Biomedical term extraction using fuzzy association

Bidyut Das
Haldia Institute of Technology

Mukta Majumder
University of North Bengal

Santanu Phadikar
Maulana Abul Kalam Azad University of Technology

Arif Ahmed Sekh (✉ SKARIFAHMED@GMAIL.COM)
XIM University Bhubaneswar https://orcid.org/0000-0003-0706-2565

Research Article

Keywords: Term extraction, Ngram, Fuzzy association, Natural language processing, GENIA dataset

Posted Date: July 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1826096/v1

Biomedical term extraction using fuzzy association

Bidyut Das · Mukta Majumder · Santanu Phadikar · Arif Ahmed Sekh*

Received: date / Accepted: date

Abstract Automatic term extraction from a biomedical text is a well-known problem in the area of natural language processing. It is carried out by employing four kinds of measures: linguistic and rule-based, dictionary-based, statistical, and machine learning. Automatic term extraction indicates whether or not two or more words come together in the text more often than by chance to form a biomedical term that automatically extracted using an automated system. A fuzzy set-theoretic approach is presented in this article that compares with the existing statistical measures. The experimental result shows that the fuzzy-measure offers better precision than the popular statistical-measures for extracting biomedical terms, especially when we have compared more than 60% ranked list of extracted terms.

Keywords Term extraction · Ngram · Fuzzy association · Natural language processing · GENIA dataset

Bidyut Das
Haldia Institute of Technology
Haldia, India
E-mail: bidyut2002in@gmail.com

Mukta Majumder
University of North Bengal
Siliguri, India
E-mail: mukta_jgec_it_4@yahoo.co.in

Santanu Phadikar
Maulana Abul Kalam Azad University of Technology
Kolkata, India
E-mail: sphadikar@yahoo.com

Arif Ahmed Sekh,
XIM University
Bhubaneswar, India
E-mail: skarifahmed@gmail.com
* Corresponding Author

1 Introduction

The biomedical data is increasing exponentially, and a biomedical knowledge base (KB) plays a significant role in the development of biomedical science. The information that takes the attention of biomedical researchers falls into three categories - extraction of **(a)** biomedical terms (Sandoval et al., 2019), **(b)** relations (Hong et al., 2020), and **(c)** events (Li et al., 2019) (See Fig. 1. This work mainly focuses on the first type, extraction of biomedical terms from the biomedical dataset. The biomedical data consists of clinical descriptions, event reports, health records, emails, or patient's feedback forms (Murdoch and Detsky, 2013). In the past several years, process this data automatically led to the advancement in the field of biomedical text mining (Lamurias and Couto, 2019).

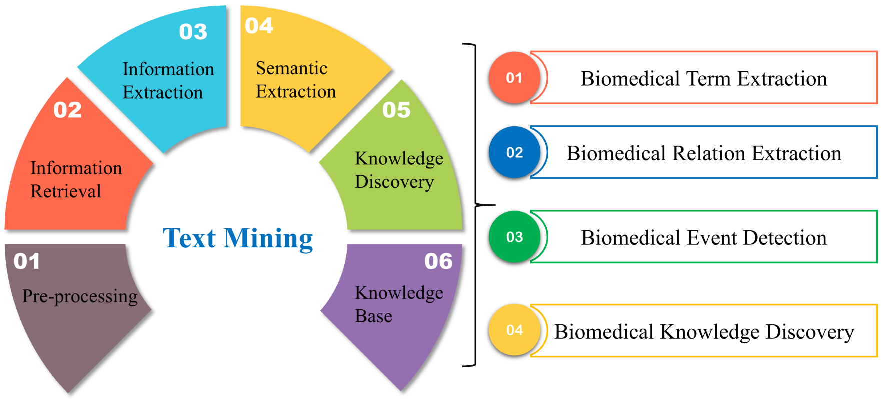

Fig. 1 Biomedical information retrieval and applications.

The identification of terms from biomedical literature is one of the challenging research topics in the last few years, both in Natural Language Processing (NLP) and biomedical communities (Herrero-Zorita et al., 2014; Samy et al., 2012). The task of manually extracting biomedical terms is time-consuming and laborious. The researchers have faced many difficulties in designing automatic methods for selecting the biomedical terms and concepts under the form of vocabularies, thesauri, terminologies, or ontologies to assist knowledge experts.

Automatic Term Extraction (ATE) involves the extraction of technical terms from domain-specific corpora (Heylen and De Hertog, 2015; da Silva Conrado et al., 2014). ATE is a task of domain knowledge retrieval because the technical terms are used for lexicon update, ontology creation, summarization, named-entity recognition, etc. (Lossio-Ventura et al., 2016). ATE is applied in several domains such as biomedical (Lossio-Ventura et al., 2014b), ecological (Conrado et al., 2013), mathematical (Stoykova and Petkova, 2012), social networks (Lossio Ventura et al., 2012), natural sciences (Dobrov and Loukachevitch, 2011), and information technology (Newman et al., 2012).

