Zaghloul and Trimi *International Journal of Quality Innovation* (2017) 3:3
DOI 10.1186/s40887-017-0012-y

International Journal of Quality Innovation

THEORETICAL ARTICLE Open Access

Developing an innovative entity extraction method for unstructured data

Waleed Zaghloul[1] and Silvana Trimi[2*]

* Correspondence: silvana@unl.edu
[2]Department of Supply Chain Management and Analytics, College of Business Administration, University of Nebraska, Lincoln, USA
Full list of author information is available at the end of the article

Abstract

The main goal of this study is to build high-precision extractors for entities such as Person and Organization as a good initial seed that can be used for training and learning in machine-learning systems, for the same categories, other categories, and across domains, languages, and applications. The improvement of entities extraction precision also increases the relationships extraction precision, which is particularly important in certain domains (such as intelligence systems, social networking, genetic studies, healthcare, etc.). These increases in precision improve the end users' experience quality in using the extraction system because it lowers the time that users spend for training the system and correcting outputs, focusing more on analyzing the information extracted to make better data-driven decisions.

Keywords: Entity extraction, Machine learning, Precision of extraction, Text analytics, Natural language processing

Background

In today's networked global world, people, goods, data, and knowledge move more widely, quickly, and freely in the speed of the light across the various boundaries. For businesses, governments, and scientific communities, this new environment has brought tremendous opportunities, as well as challenges [1]. Not only are countries' economies more connected and interdependent but also the people, governments, politics, and knowledge [2]. The advances in information and communication technology (ICT) have brought a tremendous increase in the amount of data created and shared (big data), techniques, technologies, and systems to extract value from the data. Data analytics are used for a variety of purposes (business, security and safety, scientific discovery, etc.), domains (biology, medicine, education, etc.), and stakeholders (businesses, governments, scientists, and consumers) [3]. Therefore, extracting information and value from data has become critical for academia, the industry, and governments.

Many institutions and organizations are increasingly gathering intelligence by processing and analyzing massive amounts of data that is in textual format and collected from multiple sources and languages. Processing and analyzing such data, which very often are imperfect, incomplete, and unstructured, have become increasingly difficult. Thus, development of new intelligence and analytics (I&A) technologies and improvement of the quality of data processing and analytics have become the focus of

governments, mathematicians, computer scientists, and data analysts. Methodologies, techniques, and practices of extracting value from data, i.e., I&A, have changed with the changes in types of data collected and analyzed. I&A 1.0 analyzed structured content; I&A 2.0 analyzes unstructured (text-based) content; and I&A 3.0, which is in the early stages, focuses on mobile and sensor-based content analysis [4].

The emerging areas for text analytics are (1) information extraction (IE)—automatically extracting structured information from documents; (2) topic model (TM)—discovering the main themes in a large and unstructured collection of documents by using algorithms; (3) opinion mining—access, extract, classify, and understand the opinions expressed in many sources including social networks; sentiment analysis is also used for opinion mining; and (4) question answering (Q&A)—answering factual questions (e.g., IBM's Watson, Apple's Siri, Amazon's Alexa, etc.) based on techniques from statistical natural language processing (NLP), information retrieval (IR), and human-computer interaction (HCI) [4].

The purpose of this study is to add to the current literature on the performance of named entity recognition (NER), the building block for IE systems. Specifically, we build an entity extractor for categories Person and Organization, which are entities that have a finite set of identifiers. The proposed extracting method improves (1) the extraction quality of the system by increasing the precision of entity extraction as a result of the initial extracted entities from our method; our highly precised entities will be used as a seed for training other machine learning (ML) systems, across domains and languages, thus, not only eliminating the need for manual training but also will expand the seed (through learning) and therefore will continue to increase the precision and use of entity extraction across domains and languages; and (2) the experts' experience and use: experts will spend less time for training the system (as is done by the seed entities created by us, and expended by ML) and also less time on fixing faulty entities and relationships extracted by the systems, thus, instead, will spend their time on using the system's outcome to make better data-driven decision.

This paper is organized as follows: the "Literature review" section presents a review of relevant literature for the theoretical concepts of the study; the "The proposed method" section explains the proposed method for entity extraction; the "Discussion" section provides and discusses the study results; and the "Conclusions" section concludes the study by summarizing the study's contributions, limitations, and future research needs.

