Journal of Theoretical and Applied Information Technology
20th September 2014. Vol. 67 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SURVEY ON INFORMATION EXTRACTION FROM CHEMICAL COMPOUND LITERATURES: TECHNIQUES AND CHALLENGES

[1,4]MUAWIA ABDELMAGID, [2]MUBARAK HIMMAT, [2,3]ALI AHMED

[1]Deanship of Scientific Research, University of Dammam, 31441 Dammam, KSA

[2]Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Malaysia

[3]Faculty of Engineering, Karary University, 12304, Khartoum, Sudan

[4]College of Engineering, Sudan University of Science &Technology, Khartoum, Sudan

E-mail: [1]muawiasadig@yahoo.com, [2]barakamub@yahoo.com, [3]alikarary@gmail.com

ABSTRACT

Chemical documents, especially those involving drug information, comprise a variety of types – the most common being journal articles, patents and theses. They typically contain large amounts of chemical information, such as PubMed-ID, activity classes and adverse or side effects. Techniques are used to extract information from a huge number of documents and it is presented in a useful structurally prepared format that can be applied to structured, semi-structured and unstructured texts. Numerous information extraction methods and techniques have been proposed and implemented. In principle, there are two main approaches to information extraction, the knowledge engineering approach and the learning approach. In this survey study, we first provide the historical background on information extraction approaches applied to chemical documents and discuss several kinds of information extraction tasks that have emerged in recent years. Then, we discuss the metrics used for evaluating information extraction systems, and finally the survey outlines the main issues that will shape future research in this area of study.

Keywords: *Information Extraction, Information Retrieval, Chemical Extraction, Extraction Methods*

1. INTRODUCTION

Information Extraction (IE) involves automatically identifying selected entity types, events or relations in free text [1]. "Information Extraction (IE) is the name given to any process which selectively structures and combines data which is found, explicitly stated or implied, in one or more texts" [2]. The IE is a process of producing unambiguous data based on natural language processing; the extracted data may be presented to users directly, stored in a database, or used in information retrieval (IR) for indexing purposes. We can summarize the main differences between IR and IE as follows. In IR the main issue or goal is to relate the relevant documents to the specific query of the users. This query is on a particular topic or a subject of interest to the user. The query relies on a specified set of parameters, keywords and terms, or on an initial set of documents. But in IE the researcher is concerned with analysing these relevant documents and then extracting information from them. In other words, IE is the process of extracting different types of information, such as entities, references, entity relations and involved events. In order to locate the type of information mentioned, the usual processing steps are as follows:

- Named Entity Recognition
- Co-reference
- Template Element Task
- Template Relation Task
- Scenario Template Task

Information Extraction begins with a collection of texts, and then converts these collected texts into readily digested and analysed information. It segregates related text fragments, extracts related information from the fragments, and finally pieces the targeted information together in a coherent form [3-5]. Information Extraction involves the method of identifying specific predefined classes of entities or relations in a text by this automated processing. Also, IE is used in a summarization process to generate short summaries of documents [6].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

To encourage the development of new methods of IE, the Defence Advanced Research Projects Agency initiated the Message Understanding Conferences (MUC). The MUCs were first held through 1987-1998, and then the IE was taken in several directions. By the end of MUC-7, the final MUC conference that was held in 1998, we can conclude that the MUC programme had defined five tasks for IE that included Named Entity Recognition (NER), Co-reference Resolution (CO), Template Element Construction (TE), Template Relation Construction (TR) and Scenario Template production (ST). These tasks are defined to locate and classify names, identify relations between entities, add descriptive information to NE results using CO, identify relations between TE entities, and fit TE and TR results in specified events, respectively. We noticed that each task has a different role to play. They can be combined or they can be selected independently depending on different criteria related to the field or the domain to be focused on. In other words, IE activity can be customized by choosing and developing tasks according to the application's needs.

As we separate the information extraction activity into several tasks we reveal a degree of complexity; but at the same time many advantages behind this separation can be seen, such as:

- Possibility to apply different algorithms and techniques to each task according to the objective of a particular application.
- Ease of debugging, since each task is separated from others.

Generally, IE systems are knowledge-intensive in terms of building them and they are more computationally intensive; therefore, extraction of the chemical data that is surveyed in this paper requires broad knowledge in the chemical compounds domain.