The term extraction methods are classified into four groups: (a) linguistic and rule-based, (b) dictionary-based, (c) statistical, and (d) machine learning. We propose a new fuzzy-based statistical method for biological term extraction. The method can be consider as a hybrid of statistical and linguistic system. The "terms" are helpful to get the conceptual structure of a "domain". Term contains a single word called unigram-term, or multi-words called ngram-term. The proposed study focuses on extracting the ngram biomedical terms, which are more complex than the unigram term extraction.

The objective of this paper is to present a new fuzzy-association (Martin-Bautista et al., 2004) method to extract and rank biomedical terms automatically from a biomedical dataset. The proposed approach has two parts. The first part extracts the terms having two words (bigram), and the second part shows the joining of bigrams to make trigram (three words), trigrams to quadgram (four words), and so on. Here two fuzzy sets are considered, one is for unique words, and the other is for consecutive word pairs. The membership functions of fuzzy sets are calculated using the word-occurrence knowledge in the dataset. The fuzzy membership measures are combined to calculate the fuzzy association score for extracting meaningful biomedical terms. The article's key contributions are:

- A new fuzzy association score has been proposed to extract and rank biomedical terms from a biomedical corpus.
- This approach does not restrict the number of words in a biomedical term. It automatically extracts meaningful biomedical terms in an unsupervised way.
- The method's parameters can be customized to achive desire results based on user input.

The paper is subdivided into six sections wherein related previous works are shown in section 2, the proposed methodology is presented in section 3, section 4 illustrates the preprocessing technique, experimental results are demonstrated in section 5 and section 6 depicted the conclusion.

2 Related Works

For term extraction, different authors suggested different techniques in the literature. These techniques are grouped into five clubs: linguistic and rule-based, dictionary-based, statistical, machine learning, and hybrid. A few previous works on each group are discussed below.

2.1 Linguistic and Rule-based

Rule-based methods use pattern analysis for term creation such as part-of-speech patterns, syntagmatic patterns, and grammar knowledge such as morphological analysis, lemmas, and affixes (Golik et al., 2013). In Spanish, noun phrases are used for medical terms extraction (Koza et al., 2011). In general, an efficient strategy can be obtained if the rule-based method focuses on a single language to create terms. However, this is not for all languages nor all domains (Herrero-Zorita et al., 2015).

2.2 Dictionary-based

Dictionary-based techniques use digital resources such as a list of stopwords, ontologies, glossaries, and domain thesaurus. These lists are used to filter the text and uniquely identify useful terms by eliminating not interest words. This strategy is simple and effective. But, it is inadequate, and not available for all domains. In (Segura-Bedmar et al., 2008), the UMLS meta thesaurus and name lists of other generic drugs are used to identify and classify pharmacological names in biomedical texts.

2.3 Statistical-based

Statistical methods of term extraction mainly search for sequence repetition or term frequency. The term frequency refers to the number of times a term or word sequence appears in the text of the dataset. The user specifies the frequency threshold based on the application. The strength of the statistical approach is that it does not depend on the language of the dataset; but depends on statistics, patterns, and probabilities. However, this approach has some drawbacks. In this approach, the frequency is the main factor for determining a unit as a term. But not all repeated words are terms, and not all terms repeated in a text. Sometimes, a stop list is used to deal with this problem. A stop list is a list of items that are ignored as a term. The two popular statistical techniques, log-likelihood-ratio and log-dice metric and other statistical technique such as mutual information (MI) or distributional semantics or lexical collocation is used for extracting biomedical terms (Gelbukh et al., 2010; Lossio-Ventura et al., 2013).

2.4 Machine Learning

Machine Learning is a special type of method that uses statistical techniques and consists of training algorithms with a dataset that previously annotated by human experts. Machine Learning algorithms (decision-trees, hidden-markov-model, or support vector machine) are trained using annotated terms and are applied to a test dataset to identify new terms. The classifier divides all terms in the text between true and false terms. Lastly, advanced neural network research generates encouraging ways for detecting terms using sequence modeling such as part-of-speech or named-entity recognition. Biomedical term recognition methods also use Recurrent Neural Network models (Long-Short-Term Memory) (Lyu et al., 2017) and hybrid approaches (merging with Conditional Random Fields) (Lample et al., 2016), attention mechanisms, and language modeling (Rei, 2017). These methods use vector representation of words with their frequency distribution (word embedding) (Pennington et al., 2014). The main drawback of such method is that it demands a large volume of training data set (annotated by human experts) that is difficult in many cases.

