so forth. In chemical and biomedical fields, the NER focuses on protein or gene names, compound names, drug disruptions, and potion records.

IE automatic approaches

The IE systems have taken different orientations and approaches. As an example, we outline the IE Approach of Machine Learning which is based on the automatic extraction of patterns using machine learning techniques. This approach can be categorized into four groups as shown in Figure-1. These categories are defined below:

1. Supervised learning systems approach is proposed and used in many works and requires large amounts of data to set as training data. It then uses the machine learning rules and techniques to extract the required information;
2. Semi-supervised learning systems approach (e.g., Mutual Bootstrapping);
3. Unsupervised learning systems approach is quite different from supervised learning systems for it uses corpus bootstrapping methods that depend on small seed rules that are learnt from an annotated system;
4. Hybrid NER systems; this approach uses a combination of a dictionary base and Conditional Random.

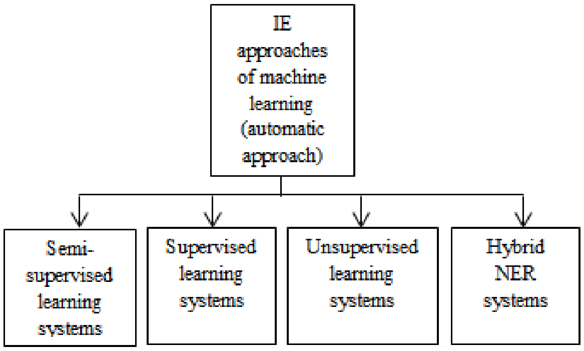
IE approaches of machine learning (automatic approach)
Semi-supervised learning systems
Supervised learning systems
Unsupervised learning systems
Hybrid NER systems

Figure-1. Automatic approach categories.

IE FROM CHEMICAL DOCUMENTS

As previously mentioned, IE automatic systems have been applied in several fields, such as news extraction, literature extraction, bioinformatics, and so forth. The development happens in Natural Language Processing and its applications increase so as to involve the extraction methods in different areas. In the areas of chemical, biomedical and other related areas, a lot of IE methods have been developed. The work done by [7] proposed a technique that is used to extract facts from journals of the American Chemical Society (ACS); the extracted facts are about chemical reactions. Another work done by [8] proposed an extraction method using the technique of nearest neighbour k(NN). A further work which was a little similar to this work was also conducted by [9] who proposed a method that automatically extracts information on protein - protein interactions from scientific literature. The method consists of different stages; firstly, they select the target which means the documents of the protein that are used as the source documents. In the next stage, they identify all the protein names. Subsequently, all compound sentences or complex sentences are processed. Their method then automatically recognizes and extracts the protein - protein interactions.

Some other related work proposed an extraction technique that referred to the text mining approach [10]. Most of mentioned methods are used to extract all frequently occurring biological relations among a pair of biological concepts to identify relevant information. In [11], a chemical IE system (ChemEx) has been developed with the main aim of extracting information from both texts and images. [12] has proposed an IE method to extract information from biomedical literature. This method was based on integration of semantic information to multiple kernels for extracting protein - protein interactions. Generally, the biomedical and chemical literature has become an important source of useful information. This useful information needs to be extracted using intelligent IE systems. Thus, much development is required to enrich the process of extracting information from chemical and biomedical domains and accordingly, we notice the on-going development in IE systems.

TYPES OF IE TECHNIQUES AND METHODS PROPOSED FOR MEDICATION, CHEMICAL AND BIOMEDICAL AREAS

The process of extracting knowledge and information automatically from text data has recently become one of the most important and active fields of research; especially IE techniques that are used to extract information from chemical and biomedical literature. The common documents that contain chemical and biomedical information are medical records, patient discharge summaries, biological documents, drug description, scientific papers, patents, and so forth. Over the last few decades, many IE methods that use different techniques have been proposed and developed for the chemical and biomedical domains. The table below Table-1 shows examples of some IE methods that have been proposed for the aforementioned fields. The table has four columns; proposed method, used techniques, the extracted document type, and the references.

Table-1. Examples of some chemical related IE methods.

Proposed method	Used techniques and tools	Used text document and dataset	Reference
Proposed method uses computational linguistics techniques to analyse the natural language text	Lexical and syntactic aspects	American Chemical Society journals	**[7]**
Developed tools to automate the problem list using NLP to extract potential medical problems from free-text documents in a patient's (EMRs) electronic medical records.	NLP	Free-text documents in a patients (EMR) Electronic Health Record	**[13]**
The proposed automatic system that uses the NLP system (MedEx) to extract structured medication information from discharge summaries.	NLP	Discharge summaries and clinic visit notes from the Synthetic Derivative (SD) database	[14]
(BIEQA) Ontology-based biological IE and answers to queries. To extract information about the likelihood of various biological relation occurrences within tagged biological documents.	Ontology-based	Biological documents	[10]
Proposal to develop a system to extract information from Polish medical texts.	Rule-based IE system	Mammography reports and hospital records of diabetic patients	[15]
The proposed automatic method for protein –protein extraction from documents of scientific literature by searching with protein names.	MEDLINE search by searching with different key words Dictionary-based NER Systems	Scientific literature. From web	[9]
The proposed system is (ChemSpot) a NER tool for identifying chemical names that are mentioned in natural language texts.	The system used a hybrid approach that combines a Conditional Random Field with a dictionary	The system could be used with any natural language texts and documents of bimiformatics.	[16]