Literature review

Entity extraction

Natural language processing (NLP), which is the "understanding" of the natural human language by computers, involves machine translation, information retrieval, and question answering. They are becoming increasingly critical in a variety of applications such as machine reading and understanding, intelligence analysis, social media analysis, etc. [5, 6]. The diversity of domains which rely on NLP (e.g., news media, law, biomedicine, pharmaceutical/pharmacogenomics, chemistry, etc. [7–12]) is growing and so is the variety of languages (other than English). The long-term goal of NLP is to have algorithms capable of automatically reading and obtaining knowledge from the text [13].

Advanced NLP applications rely on entity and relationship extraction and on machine learning (ML). The most common form of hidden information in the text is in the form of entities: names, places, dates, and other words and phrases that give the meaning in a text. Most commonly used forms of extractions are entity extraction and their relationship or association extraction. Extraction systems are used to identify elements in a text document that belong to predefined categories of entities and to extract relationships or associations among/between entities [14]. The main method for entity extraction is named entity recognition (NER): automatically identifying names in text and classifying them into predefined set of categories [13, 15, 16]. The most widely used categories are Person, Organization, Location, and Date. As domains of applications have grown, so have the entity categories, with such newer categories as Time, Facility, Equipment, Weapon, Animal, Plant, Medicine, Protein, and Gene, among many others [16]. Relationship extraction is the process of identifying two entities that are associated together within a text document. Co-references (and links) are used to detect and extract relationships between and among entities. The analysis of such relationships is important for a variety of different applications including machine reading and understanding, intelligence analysis, and social media analysis.

NER approaches can be grouped into rule-based and statistical approaches, and more recently is a combination of both (hybrid NER). The "rule-based" hand-written approaches are the earliest form of NER. Rule-based systems define a set of rules that would determine the presence of an entity and its classification. The rules can be grammar-based, gazetteers of personal and company names, and higher level based, such as name co-reference [16, 17]. Ontologies are also used to represent a group of related independent categories. The rule-based systems are particularly useful for categories that have highly specialized entities (e.g., biomedical) and a finite number of members. Their performance is determined by the quality of the rules [18–20].

Machine learning

In the earliest form of ML, rules were used to produce decision trees needed to build algorithms that extracted named entities in unstructured texts. Nowadays, ML uses algorithms that are designed to allow a computer to learn from statistical regularities or other patterns found in data. Statistical models are built based on a large training set of documents, *corpora*, which are used to "supervise" the learning of the classification process. Algorithms learn and adjust rules through real world data. Learning can be accomplished in a fully supervised, semi-supervised, or unsupervised manner [15]. In supervised ML, to classify entities, a model is trained on an annotated (manually tagged by human experts) set of documents (*corpus*). The semi-supervised ML is a hybrid system that uses a combination of annotated and non-annotated data, machine learning, and rule-based approach [6]. To improve the training process, human experts are selectively given training examples to label while other training examples are automatically labeled by the machines [19, 21]. ML algorithms that are fully automated learn rules in an unsupervised way from data that has not been pre hand-annotated.

Machine-based entity extraction works well as long as the classifications are correct. Thus, precision is very important, and building ML-based extraction systems/methods with high precision is challenging. The supervised ML requires a large number of (manually or

automatically tagged) training documents. Thus, they are expensive—developing training examples require the input of human experts as automatically generated examples are not always very accurate. Increasing the size of the training set improves the system's precision, but it is also expensive because of experts' time [21]. Manually tagging text documents is a very slow and tedious process; tagging enough documents that could be used to train an entity extraction system could easily take months and involve a large number of human experts' and wastes' valuable time for system users. Thus, the need for large annotated corpus makes the supervised ML systems expensive and often times impractical. Even though the supervised ML methods currently have the best performance, they are mostly focused heavily on English language, with very few other languages (such as Dutch, Arabic, etc.). In addition, many of the NER corpora are from a single genre (e.g., newswire), which is an issue when it comes to their robustness and generalizability in terms of their wide usage and the level of performance in other domains.

On the other hand, the unsupervised or semi-supervised NER approaches do not need a large set of annotated corpora. In an unsupervised approach, entities are clustered based on the context or on entities' simultaneous occurrences (co-occurrence) in the source being analyzed [16]. However, as of now, these systems are not very accurate and many entities are missed.