2.5 Hybrid

Hybrid approaches combine two or more approaches. In most cases, a dictionary-based or rule-based approach is combined with a statistical approach. Zehtab-Salmasi et al. (2021) presented a hybrid term extraction method called FRAKE. The FRAKE method fuses two approaches, graph and textural features to extract useful terms. Perez-Guadarrama et al. (2018) proposed a new unsupervised fuzzy clustering method for term extraction. It fuses fuzzy logic with machine learning approach. Piskorski et al. (2021) proposed an unsupervised and linguistic unsophisticated term-extraction algorithm. It combined statistic, graph, and embedding-based approaches.

3 Proposed Method

In this paper, we propose a new association technique for finding biological terms using a fuzzy set-theoretic approach (Torres and Nieto, 2006). It fuses fuzzy-logic and statistical association methods and outperforms state-of-the-art association methods.

The term is just a word or set of words occurring together and having a relevant meaning in the dataset (Periñán-Pascual, 2018; Lossio-Ventura et al., 2014a). The simplest technique of extracting bigram terms are frequency count of word-pair in the dataset. A collection of two or more consecutive words is extracted as a term when the joint frequency of words is high compared to the frequency of individual words. Most of the researchers in literature modeled this idea as the ratio of joint frequency to individual frequency, explicitly or implicitly. The concept of high frequency is imprecise or vague; because it depends on various factors such as dataset size, word combinations, or appearances of other words in the dataset. The fuzzy logic is applied when the actual or boundary value is not clear such as cold, hot, young, old, low, or high are belongs to the fuzzy concepts. The fuzzy set theory manages such impreciseness for describing the notion of high frequency.

We have proposed two fuzzy sets - Fuzzy Word Membership (FWM) and Fuzzy Word-Pair Membership (FWPM). The FWM set correlates with the individual word, and the FWPM set corresponds with the adjacent word-pair. The word occurrence statistics in the dataset determines the membership functions of the two fuzzy sets. The two fuzzy memberships are combined to establish the score of the Fuzzy Bigram Association(FBA). The FBA assigns a range in [0, 1] that decides the score of adjacent word-pair for extracting bigrams. The adjacent bigrams are analyzed later to determine higher length ngram terms, where $(n > 2)$. The overall diagram of the proposed work is shown in Fig.2. The modules are marked using circles. Next, we discuss each module extensively.

3.1 Fuzzy Word Membership (FWM)

The $F_i(n)$ is the number of unique words present n times in the dataset. Usually, the value of $F_i(n)$ deteriorates with the increment value of n. We noticed that the value of $F_i(n)$ is high for $n = 1$. The value of $F_i(1)$ is substituted by the average score of

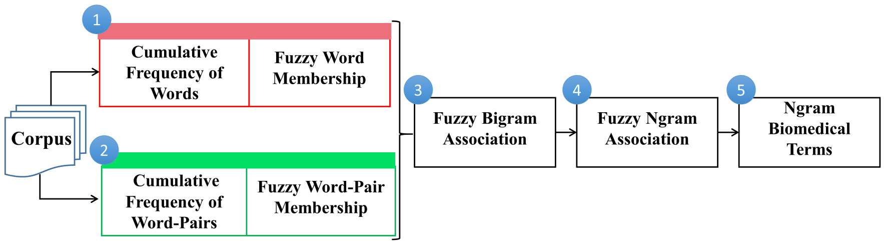

Fig. 2 The overall view of the proposed system

other word occurrences, i.e

$$F_i(1) \leftarrow \frac{1}{(n-1)} \sum_{i=2}^{n} F_i(n) \tag{1}$$

The FWM is a fuzzy set, and its membership score denotes the degree of word occurrence in the dataset. The membership function of FWM is calculated in Equation 2 based on the appearance of all unique words in the dataset.

$$\mu_{FWM}(w_i) = C_i(n)/C_i(N_{max}) \tag{2}$$

Where C_i is the cumulative frequency of each unique word w_i, and N_{max} is the maximum occurrence of an individual word in the dataset. Here,

$$C_i(n) = \sum_{i=1}^{n} F_i(n) \tag{3}$$

When $F_i \neq 0$ and $0 \leq \mu_{FWM}(w_i) \leq 1$. The μ_{FWM} is 0 for all words that not present in the dataset. It is 1 for the words that appear highest in the dataset.