This survey presents the latest methods and techniques applied in IE generally and focuses on those methods and techniques used to extract information about chemical compounds and drugs. The rest of this paper is organized as follows: In the next section we provide a review of the latest works that have been undertaken to extract information from different chemical documents and other sources. In Section 3 we discuss methods and techniques used to extract information and different types or components of information. Learning and Knowledge Engineering approaches are discussed in Section 4 and in Section 5 we introduce the main ideas on Ontology-Based Information Extraction. In Section 6 we discuss the methods used in Evaluating Information Extraction Systems. Future Directions and Challenges are illustrated in Section 7, and finally the paper is concluded in Section 8.

2. RELATED WORKS ON IE FROM DIFFERENT CHEMICAL SOURCES

It is easier to find too many documents on a chemical topic than to find the right information in these documents. The scientist still has to read large collections of research papers, theses, books and patents in order to extract the embedded chemical structures. Therefore, it is essential to adopt and implement automatic extraction systems. In spite of the fact that many existing methods have been developed for this purpose, it is still considered as a hot area of interest. Furthermore, the rapid growth of publications leads to the complication of maintaining up-to-date data sets.

The process of locating and extracting knowledge and information automatically from text data has become one of the most important and active fields of database study. Chemical documents such as articles, patient's documents, newspapers, prescriptions, and thesis have considerable value in terms of information. The IE system in chemical data is concerned with automatic extraction of different types of information from the mentioned documents. Many IE methods have been applied in several fields, such as bioinformatics, and many useful studies have been published in this area of research. The rapid development of natural language processing, and its applications, helps in involving the extraction method in many fields. In the area of chemical extraction, many works have been proposed. In the early 80s [7] proposed a technique that is used to extract facts about chemical reactions from the texts of primary journals of the American Chemical Society (ACS).

Mani and Zhang (2003) proposed and discussed a method of extracting protein names from biological literature, and they were studying the effectiveness of using nearest neighbour k(NN) and other different methods [8]. Also, there are several works that addressed the challenges of IE in chemical documents [9]. Other work proposed an extraction system through text mining approaches. Abulaish and Dey (2007), in their proposed system, use text mining to extract all frequently occurring biological relations among a pair of biological concepts, then they identify the relevant information by using their extraction system [10]. In their research project UIMA-HPC [11] proposed

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

an IE automated system that speeds-up the process of knowledge mining when dealing with patents; they used thousands of documents and images and processed them in 'patroller'. In addition, an IE software system was developed by [12]; his system is used to extract chemical reactions from chemical texts. Their devolved system relies on the output of Chemical Tagger; the system has a tool for tagging words and identifying phrases of importance in experimental chemistry texts. Tharatipyakul et al. (2012) developed a chemical information-extraction system (ChemEx). Their system works on the text and also on images in publications. They developed a text annotator that is able to extract compound, organism and assay entities from text document content [13]. Many methods have also been used to extract protein interaction relations from literatures. Li et al. (2014) proposed an IE system that extracts information from biomedical literature documents. The system is concerned with integrating the semantic information of biomedical literatures into multiple kernels for protein-protein IE [14]. Recently the biomedical literature has become an important source that has useful information on a range of chemical compound data. However, these documents provide a large amount of information on the compound nomenclatures and representations and these data of chemical entities are sometimes ambiguous. Many systems are proposed for gene and protein entity recognition, and few were proposed for chemical entities. This occurred due to the lack of a corpus to train named entity recognition systems and perform evaluation as mentioned by **[15].**

3. METHODS USED TO EXTRACT DIFFERENT TYPES OF INFORMATION

As mentioned in the previous section, any IE system used to extract valuable information or knowledge from texts may include the following different types, entities, references, relations between entities and events involving the entities [16].

Biomedical text mining has become very important; therefore, biomedical named entity recognition systems and opinion extraction have become highly attractive [17]. Most biomedical named entities were focused on abstracts, but recently the development of named entities that deal with full papers has become a popular area of research. Early works considered ENR as a method for recognizing proper names such as names of persons, locations and organizations or a set of extraction patterns [18]. Then subtypes were generated; for instance, in MUC-6, 'location' was divided into subtypes such as city, state and country.