BIOMEDICAL LANGUAGE PROCESSING AND ONTOLOGY

With its rapid growth, scientific literature contains a wealth of information on many different fields and numerous methods have been developed to identify, summarise, extract, analyse, and classify. In the IE field from biomedical literature, we note that the enrichment occurs when using the biomedical language processing system, which is considered a very useful tool in this field. Recently, enhancements have been made in the biomedical language processing (BLP) area. In BLP, a lot of new computational tools and methods have been proposed to provide the process of taking the human generated texts as input and generating different methods and systems to retrieve, classify, detect plagiarism, and extract information. The system extracts factual information, and now there are many works that focus on molecular biology and chemical related documents. The work that has been conducted by [17] describes the effort that has been made in respect of the free text in biomedical databases. Other important work carried out by [18] defines the corpora which are annotated with respect to structural and linguistic characteristics in the biomedical field. [19] proposed a framework that used conditional random fields to recognise multiple entity classes in biomedical abstracts.

Ontology

Ontology has several definitions "a formal explicit specification of a shared conceptualise-action. 'Conceptualisation' refers to an abstract model of some phenomenon in the world formed by identifying the relevant concepts of that phenomenon. 'Explicit' means that the type of concepts used and the constraints on their use are clearly defined. 'Formal' refers to the fact that the ontology should be machine understandable. 'Shared' reflects the notion that ontology captures consensual knowledge" [20]. Another definition defined by [21] which is considered more accurate is "a system that processes unstructured or semi-structured natural language text through a mechanism guided by ontologies to extract certain types of information and presents the output using ontologies". The ontology structure is organised in a hierarchical way to identify any relation between concepts. The result of this property may be used to quantify the similarity between the concepts and, implicitly, the terms semantic similarity between these terms are then used to designate these concepts [22].There are three different ontology approaches as shown in Figure-2 which cover all types of text from free text, semi-structured and structured text.

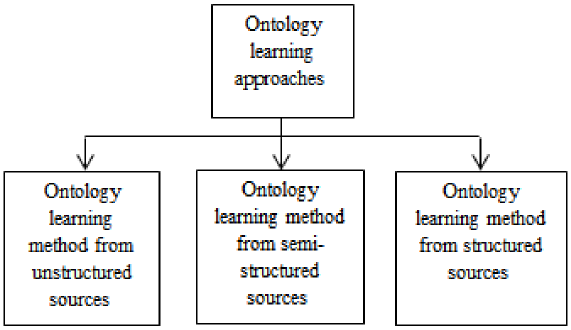
Ontology learning approaches
Ontology learning method from unstructured sources
Ontology learning method from semi-structured sources
Ontology learning method from structured sources

Figure-2. Ontology learning approaches.

Within these approaches, there are many techniques that have been proposed, such as linguistic techniques, statistical techniques and machine learning algorithms. The work done by [20, 23, 24] describes and covers many different approaches and techniques in ontology-based extraction. Natural language analysis techniques are applied to develop and enrich ontology method techniques. There are three main groups of methods.

- Methods of learning from unstructured sources.

This group is concerned with applying natural language analysis techniques to develop ontologies, and it is dependent on NLP obtained linguistic annotation processes with selected corpus.

- Method of learning from semi-structured.

This method is concerned with eliciting ontology from documents that have a predefined structure.

- Method of learning from structured text

This method relies on the ontology that is built by extracting relevant relations and concepts. The type of documents and sources here are fully structured as data from databases.

CHALLENGES

The IE processes are very complicated because they are mainly based on the automatic recognition of human language terms. Further more, the huge amount of dynamic collections of diverse materials is considered one of the critical challenges that is facing IE. Generally, challenges open the research door for IE in general; especially in the biomedical domain because the chemical text is quite different. Usually the chemical text includes complicated symbols and images which are difficult for the NLP to understand. Until now, there has been no full and complete dictionary that contains all, or even most, of the biological named entities and some chemical terms and names have multi-word meanings. All these challenges motivate researchers to work hard in this field to provide appropriate solutions for enhancing the automation process of IE systems.

CONCLUSIONS

The automation process of IE needs to be implemented in different domains. Due to the exponential growth in the number of electronic documents containing a lot of information that needs to be extracted and managed usefully, numerous methods and techniques are being proposed in the IE field to automate these processes. This study introduces some of the important IE concepts and focuses first on the classification of IE tasks, before introducing the ontology-based IE. A discussion on the different types of IE techniques and methods that are proposed for medication and chemical areas is also provided. In addition, the paper discusses and illustrates some works that have conducted in the chemical IE area. Finally, some of the challenges faced by the IE field have been introduced.