3.2 Fuzzy Word Pair Membership (FWPM)

The $F_{ij}(p)$ is the number of word-pairs appeared p times in the dataset. For the same logic, the value of $F_{ij}(1)$ is substituted by the average of other occurrence values. The FWPM is the fuzzy set and its membership function presents the degree of occurrence of the word-pair in the dataset. The membership value is decided in Equation 4 based on the appearance of all word-pairs in the dataset.

$$\mu_{FWPM}(w_i, w_j) = C_{ij}(p)/C_{ij}(M_{max}) \tag{4}$$

Where C_{ij} is the cumulative frequency of the word pair (w_i, w_j), and M_{max} is the maximum occurrence of an word-pair in the dataset. Here,

$$C_{ij}(p) = \sum_{i=1}^{p} F_{ij}(i) \tag{5}$$

When $F_{ij} \neq 0$ and $0 \leq \mu_{FWPM}(w_i, w_j) \leq 1$. The μ_{FWPM} is 0 for all word pairs that absent in the dataset, and it is 1 for the word pairs that appear highest in the dataset.

3.3 Fuzzy Bigram Association (FBA)

The Fuzzy Bigram Association (FBA) score is measured by combining the two fuzzy sets FWM and FWPM. It is observed that the degree of a word-pair (w_i, w_j) considering as bigram is directly proportional to the value of $\mu_{FWPM}(w_1, w_2)$ and inversely proportional to $\mu_{FWM}(w_i)$ and $\mu_{FWM}(w_j)$. Similarly,

$$FBA(w_j, w_j) = \mu_{FWPM}(w_i, w_j)[1 - \alpha(\mu_{FWM}(w_i) + \mu_{FWM}(w_j))] \tag{6}$$

Where $\alpha \in$ [0, 0.5] and $FBA(w_i, w_j) \in$ [0, 1]. The more score of $FBA(w_i, w_j)$ indicates the more possibility of the word-pair (w_i, w_j) to be a bigram. The word pair (w_i, w_j) is identified as a bigram if the score of $FBA(w_i, w_j)$ is greater than a threshold. One may find any threshold depending on his requirements. If threshold is high the accuracy of precision is high and recall is low, otherwise, precision is low and recall is high. Fig. 3 shows precision / recall scores of FBA for bigram term extraction.

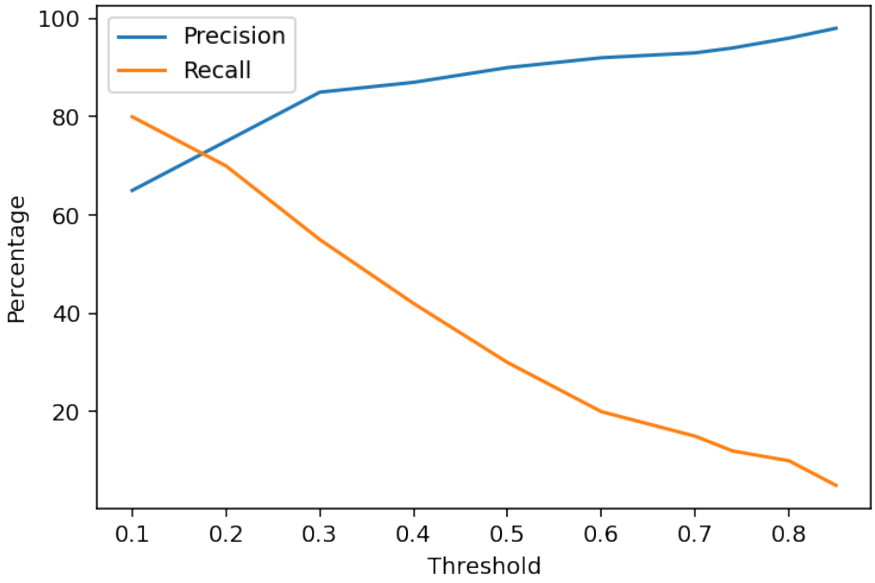

Fig. 3 Precision / recall score varying FBA threshold for bigram term extraction

3.4 Fuzzy Ngram Association(FNA)

The extractor algorithm has two steps. In the first step, it extracts a list of bigrams from the dataset. Optimally, this list contains all bigrams as well as fragments of all ngrams. In the second step, it extends each bigram to ngrams. The list T contains all

Algorithm 1 Fuzzy Bigram Association

```
Require: Dataset G
Ensure: A list of bigrams
$S_1 = (W_1, W_2, W_3, W_4, ...)$    ▷ $S_1$ is the each sub-sentence of G that not contains any stop-word
for all $(word - pair(W_i, W_j))$ do
    $F_{ij} \leftarrow frequency(W_i, W_j)$
    $F_i \leftarrow frequency(W_i)$
    $F_j \leftarrow frequency(W_j)$
    $DatabaseDB \leftarrow ((W_i, W_j), F_{ij}, F_i, F_j)$
end for
for all $((W_i, W_j)inDB)$ do
    if $(FBA(W_i, W_j) > 0.25)$ then
        $T \leftarrow ((W_i, W_j), FBA(W_i, W_j))$    ▷ T contains all extracted bigrams with their fuzzy association value
    end if
end for
```

bigrams extracted from the first step. We extend the bigram when two adjacent bigrams appear in the text adjacently, such as ‘protein kinase’ and ‘kinase activity’ are two adjacent bigrams. The second word of the first bigram and the first word of the next bigram are same; then we combine them. If ’protein kinase activity’ is appeared in the text, then we removed the bigrams with the trigram (n=3) and average association value is assigned for the trigram term. This procedure is recursively applied for extracting ngrams. Our ngram term extraction algorithm is as follows.