A previous study was conducted to compare two different biomedical named entities. The comparison system was based on abstracts and full paper corpora using training data that generated automatically and no features or resources of the biomedical domain were considered in the mentioned study. The results showed that bootstrapping using annotated abstracts that are automatically generated is efficient even when evaluating a full paper. In addition, as a future work for reducing the noise, the study was generated important interest in combining dictionary-based matching with the existing NER system [19].

Information extraction uses many aspects of document structure and content. Simple and important examples include commonly used words and phrases (entities) that identify instances of essential concepts. In chemistry these include:

- Bibliographic components (authors, journals)
- Molecular identity (name, connection table, synonym)
- Properties (units, physical properties, colours, form/nature)
- Procedures (solvents, amounts, colours, reagents, techniques)
- Instruments (manufacturer specification)

4. LEARNING AND KNOWLEDGE ENGINEERING APPROACHES

In the knowledge engineering approach grammar and extraction rules are constructed by hand. Also, the domain patterns are discovered by human experts through introspection and inspection of a corpus. At an earlier MUC conference several IE systems were proposed and prepared manually to provide and define a range of tasks, such as FASTUS [20], GE NLTOOLSET [21], PLUM [22] and PROTEUS [23]. Knowledge engineering has many disadvantages and it is a very laborious development process and the required expertise may not be available.

Due to the difficulty of generating extraction rules manually by knowledge engineers, and the fact that it is time consuming, information extraction research was initiated and tends towards automating this task. The automated tasks are applied by two approaches. First is supervised learning; in this approach firstly the training data is required and it must consist of large data. This

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

training data is used to learn the rules automatically based on machine learning techniques. The second approach is unsupervised learning, and this approach is a little different from the first one where the bootstrapping methods are used by a small set of seed rules and an annotated corpus [24]. In the past ten years, many hybrid machine learning approaches have been used to extract information from documents of free text [25]. Recently, machine learning and rule-based and hybrid systems are used to extract information from clinical text documents [26].

5. ONTOLOGY-BASED INFORMATION EXTRACTION

Ontology-based Information Extraction is considered as one of the important subfields of IE. Ontology is the explicit representation of the meaning of terms in vocabularies and their interrelationships among them. In their studies, [27, 28] defined ontology as formal and explicit specification of a shared conceptualization. Since ontology is dedicated to a certain domain and specifying its concepts, it would be useful in information extraction considering the fact that IE is mainly based on extracting information from a particular domain. In spite of the fact that the term ontology-based information extraction has only been recently conceived [29], some works in this field had been undertaken previously, such as constructing ontologies from text as published by [30-32].

Generally, many recent works are on-going related to this field, and many workshops have been recently organized [33]. Recently, this year, there are many new works on ontologies based on IE [34, 35]. Also, there are many works that agree with the growth of information retrieval in the near future and recently [33, 34, 36-38]. From the previous proposed works we can say that the process of discovering and extracting the knowledge and information automatically from text data has become one of the most important and active fields of database study and it has been applied in several fields, such as bioinformatics. Recently, many works have been published and it takes advantage of the fast development of natural language processing and its applications.

6. EVALUATING INFORMATION EXTRACTION SYSTEMS

Information extraction or the automatic method of retrieving information from a text or document is considered very useful when the extracted information is accurate and reliable. Most IE works are tested based on two common tasks performed by human experts who extract these chemical or other data from documents (journal articles, papers, magazines etc.). Then they attempt to create an automatic categorization of the papers according to the types of chemical data that they contain. The second task focuses on retrieving sentences containing a specific type of chemical or specific data from different documents and then compares them to their system results. Most of these works use precision as the evaluation method. Most evaluation system results rely on automated language processing techniques against human performance of extraction on the same real document. In a message understanding conference they detail the primarily evaluation measure recall and precision as follows [39].

$$\text{Recall} = N_{correct} / N_{key}$$

$$\text{Precision} = N_{correct} / (N_{correct} + N_{incorrect})$$

Precision: is one of the common evaluation rules in IE and it is defined as the ratio of the statements extracted correctly to the total number of extracted statements.

Recall: is defined in IR as the fraction of the documents that represent those relevant to the query that are successfully retrieved.