ACKNOWLEDGEMENTS

This work was funded by the Deanship of Scientific Research of the University of Dammam (Kingdom of Saudi Arabia), for the Project Grant Number (2014329).

REFERENCES

[1] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," IBM Journal of research and development, vol. 1, pp. 309-317, 1957.

[2] S. Soderland, "Learning information extraction rules for semi-structured and free text," Machine learning, vol. 34, pp. 233-272, 1999.

[3] N. Ashish and C. A. Knoblock, "Wrapper generation for semi-structured internet sources," ACM Sigmod Record, vol. 26, pp. 8-15, 1997.

[4] P. Cimiano, U. Reyle, and J. Šarić, "Ontology-driven discourse analysis for information extraction," Data & Knowledge Engineering, vol. 55, pp. 59-83, 2005.

[5] C. De Bolós, M. Garrido, and F. X. Real, "MUC6 apomucin shows a distinct normal tissue distribution that correlates with Lewis antigen expression in the human stomach," Gastroenterology, vol. 109, pp. 723-734, 1995.

[6] P. Kluegl, M. Toepfer, P.-D. Beck, G. Fette, and F. Puppe, "UIMA Ruta: Rapid development of rule-based information extraction applications," Natural Language Engineering, pp. 1-40, 2014.

[7] E. M. Zamora. and P. E. Blower Jr, "Extraction of chemical reaction information from primary journal text using computational linguistics techniques. 1. Lexical and syntactic phases," Journal of chemical information and computer sciences, vol. 24, pp. 176-181, 1984.

[8] I Mani. and I Zhang., "kNN approach to unbalanced data distributions: a case study involving information extraction," in Proceedings of Workshop on Learning from Imbalanced Datasets, 2003.

[9] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi, "Automated extraction of information on protein–protein interactions from the biological literature," Bioinformatics, vol. 17, pp. 155-161, 2001.

[10] M. Abulaish and L. Dey, "Biological relation extraction and query answering from medline abstracts using ontology-based text mining," Data & Knowledge Engineering, vol. 61, pp. 228-262, 2007.

[11] A Tharatipyakul., S Numnark., D Wichadakul. and S Ingsriswang., "ChemEx: information extraction system for chemical data curation," BMC bioinformatics, vol. 13, p. S9, 2012.

[12] L Li., P Zhang., T Zheng., H Zhang., Z Jiang. and D Huang. "Integrating Semantic Information into Multiple Kernels for Protein-Protein Interaction Extraction from Biomedical Literatures," PloS one, vol. 9, p. e91898, 2014.

[13] S. Meystre. and P. J. Haug., "Natural language processing to extract medical problems from electronic clinical documents: performance evaluation," Journal of biomedical informatics, vol. 39, pp. 589-599, 2006.

[14] H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny, "MedEx: a medication information extraction system for clinical narratives," Journal of the American Medical Informatics Association, vol. 17, pp. 19-24, 2010.

[15] A. Mykowiecka, M. Marciniak, and A. Kupść, "Rule-based information extraction from patients' clinical data," Journal of biomedical informatics, vol. 42, pp. 923-936, 2009.

[16] T. Rocktäschel, M. Weidlich, and U. Leser, "ChemSpot: a hybrid system for chemical named entity recognition," Bioinformatics, vol. 28, pp. 1633-1640, 2012.

[17] A. T. McCray, A. R. Aronson, A. C. Browne, T. C. Rindflesch, A. Razi, and S. Srinivasan, "UMLS knowledge for biomedical language processing," Bulletin of the Medical Library Association, vol. 81, p. 184, 1993.

[18] K. B. Cohen, P. V. Ogren, L. Fox, and L. Hunter, "Corpus design for biomedical natural language processing," in Proceedings of the ACL-ISMB workshop on linking biological literature, ontologies and databases: mining biological semantics, 2005, pp. 38-45.

[19] B. Settles, "Biomedical named entity recognition using conditional random fields and rich feature sets," in Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, 2004, pp. 104-107.

[20] A. Gómez-Pérez and D. Manzano-Macho, "An overview of methods and tools for ontology learning from texts," The knowledge engineering review, vol. 19, pp. 187-212, 2004.

[21] D. C. Wimalasuriya and D. Dou, "Ontology-based information extraction: An introduction and a survey of current approaches," Journal of Information Science, 2010.

[22] P. W. Lord., R D. Stevens., A Brass. and C A. Goble. "Semantic similarity measures as tools for exploring the gene ontology," in Pacific Symposium on Biocomputing, 2003, pp. 601-612.

[23] P Buitelaar., P Cimiano. and B Magnini., Ontology learning from text: An overview vol. 123, 2005.

[24] A Maedche. and S. Staab, Ontology learning: Springer, 2004.