In a semi-supervised NER, a small starter supervised set of seeds/rules is used, which is expanded as the system is used and therefore learning occurs. Tagging fewer documents for training decreases the tagging time, without lowering the precision, as the system keeps expending and perfecting through ML and while it is used, in time, across domains, and different languages. This leaves the semi-supervised methods to be the best option for NER. Therefore, our proposed method utilizes it, to create a high-precision initial training seed that will be utilized, while being improved through ML, for semi-supervised training in cross-domain, and multiple languages systems. In our proposed method, we create a high-precision initial seeds of entities for training which improves several of the abovementioned inadequacies of the supervised and unsupervised methods: eliminate the need for manually tagged documents for training (no experts' time needed); can be used across multiple domains and languages, thus, increases its generalizability; and by being used across domains, languages, and in time, it will keep improving its extraction quality as a result of increase in recall (categories will keep expanding) and precision (as initial seeds are highly precised, and method focuses on precision).

Methods

The primary motivations for this study are to (1) improve the precision (eliminate noise) of identifying elements that belong to the two entities, Person and Organization; (2) eliminate the need for tagged documents used in training systems to identify entities of those two categories; this decreases experts' time for both training the system and for evaluating the accuracy of extraction outcome (entities and relationships), thus instead allowing them spend more time in analyzing the system's outcome and decision making; (3) create a high-precision seed entity list (for each of the two categories) that can be used to train ML systems over time and across multiple domains and languages.

The main goal of this study was to increase the precision of extracting elements of certain categories. In this paper, we focused on only two of those, Person and

Organization. The target documents for this NER system are news reports, and the target audience is the intelligence analyst group tasked with analyzing these documents and making decisions accordingly. Analyzing the relationships and generating a graph that represent the relationships of interest (persons, organizations, etc.) among identified elements from the NER system is particularly important to the analyst. Hence, precision of entity extractors is extremely important, as wrong extracted entities will produce graphs that are highly cluttered and inaccurate, and therefore, highly undesirable for analysts. The presence of noisy data will require analysts to spend considerable amount of time in reviewing the system's generated graphs to determine the useful nodes from the noisy ones, and manually fixing the problems and reconstructing the graph. This manual process of graphing, which is required to be repeated to make sure that there is no noise in nodes and fix their relationships, could render the automated system useless and cumbersome in the long term. Thus, focusing in the accuracy, ensuring that all entity extractors used have very high precision that would result in relationship graphs that are precise and useful, is more important than some missing nodes due to the decrease of the recall while attempting to maximize precision.

Our method focuses on only Person and Organization as entity types for extraction because they are the most widely used entity types. Moreover, these selected entity types usually do have detectable cursors such as honorifics for Person entities. Thus, instead of looking for every person or organization in the text, our method simply looks for types or categories of persons or organizations respectively. Since the documents that our system would analyze are formal reports, many details are usually present in the text. For example, Persons are usually identified with their proper salutations (such as Mr., Dr., General, etc.) or positions (President, Prime Minister, Secretary of Defense, etc.). The rule-based system searches for such indicators to find the appropriate type of entity. For identifying Organizations, organization types like banks, companies, corporations, and universities are sought in the text.

The steps we used were (1) we built a rule-based independent extractor for each of the two entities; (2) we used our entity extractors on a set of 52,000 documents and created a highly reliable entity types seed, which can be used for training other ML-based entity extraction systems; (3) we measured the precision of our extraction systems and were very satisfied with the results.

Results

Since the target users for our proposed method are the news/intelligence analysts who seek more automated analysis of documents, we used 52,000 unclassified training news reports and stories, similar in format with the real documents analyzed by analysts' NER systems (unstructured text, and in a variety of lengths, from a few pages to over 200 pages in some cases).

First, we built a rule-based system to classify entities in documents (52,000 of them) based on the predefined categories. Two independent extractors were built (with the aim of maximizing precision) to extract entities of the types Person and Organization. Each extractor extracted entities and some metadata needed for further processing. First, we extracted entities by running our two extractors on the document set (52,000).

Second, we measured the proposed method's *precision*—exactness or the *quality* of retrieved instances (retrieved instances that are relevant). To determine if each of the entities obtained by the NER system was actually of the correct classification, a sample of the appropriate size (Table 1) was randomly taken from the extracted entities of each of the two categories. The appropriate sample size was determined statistically, based on a confidence level of 95% and a confidence interval of 4. These samples were manually analyzed to determine if each of the entities obtained by the systems was actually correctly classified. A closer look at the erroneously identified examples showed the possibility of some improvements in the implementation of the algorithm, which would result in a further improvement in precision. However, this is beyond the aim of this paper and therefore, we did not do it. The precision results were very satisfactory: extraction precision for the Person category was 98.1%, and for the Organization category 97.5% (Table 1).