Algorithm 2 Fuzzy Ngram Association

```
Require: Bigrams
Ensure: A list of ngrams
while (extend $(k)$) do ▷ Extend $(k)$ returns true when any add / remove operation is performed in the following for-loop
    for all $((W_1, ...W_p)$ and $(W_q, ...W_n)inT)$ do
        if $((W_p == W_q))$ then
            combine $(W_1, ...W_p...W_n)$
        end if
        if $((W_1, ...W_p...W_n)appearsintext)$ then
            remove$(W_1, ...W_p)$ and $(W_q, ...W_n)$ from T
            avg=$(FNA(W_1, ...W_p) + FNA(W_q, ...W_n))/2$
            T $\leftarrow$ add $\{(W_1, ...W_p...W_n), avg\}$
        end if
    end for
end while
```

4 Dataset

We found various biomedical corpora from the previous literature related to term extraction. Some of them are listed in Table 1. The CRAFT Corpus (Bada et al., 2012) is a collection of 97 full-length, open-access biomedical journal articles. It is developed to serve as a high-quality gold standard for the training and testing of advanced biomedical NLP systems. The CorpusDT (da Silva Conrado et al., 2014) dataset is designed for Term extraction in the Brazilian Portuguese language. The ACL RD-TEC dataset (QasemiZadeh and Handschuh, 2014) is used for evaluating the extraction and classification of terms from literature in the domain of computational linguistics. The dataset is derived from the Association for Computational Linguistics anthology reference corpus (ACL ARC). It consists of more than 82,000 manually annotated terms. The Annotated Corpora for Term Extraction Research (ACTER) dataset (Rigouts Terryn et al., 2020) is developed and annotated manually for term extraction. This corpus is domain-specific, and it covers three languages (English, French, and Dutch). The GENIA corpus (Kim et al., 2003) contains approximately 2,000 MEDLINE abstracts. It contains more than 400000 words and almost 100000 annotations for biological terms. It is the primary collection of biomedical literature tagged by XML with biomedical terms. We have mentioned a few datasets in this article in Table 1, but most of them are domain-specific and tagged manually. The GENIA corpus is hand-coded for biological terms. It is a standard dataset for testing the biomedical term. Therefore, we have considered the GENIA corpus for our experimental work.

The dataset is first pre-processed and the frequency of word-pair of consecutive words is counted that simplify to extract biomedical terms. Secondly, the XML tags are removed to extract the plain text from the dataset. In third step, the punctuation marks are discarded. It is mentioned that if a punctuation mark is present between two continuous words, then they are not considered as a word-pair of consecutive words. So, each sentence is broken many times for the appearance of punctuation marks. Finally, the word and ward-pair are identified from each sentence and counted the frequency.

Table 1 Publicly available biomedical corpora

Dataset Name	Reference	Description	Application
CRAFT	(Bada et al., 2012; Cohen et al., 2017)	BioNLP Corpora: Manually annotated corpus consisting of 67 full-text biomedical journals	Concept annotation
CorpusDT	(da Silva Conrado et al., 2014)	To build corpora for supporting NLP researches, especially on Brazilian Portuguese	Term extraction in Brazilian Portuguese
ACL RD-TEC	(QasemiZadeh and Handschuh, 2014)	A dataset for evaluation of term and entity recognition in computational linguistics	Terminology extraction
ACTER	(Rigouts Terryn et al., 2020)	Annotated corpora for term extraction research	Term extraction
GENIA	(Kim et al., 2003; Terryn et al., 2019)	A semantically annotated corpus of biological literature	Term extraction, ontology creation, Part-of-speech tagger

5 Experimental Results

The experimental evaluation of proposed biomedical terms extraction methods is presented in this section. The performance of the scheme is demonstrated on GENIA Version 3.02 text collection (Kim et al., 2003). The precision and recall measures are widely used for term extraction. So, the results are evaluated on these two measures. The α is set to 0.25 and the FBA threshold is set to 0.5. The precision and recall score is shown in table 2. It is notices that the propose method gives high precision and low recall value that means its extraction ratio is low but most of the extracted terms are correct. Therefore this method is more applicable where high precision score is required and recall is not important. The α is set to 0.25 and FBA threshold is set to 0.25 to increase the recall value in table 3. The recall value is also increased when the dataset size is large. The F-measure is high at the last row of the table indicate it gives better result when consider the whole dataset. The overall best result was achieved by FNA: 63.16 recall with 53.77 precision (F-measure 58.09) in table 4.