7. FUTURE DIRECTIONS AND CHALLENGES

There are many challenges facing IE in general and the chemical field specifically. On the other hand, the factor that plays an important role in complicating the extraction process is the extraction of text from images. Recent studies in the field of image processing showed a great amount of interest in content retrieval from images and videos, since the locating of text from complex background images is still difficult and challenging [40].

Machine learning has introduced advances in the use of IE; but still some challenges remain open for the new themed issue. Many tasks that have been performed by IE systems involve a strong knowledge dependency in the thematic domains that are considered most frequently. Many of the tasks performed by a traditional IE system (especially those that relate to templates) have a strong dependency on knowledge about the thematic domain, which is very frequently dispersed among the various tasks. All of the ontology-based information extraction systems are proposed to avoid and reduce these problems by the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

use of ontologies. Ontology is to provide all the means that perform very well in an IE system by using the domain knowledge that is required for the operation of IE [41]. It is extremely difficult to deal with biomedical terminology because effective standards have not been introduced in the field [9]. This will motivate researchers to work hard in this directions in order to enhance the automation process of annotation systems. On the other hand, the factor that plays an important role in complicating the extraction process is the extraction of text from images.

8. CONCLUSION

With the huge growth in the number of documents, several works and systems are proposed in IE to automate the IE methods and techniques and make them more flexible. This study introduced the general concepts of information extraction and focused on methods and techniques used to extract information from different sources of chemical literatures. This study also illustrates the works that have been undertaken in chemical information extraction, the future methods direction and the useful ideas that will be used to extract chemical information. Finally, this survey also introduced the two basic measures or metrics used to evaluated the extraction methods and techniques.

ACKNOWLEDGMENTS

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).

REFRENCES:

[1] Appelt, D.E., et al. *FASTUS: A finite-state processor for information extraction from real-world text*. in *IJCAI*. 1993.

[2] Cowie, J. and W. Lehnert, *Information extraction.* Communications of the ACM, 1996. **39**(1): p. 80-91.

[3] Muslea, I. *Extraction patterns for information extraction tasks: A survey*. in *The AAAI-99 Workshop on Machine Learning for Information Extraction*. 1999.

[4] Frakes, W.B. and R. Baeza-Yates, *Information retrieval: data structures and algorithms.* 1992.

[5] Baeza-Yates, R. and B. Ribeiro-Neto, *Modern information retrieval*. Vol. 463. 1999: ACM press New York.

[6] Amitay, E. and C. Paris. *Automatically summarising web sites: is there a way around it?* in *Proceedings of the ninth international conference on Information and knowledge management*. 2000. ACM.

[7] Zamora, E.M. 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, 1984. **24**(3): p. 176-181.

[8] Mani, I. 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] Townsend, J.A., et al., *Chemical documents: machine understanding and automated information extraction.* Organic & biomolecular chemistry, 2004. **2**(22): p. 3294-3300.

[10] Abulaish, M. and L. Dey, *Biological relation extraction and query answering from medline abstracts using ontology-based text mining.* Data & Knowledge Engineering, 2007. **61**(2): p. 228-262.

[11] Bergmann, S., et al., *Information extraction from chemical patents.* Computer Science, 2012. **13**(2): p. 21-32.

[12] Lowe, D.M., *Extraction of chemical structures and reactions from the literature*, 2012, University of Cambridge.

[13] Tharatipyakul, A., et al., *ChemEx: information extraction system for chemical data curation.* BMC bioinformatics, 2012. **13**(Suppl 17): p. S9.

[14] Li, L., et al., *Integrating Semantic Information into Multiple Kernels for Protein-Protein Interaction Extraction from Biomedical Literatures.* PloS one, 2014. **9**(3): p. e91898.

[15] Grego, T., et al., *Identification of chemical entities in patent documents*, in *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living*. 2009, Springer. p. 942-949.

[16] Tombros, A. and M. Sanderson. *Advantages of query biased summaries in information retrieval*. in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 1998. ACM.

[17] Ku, L.-W., Y.-T. Liang, and H.-H. Chen. *Opinion Extraction, Summarization and Tracking in News and Blog Corpora*. in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. 2006.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[18] Jung, K., K. In Kim, and A. K Jain, *Text information extraction in images and video: a survey.* Pattern recognition, 2004. **37**(5): p. 977-997.