The recall value for each particular category (for example, Person) is used to measure the sensitivity of NER—the completeness or the *quantity* of relevant instances retrieved (the percentage of all extracted entities over all actually existing entities in the corpus being analyzed). Considering that we have two entities, and a large number of documents (over 50,000), it would be a challenge to determine the exact recall value for the two of them. We could estimate recall to help provide an idea about the F measure. One possible approach to estimate the recall value is to take a random sample of documents. In our method, we determined the precision by utilizing a large representative sample (determined statistically) of extracted entities which were drawn from a large size of documents. Manually tagging entities of all two types in such large set of documents and comparing these entities to those found automatically by the extractor would require a lot of work. Since this will only calculate an estimate of recall, and it was not the main goal of this study (we focused on improving precision), we did not measure the recall.

Discussion

The main goal of our study was to build very high-precision entity extractors for the Person and Organization categories that would minimize the noisy output (entities and their relationships). We used the two specific categories, Person and Organization, they are among the top most heavily utilized categories in information retrieval systems across domains, thus they can be used to further improve the NER system's precision and expand its scope through machine learning.

First, our method improved the precision of entity extraction of two categories and created a good initial seed that can be used for machine learning in the future. The

Table 1 Precision testing of extraction systems

Test results	Entity	
	Person	Organization
Entities found (by system)	45,487	115,967
Unique entities found	14,851	22,820
Testing (random) sample size	577	585
Correct entities identified (by system)	570	570
NER precision	98.4%	97.5%

initial seed, not only has high extraction precision, but because it was created based on a large set of documents (52,000), it can be utilized as a highly reliable training set for ML algorithms across domains, thus eliminating the need for manually tagged training documents. ML-based extractors are trained to classify an extracted entity into one of many predefined categories; misclassification error however can be considerable. Utilizing specialized identifier lists (seed) that help extract entities of a certain type greatly minimizes the misclassification of entities. Our proposed method ensures that each extractor is specialized in one and only one category: if the rules of a specific classifier do not recognize an entity, it will be ignored and not extracted at all instead of being misclassified. Even though missing a considerable number of potential entities will lower the extractor recall, we have designed (not part of this paper) a second step for ML that will counteract this weakness and improve recall (while not affecting precision).

Furthermore, these categories (e.g., Person and Organization) are general enough so that the extractors can be used and continuously improve entity extractions on the two aforementioned categories across domains and applications. For example, identifying a hospital in an intelligence report is similar to identifying a hospital in a medical report. After an initial investment into building a good seed list and training the ML models, the extractors can become stable over time and be useful in multiple domains. Cross-domain ML will further refine (increase precision and recall) our entity extractors, by perfecting its rules and expanding the entity seed for the two involved categories. The improvement in entity extraction quality will also increase the precision of entity relationship extraction. Noisy data becomes less of a roadblock. The end users of the extractor systems will be spending less time in evaluating the quality of the extracted entities and relationships and spend more time instead in analyzing the information retrieved from the system and make better data-driven decisions. Relationship extraction accuracy is particularly important in certain domains, such as intelligence systems, social networking, genetic studies, healthcare, and the like.

Our method is simple enough that it can be used for different languages other than English, especially Germanic languages like German and Dutch, and Romance languages like Spanish, French, and Italian. Lastly, since our method does not require ML or training of any sort in the first stage, it can be applied across application areas without the need for any major changes. However, building separate, statistically based models for each application area would be needed.

Conclusions

Contributions

The greatest benefit of our proposed method is that it creates high-precision entity extractors for the Person and Organization categories. These extracted entities will make a highly reliable training set for machine learning algorithms to learn the extraction rules for the two categories, and thus improve the extraction quality (F measure) of all two extractors. The number of documents needed for training is usually very large and would require hundreds of hours from each analyst involved. Specialists' time is not only very expensive but also often difficult to find. Thus, eliminating specialists' time to train systems not only results in huge cost savings and but also increases specialists' efficiency as they spend their time performing their job by using and

interpreting the outcome of NER systems, instead of training, checking outcomes, and correcting errors. Improvement in precision in extraction of entities improves the precision of entity relationship extraction, thus minimizing the system users' time spent on searching through graphs and fixing faulty relationships.