Table 2 Precision / Recall Scores of FNA (When $FBA = 0.5$ and $\alpha = 0.25$)

				Percentage of Ranked List					
10	20	30	40	50	60	70	80	90	100
				Precision					
93.93	89.31	84.77	85.11	83.18	83.96	85.15	84.73	85.39	86.23
				Recall					
0.15	0.28	0.41	0.54	0.66	0.81	0.96	1.09	1.23	1.38
				F-Measure					
0.30	0.57	0.81	1.09	1.32	1.60	1.89	2.15	2.44	2.73

Table 3 Precision / Recall Scores of FNA (When $FBA = 0.25$ and $\alpha = 0.25$)

				Percentage of Ranked List					
10	20	30	40	50	60	70	80	90	100
				Precision					
87.35	82.91	79.87	77.33	74.98	73.14	71.36	70.37	69.22	67.65
				Recall					
2.11	4.00	5.78	7.46	9.05	10.59	12.06	13.59	15.04	16.33
				F-Measure					
4.12	7.63	10.79	13.62	16.15	18.51	20.63	22.78	24.71	26.31

The comparative analysis of our FNA with other popular statistical association measures (da Silva Conrado et al., 2014) is shown in Fig. 4 and Fig. 5. The green curve shows the extraction result of the biomedical term using the Log-likelihood association method. It got comparatively low results than the other two association methods. T-test (orange curve) and Mutual Information (red curve) got almost the same results. There are many associations available in the literature to extract keywords. Log-likelihood, T-test, and Mutual-information are used popularly as statisti-

Table 4 Precision / Recall Scores of FNA (When $FBA = 0$ and $\alpha = 0.25$)

Percentage of Ranked List									
10	20	30	40	50	60	70	80	90	100
Precision									
78.91	74.21	69.45	71.14	71.32	70.69	69.59	67.91	65.86	63.16
Recall									
6.72	12.63	17.73	24.22	30.36	36.11	41.47	46.25	50.46	53.77
F-Measure									
12.38	21.59	28.26	36.14	42.59	47.80	51.97	55.03	57.14	58.09

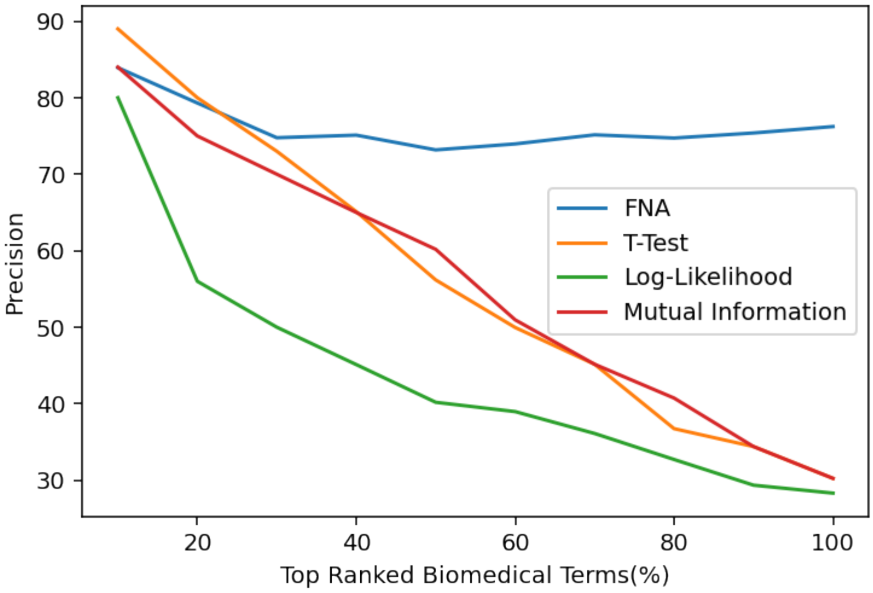

Fig. 4 Comparative analysis of precision for biomedical term extraction

cal methods for term extraction. Our approach got the highest (blue curve) accuracy than the other association, especially, when considering the whole dataset. Fig. 4 shows the comparative analysis of the precision, and Fig. 5 depicts the recall results.