[19] Vlachos, A. *Evaluating and combining biomedical named entity recognition systems*. in *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*. 2007. Association for Computational Linguistics.

[20] Hobbs, J.R., et al. *SRI International: Description of the FASTUS system used for MUC-4*. in *Proceedings of the 4th conference on Message understanding*. 1992. Association for Computational Linguistics.

[21] Krupka, G.R. *SRA: Description of the SRA system as used for MUC-6*. in *Proceedings of the 6th conference on Message understanding*. 1995. Association for Computational Linguistics.

[22] Ayuso, D., et al. *BBN: Description of the PLUM System as Used for MUC-4*. in *Proceedings of the 4th conference on Message understanding*. 1992. Association for Computational Linguistics.

[23] Grishman, R., S. Huttunen, and R. Yangarber, *Information extraction for enhanced access to disease outbreak reports.* Journal of Biomedical Informatics, 2002. **35**(4): p. 236-246.

[24] Kaiser, K., C. Akkaya, and S. Miksch, *How can information extraction ease formalizing treatment processes in clinical practice guidelines?: A method and its evaluation.* Artificial intelligence in medicine, 2007. **39**(2): p. 151-163.

[25] Neumann, G., *A hybrid machine learning approach for information extraction from free text*, in *From Data and Information Analysis to Knowledge Engineering*. 2006, Springer. p. 390-397.

[26] Uzuner, Ö., I. Solti, and E. Cadag, *Extracting medication information from clinical text.* Journal of the American Medical Informatics Association, 2010. **17**(5): p. 514-518.

[27] Gruber, T.R., *A translation approach to portable ontology specifications.* Knowledge acquisition, 1993. **5**(2): p. 199-220.

[28] Studer, R., V.R. Benjamins, and D. Fensel, *Knowledge engineering: principles and methods.* Data & knowledge engineering, 1998. **25**(1): p. 161-197.

[29] Metkewar, P., S. Mapari, and S. Naik, *Conceptual model of Ontology in Education Era.*

[30] Hwang, C.H. *Incompletely and Imprecisely Speaking: Using Dynamic Ontologies for Representing and Retrieving Information*. in *KRDB*. 1999. Citeseer.

[31] Maedche, A., G. Neumann, and S. Staab, *Bootstrapping an ontology-based information extraction system*, in *Intelligent exploration of the web*. 2003, Springer. p. 345-359.

[32] Saggion, H., et al., *Ontology-based information extraction for business intelligence*, in *The Semantic Web*. 2007, Springer. p. 843-856.

[33] Wimalasuriya, D.C. and D. Dou, *Ontology-based information extraction: An introduction and a survey of current approaches.* Journal of Information Science, 2010. **36**(3): p. 306-323.

[34] Shah, R. and S. Jain, *Ontology-based Information Extraction: An Overview and a Study of different Approaches.* International Journal of Computer Applications, 2014. **87**.

[35] Tao, J., A.V. Deokar, and O.F. El-Gayar. *An Ontology-Based Information Extraction (OBIE) Framework for Analyzing Initial Public Offering (IPO) Prospectus*. in *System Sciences (HICSS), 2014 47th Hawaii International Conference on*. 2014. IEEE.

[36] Rastegar-Mojarad, M., B. Harrington, and S.M. Belknap. *Automatic detection of drug interaction mismatches in package inserts*. in *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*. 2013. IEEE.

[37] Thakare, P.D., et al., *A REVIEW ON INFORMATION REPRESENTAION AND RETRIVAL IN SEMANTIC WEB.* 2013.

[38] Holey, P.A. and S. Prabhune, *Review of Content-based Recommendation System.*

[39] Sundheim, B.M. *Overview of the third message understanding evaluation and conference*. in *Proceedings of the 3rd conference on Message understanding*. 1991. Association for Computational Linguistics.

[40] Giri, P.S., *Text Information Extraction And Analysis From Images Using Digital Image Processing Techniques.* 2013.

[41] Karkaletsis, V., et al., *Ontology based information extraction from text*, in *Knowledge-driven multimedia information extraction and ontology evolution*. 2011, Springer. p. 89-109.