Another important benefit of our system is its generalizability: (1) There is no need for knowledge in any specific domain, as the two categories we applied our system to are both general in nature and widely used in many knowledge domains. Identifying a person or organization is similar across domains. Thus, there is no need for domain experts to build their own extraction system. After an initial investment in building a set of good extractors, the extractors can become stable over time through applications and be useful in multiple domains. The method can be further refined through the feedback from users. (2) The proposed method for extracting entities' type of Person and Organization can be effectively applied to any other category of entities that have a finite set of sub-types or identifiers. If the target category is not directly applicable, it is possible to generate similar initial high-precision seed entity lists (as long as the number of seed items is large enough), which can be used to train machine-learning systems to learn the extraction rules. The easiest way to obtain such seed entity lists is through the use of specialized dictionaries specifically built with a subset of the known entity set. For example, a category like Locations could benefit from the use of a specialized dictionary or a gazetteer. A seed list of locations could be extracted from the document set; then the feature surrounding each seed could be used by a ML algorithm to generate extraction rules for finding entities. (3) In our proposed method, there is no need for manually tagged training set of documents in any sort or ML in the first stage, increasing the generalizability of applications (not just domain) of the method across many areas, without the need for major modifications.

Finally, our proposed method could also be easily used in different languages that have similar features to English, especially Germanic languages (such as, German, Dutch, Danish, Norwegian, etc.) and Romance languages (such as, Spanish, French and Italian). Replacing the English list of identifiers for each category with the equivalent list from the target language would yield similar results with little to no complications.

Limitations

The main aim of our study was to improve the precision of entity extraction. The quality of our extraction system, however, depends on the quality of the determiner lists. If the categories do not have a finite number of "members," our method would not achieve similar high-precision results. Creation of such lists requires research and time and could vary from one language to another. This method could be a challenge for very large data intensive systems. It would not be a very difficult task, however, to take an English list of determiners and find the equivalent list in other languages.

Another weakness of our method is that it is not applicable to every possible entity extraction category. A category like Locations could highly benefit from the use of a specialized dictionary or a gazetteer (reference). As discussed earlier, a seed list of locations could be extracted from the document set, and then, the feature surrounding each seed could be used by ML algorithms to generate extraction rules or to find

similar entities. Such features could be based on part of speech (POS) tags, grammatical constructs, semantics, and many other possible features. The efforts to create such lists will increase the ability to utilize them across application domains, with little to no changes.

Last limitation in using our proposed method comes from the fact that, because it is aimed at maximizing precision, it could lower the recall as an indirect result. Measuring recall in a system that processes a very large number of documents is always a challenge, and therefore, recall in such systems is usually estimated. Such deficiency could be overcome by accurately measuring the recall based on data sets such as CoNLL03 [22] or other similar datasets. This recall evaluation (not just estimation) on a common data set, even though the data sets might not belong to the same domain as that of the extractor system that is being built, allows valid extraction quality comparisons across systems and methods.

Future research needs

The next logical step in the stream of this research is building appropriate learning models and training machines that utilize our method's high-precision entity extraction. Because the entities extracted using the method discussed in this paper have very high precision and are run against a decent size of document set, they will make a very good training set for ML algorithms. Different types of ML methods could be used, such as Neural Networks [23], Support Vector Machines [24], Decision Trees, Bayesian Networks, Automated Rule Construction, Linear and Extended Linear Models, Clustering or Ensemble Learning (combination of a variety of ML methods), etc. Exploring different techniques for which one of them will give the best results, and whether different techniques can capture different (or similar) sets of previously undiscovered entities, can be interesting future research projects. Utilizing entities extracted in this paper acting as a highly reliable training set for ML will be a step toward building higher quality entity extractors. Measuring how much recall and, consequently, the F measure can be improved through different machine learning can be another aim for future research.

Finally, our proposed algorithm can be applicable to Germanic and Romance languages because they have similar features to English. An interesting research direction would be to investigate if a similar method could be developed for languages with features that are different from English, such as Asian and Semitic languages.

Acknowledgements
N/A.

Funding
No funding was used for this research.

Authors' contributions
Both authors contributed equally to this paper. Both authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Valera Intelligent Systems, Fairfax, VA, USA. [2]Department of Supply Chain Management and Analytics, College of Business Administration, University of Nebraska, Lincoln, USA.