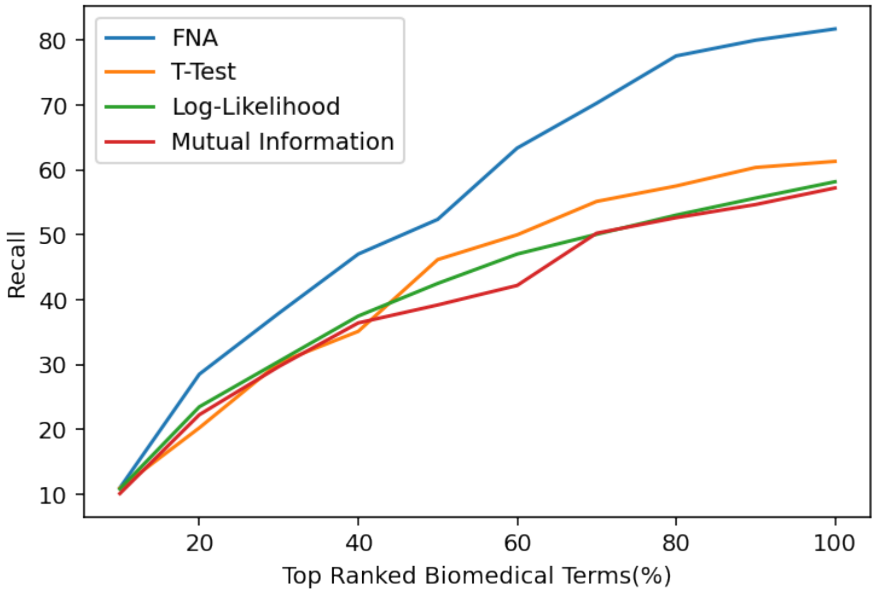

Fig. 5 Comparative analysis of recall for biomedical term extraction

6 Conclusion

Meaningful terms are linguistically motivated. The fuzzy approach deals with the linguistic properties of elements. Based on this intuition, a fuzzy-based biomedical term extraction technique is described in this article. The membership functions of two different fuzzy sets are combined to extract biomedical terms. The experimental results prove the utility of the fuzzy approach for biological term extraction, especially when a considered rank list is heigh. Future work will be focused on extracting the new biomedical terms from recently published articles on the web, which are not present in the current biomedical dictionary. Another work will be focused on answering biomedical questions automatically using biological terms.

Acknowledgment

Ethics approval and consent to participate: This article does not contain any studies with human participants or animals performed by any of the authors. Informed consent: Informed consent was obtained from all individual participants included in the study.
Consent for publication: All authors reviewed and provides consents for publication.
Availability of data and material: The dataset used is a public dataset.
Competing interests: The authors declare that there is no conflict of interest regarding the publication of this paper.
Funding: This study is not Funded from anywhere.

Authors' contributions: B. Das conceptualize, experiments, and write the manuscript. M. Majumdar, S. Pakida, and A. Ahmed supervised, validate, and write the manuscript.

References

Bada M, Eckert M, Evans D, Garcia K, Shipley K, Sitnikov D, Baumgartner WA, Cohen KB, Verspoor K, Blake JA, et al. (2012) Concept annotation in the craft corpus. BMC bioinformatics 13(1):161

Cohen KB, Lanfranchi A, Choi MJy, Bada M, Baumgartner WA, Panteleyeva N, Verspoor K, Palmer M, Hunter LE (2017) Coreference annotation and resolution in the colorado richly annotated full text (craft) corpus of biomedical journal articles. BMC bioinformatics 18(1):1–14

Conrado MS, Pardo TA, Rezende SO (2013) Exploration of a rich feature set for automatic term extraction. In: Mexican International Conference on Artificial Intelligence, Springer, pp 342–354

Dobrov BV, Loukachevitch N (2011) Multiple evidence for term extraction in broad domains. In: Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, pp 710–715

Gelbukh A, Sidorov G, Lavin-Villa E, Chanona-Hernandez L (2010) Automatic term extraction using log-likelihood based comparison with general reference corpus. In: International conference on application of natural language to information systems, Springer, pp 248–255

Golik W, Bossy R, Ratkovic Z, Nédellec C (2013) Improving term extraction with linguistic analysis in the biomedical domain. Res Comput Sci 70:157–172

Herrero-Zorita C, Campillos-Llanos L, Moreno-Sandoval A (2014) Collecting and pos-tagging a lexical resource of japanese biomedical terms from a corpus. Procesamiento del Lenguaje Natural 52:29–36

Herrero-Zorita C, Molina C, Moreno-Sandoval A (2015) Medical term formation in english and japanese. Review of Cognitive Linguistics Published under the auspices of the Spanish Cognitive Linguistics Association 13(1):81–105

Heylen K, De Hertog D (2015) Automatic term extraction. Handbook of terminology 1(01)

Hong L, Lin J, Li S, Wan F, Yang H, Jiang T, Zhao D, Zeng J (2020) A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories. Nature Machine Intelligence pp 1–9

Kim JD, Ohta T, Tateisi Y, Tsujii J (2003) Genia corpus—a semantically annotated corpus for bio-textmining. Bioinformatics 19(suppl_1):i180–i182