Received: 22 March 2017 Accepted: 5 May 2017
Published online: 22 May 2017

References

1. Lee SM (2015) The age of quality innovation. Int J Qual Innov 1(1):1–8
2. Lee SM, Olson D (2010) Convergenomics: strategic innovation in the convergence era. Gower Survey, UK
3. Kim GH, Trimi S, Chung JH (2014) Big data applications in the government sector: a comparative analysis among leading countries. Commun ACM 57(3):78–85
4. Chen H, Chiang RH, Storey VC (2012) Business intelligence and analytics: from big data to big impact. MIS Q 36(4): 1165–1188
5. Hossain MS, Butler P, Boedihardjo AP, Ramakrishnan N (2012) Storytelling in entity networks to support intelligence analysts. In: KDD '12: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining., pp 1375–1383
6. Zayed O, El-Beltagy S, Haggag O (2013) An approach for extracting and disambiguating Arabic persons' names using clustered dictionaries and scored patterns. In: Métais E, Meziane F, Saraee M, Sugumaran V, Vadera S (eds) Natural Language Processing and Information Systems, vol 7934, NLDB 2013. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg
7. Usami Y, Cho HC, Okazaki N, Tsujii J (2011) Automatic acquisition of huge training data for bio-medical named entity recognition. In: BioNLP '11: Proceedings of Biomedical Natural Language Processing Workshop., pp 65–73
8. Habib MS, Kalita J (2010) Scalable biomedical named entity recognition: investigation of a database-supported SVM approach. Int J Bioinforma Res Appl 6(2):191–208
9. Hakenberg J, Leaman R, Ha VN, Jonnalagadda S, Sullivan R, Miller C, Tari L, Baral C, Gonzalez G (2010) Efficient extraction of protein-protein interactions from full-text articles. IEEE/ACM Trans Comput Biol Bioinform 7(3):481–494
10. Sutton N, Wojtulewicz L, Mehta N, Gonzalez G (2012) Automatic approaches for gene-drug interaction extraction from biomedical text: corpus and comparative evaluation. In: BioNLP '12: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing., pp 214–222
11. Corbett P, Copestake A (2008) Cascaded classifiers for confidence-based chemical named entity recognition. In: BioNLP '08: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing., pp 54–62
12. Dozier C, Kondadadi R, Light M, Vachher A, Veeramachaneni S, Wudali R (2010) Named entity recognition and resolution in legal text. Springer, Berlin
13. Sharda R, Delen D, Turban E (2013) Business intelligence and analytics systems for decision support. Pearson Education, New Jersey
14. Zaghouani W (2012) RENAR: a rule-based Arabic named entity recognition system. ACM Trans Asian Lang Inf Process 11(1):2:1–2:13
15. Nedeau D, Sekine S (2007) A survey of named entity recognition and classification. Linguisticae Investig 30(1):3–26
16. Desmet B, Hoste V (2013) Fine-grained Dutch named entity recognition. Lang Resour Eval 48(2):307–343
17. Gotoh Y, Renals S (2000) Information extraction from broadcast news. Philos Trans 359(1769):1295–1310
18. Petasis G, Vichot F, Wolinski F, Paliouras G, Karkaletsis V, Spyropoulos CD (2001) Using machine learning to maintain rule-based named-entity recognition and classification systems. In: ACL '01: Proceedings of the 39th Annual Meeting of Association for Computational Linguistics., pp 426–433
19. Witten IH, Frank E, Hall M (2011) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, Burlington
20. Zhou G, Su J (2003) Named entity recognition using an hmm-based chunk tagger. In: ACL '02: Proceedings of the 40th Annual Meeting of Association for Computational Linguistics., pp 473–480
21. Ekbal A, Saha S, Singh D (2012) Active machine learning technique for named entity recognition. In: ICACCI '12 Proceedings of the International Conference on Advances in Computing, Communications and Informatics., pp 180–186
22. Ratinov L, Roth D (2009) Design challenges and misconceptions in named entity recognition. In: CoNLL'09: Proceedings of the 13th Conference on Computational Natural Language Learning of Association for Computational Linguistics., pp 147–155
23. Zaghloul W, Lee SM, Trimi S (2009) Text classification: neural networks vs. support vector machines. Ind Manag Data Syst 109(5):708–717
24. Isozaki H, Kazawa H (2002) Efficient support vectors for named entity recognition. In: COLING '02: Proceedings of the 19th international conference on Computational linguistics., pp 1–7