Koza W, Solana Z, Conrado MdS, Rezende SO, Pardo TA, Díaz-Labrador J, Abaitua J (2011) Extracción terminológica en el dominio médico a partir del reconocimiento de sintagmas nominales [0]. Revista de Lingüística Informática, Modelización e Ingeniería Lingüística-INFOSUR pp 27–40

Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C (2016) Neural architectures for named entity recognition. arXiv preprint arXiv:160301360

Lamurias A, Couto FM (2019) Text mining for bioinformatics using biomedical literature. Encyclopedia of bioinformatics and computational biology 1:602–611

Li D, Huang L, Ji H, Han J (2019) Biomedical event extraction based on knowledge-driven tree-lstm. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp 1421–1430

Lossio Ventura JA, Hacid H, Ansiaux A, Maag ML (2012) Conversations reconstruction in the social web. In: Proceedings of the 21st International Conference on World Wide Web, pp 573–574

Lossio-Ventura JA, Jonquet C, Roche M, Teisseire M (2013) Combining c-value and keyword extraction methods for biomedical terms extraction. In: LBM: Languages in Biology and Medicine

Lossio-Ventura JA, Jonquet C, Roche M, Teisseire M (2014a) Biotex: A system for biomedical terminology extraction, ranking, and validation. In: ISWC: International Semantic Web Conference

Lossio-Ventura JA, Jonquet C, Roche M, Teisseire M (2014b) Yet another ranking function for automatic multiword term extraction. In: International Conference on Natural Language Processing, Springer, pp 52–64

Lossio-Ventura JA, Jonquet C, Roche M, Teisseire M (2016) Biomedical term extraction: overview and a new methodology. Information Retrieval Journal 19(1-2):59–99

Lyu C, Chen B, Ren Y, Ji D (2017) Long short-term memory rnn for biomedical named entity recognition. BMC bioinformatics 18(1):462

Martin-Bautista M, Sánchez D, Serrano J, Vila M (2004) Text mining using fuzzy association rules. In: Fuzzy logic and the internet, Springer, pp 173–189

Murdoch TB, Detsky AS (2013) The inevitable application of big data to health care. Jama 309(13):1351–1352

Newman D, Koilada N, Lau JH, Baldwin T (2012) Bayesian text segmentation for index term identification and keyphrase extraction. In: Proceedings of COLING 2012, pp 2077–2092

Pennington J, Socher R, Manning C (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543

Perez-Guadarrama Y, Simón-Cuevas A, Hojas-Mazo W, Olivas JA, Romero FP (2018) A fuzzy approach to improve an unsupervised automatic keyphrase extraction process. In: 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, pp 1–6

Periñán-Pascual C (2018) Dexter: A workbench for automatic term extraction with specialized corpora. Natural Language Engineering 24(2):163–198

Piskorski J, Stefanovitch N, Jacquet G, Podavini A (2021) Exploring linguistically-lightweight keyword extraction techniques for indexing news articles in a multilingual set-up. In: Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pp 35–44

QasemiZadeh B, Handschuh S (2014) The acl rd-tec: a dataset for benchmarking terminology extraction and classification in computational linguistics. In: Proceedings of the 4th International Workshop on Computational Terminology (Computerm), pp 52–63

Rei M (2017) Semi-supervised multitask learning for sequence labeling. arXiv preprint arXiv:170407156

Rigouts Terryn A, Hoste V, Drouin P, Lefever E (2020) Termeval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset. In: 6th International Workshop on Computational Terminology (COMPUTERM 2020), European Language Resources Association (ELRA), pp 85–94

Samy D, Moreno-Sandoval A, Bueno-Díaz C, Salazar MG, Guirao JM (2012) Medical term extraction in an arabic medical corpus. In: LREC, pp 640–645

Sandoval AM, Díaz J, Llanos LC, Redondo T (2019) Biomedical term extraction: Nlp techniques in computational medicine. IJIMAI 5(4):51–59

Segura-Bedmar I, Martínez P, Samy D (2008) Detección de fármacos genéricos en textos biomédicos. Procesamiento del lenguaje Natural 40

da Silva Conrado M, Di Felippo A, Pardo TAS, Rezende SO (2014) A survey of automatic term extraction for brazilian portuguese. Journal of the Brazilian Computer Society 20(1):12

Stoykova V, Petkova E (2012) Automatic extraction of mathematical terms for precalculus. Procedia Technology 1:464–468

Terryn AR, Hoste V, Lefever E (2019) In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora. Language Resources and Evaluation pp 1–34

Torres A, Nieto JJ (2006) Fuzzy logic in medicine and bioinformatics. BioMed Research International 2006

Zehtab-Salmasi A, Feizi-Derakhshi MR, Balafar MA (2021) Frake: Fusional realtime automatic keyword extraction. arXiv preprint arXiv:210404830