IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 April 2023, accepted 21 April 2023, date of publication 1 May 2023, date of current version 15 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3272056

RESEARCH ARTICLE

Weighted Co-Occurrence Bio-Term Graph for Unsupervised Word Sense Disambiguation in the Biomedical Domain

ZHENLING ZHANG[1,4], YANGLI JIA[1,4], (Member, IEEE), XIANGLIANG ZHANG[1,4], MARIA PAPADOPOULOU[2,3,4], AND CHRISTOPHE ROCHE[3,4], (Member, IEEE)

[1]School of Computer Science, Liaocheng University, Liaocheng 252059, China
[2]School of Philosophy, National Kapodistrian University of Athens (NKUA), 15784 Athens, Greece
[3]Computer Science, Systems, Information and Knowledge Processing Laboratory (LISTIC Laboratory), Université Savoie Mont Blanc, 73376 Chambéry, France
[4]Knowledge Engineering and Terminology Research Centre (KETRC), Liaocheng University, Liaocheng 252059, China

Corresponding author: Yangli Jia (jiayangli@lcu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 81973695.

ABSTRACT Word Sense Disambiguation (WSD) is a significant and challenging task for text understanding and processing. This paper presents an unsupervised approach based on Weighted Co-occurrence bio-Term Graph (WCOTG) for performing WSD in the biomedical domain. The graph is automatically created from biomedical terms that are extracted from a corpus of downloaded scientific abstracts. Two kinds of weights are introduced on the links of the built bio-term graph and are taken as important factors in the process of disambiguation. The modified Personalised PageRank (PPR) algorithm is used for performing WSD. When evaluated on the NLM-WSD and MSH-WSD test datasets, and an acronym test set, the method outperforms the widely used unsupervised ones addressing the same problem, and the average result is almost equal to that of the BlueBERT_LE-based method. In contrast, our method has no additional enhancement or training for BERT-based models. Comparative experiments validate the positive effect of links' weight on disambiguation efficiency. Last, the statistical experiments on the relation among system accuracy, the numbers of medical abstracts in the corpus, and the corresponding extracted terms suggest an excellent minimum corpus scale, when resources are limited.

INDEX TERMS Biomedical informatics, biomedical natural language processing, word sense disambiguation, unified medical language system, personalised PageRank algorithm.

I. INTRODUCTION

Word Sense Disambiguation (WSD) systems attempt to automatically identify the proper sense of ambiguous words in context [1], [2]. For example, WSD would aim to identify the meaning of the word ''cold'' to be ''cold temperature'' or ''common cold'' depending on the context in which it occurs. WSD is often characterized as an intermediary step in the process of understanding natural language texts [2], [3]. It is beneficial for applications in the biomedical domain, such as information extraction, automated knowledge discovery, question-answering [4], etc.

The associate editor coordinating the review of this manuscript and approving it for publication was Giovanni Dimauro.

The disambiguation task in the biomedical field is aimed at medical texts and terms. Combining domain knowledge can better improve the system's performance. Currently, the Unified Medical Language System (UMLS) Knowledge Sources are widely used in tasks such as disambiguation and automatic question-answering in the biomedical field. However, existing work such as [5] and [6], etc., typically involves extracting concepts (CUIs) from UMLS to build systems. In disambiguation tasks, this can lead to a mapping bottleneck when converting polysemy in biomedical documents into concepts. Therefore, we build term graphs directly, not concept graphs, the latter requires the use of term-to-concept mapping tools. This method can be considered an improvement of the existing work by Duque et al. [5], whose

knowledge base is an unweighted concept graph. We not only directly construct the term graph, but also add weights to the edges of the graph to improve the system's disambiguation performance. Transformer-based neural network models in the biomedical domain and co-occurrence probability of terms are used to generate weights in graphs.

In summary, this paper proposes a novel approach based on Weighted Co-occurrence bio-Term Graphs (WCOTG) for performing WSD in the biomedical domain. The graphs are automatically created in an unsupervised way from biomedical terms that are extracted from a corpus of downloaded scientific abstracts. Each graph represents relations between the ambiguous term and the terms that appear frequently in the same document. The corresponding weight depends on the co-occurrence probability or relatedness between them.

The most important contributions of this paper are:

1) We build graphs from terms instead of concepts, which bypasses the problem of manual disambiguation of the mapping tools when transforming polysemous terms from biomedical documents into concepts.

2) We add weights to the links of the built bio-term graphs and take them as important factors in the process of disambiguation. Two weighting methods are applied to weight the co-occurrence term graph. The co-occurrence probability of terms and the relatedness value between BERT-based vectors of terms are taken as weight respectively. The disambiguation results of weighted and unweighted bio-term graphs show that weight has a positive impact on disambiguation performance.

3) We extract sub-term graphs with a relatively small amount of data for performing WSD, so as to improve the running efficiency of the system.

4) The PPR algorithm which is normally applied over the unweighted directed graph is modified for the WSD task over the weighted undirected term graph.

The rest of the paper is organized as follows. Section II describes previous approaches to biomedical WSD. Section III presents the proposed system, describing the steps for building a bio-term graph and the disambiguation process. In section IV, corpus building is discussed. In section V, several comparative evaluations and further experiments on different test datasets are carried out. Finally, conclusions and future work are presented in Section VI.

II. PREVIOUS WORK

Existing methods to automatically resolve ambiguity can be classified into two different perspectives: knowledge-based (or not), and supervised (or not). The former can be further distinguished into knowledge-based only, corpus-based only, and (both) knowledge & corpus-based. For example, the Enhanced WSD Integrating Synset Embeddings and Relations (EWISER) is a hybrid knowledge-based and supervised approach to WSD that integrates explicit relational information from the WordNet LKB [7]. The latter can be further subdivided into supervised, unsupervised, and semi-supervised. Supervised methods normally apply SVM (Support Vector Machine) [8], [9] and neural networks [7], [10], [11] to identify the proper sense of the ambiguous target word. Recently, in the biomedical domain, the recurrent network LSTM (Long Short-Term Memory) [8] and deepBioWSD network [12] for WSD have achieved state-of-the-art accuracy. The disadvantage of these supervised approaches is that they require lots of labeled training data, which are extremely time-consuming and expensive to create [13]. Semi-supervised approaches offer solutions to that problem by using techniques to generate sense-tagged examples automatically [9], [14]. Unsupervised methods do not require labeled training examples and typically use graph-based clustering techniques [15]. Recently, word embedding models [16] and pre-trained language model BERT (Bidirectional Encoder Representation from Transformers) [17], [18] and its variant BERT models [19], [20], [21] all pre-trained on large corpora were introduced to conduct unsupervised learning for WSD. For example, Mao and Wah [6] generate semantic relatedness measurements between UMLS concepts to achieve disambiguation by applying the word embedding models and various flavors of BERT.

The work by Duque et al. [5] is an unsupervised method. It introduces a graph-based approach for WSD in the biomedical domain, wherein, the UMLS database is explored to convert text from the original document set to biomedical concepts, and a corpus of biomedical documents is used to build a co-occurrence graph that is then analyzed to identify the meanings of ambiguous words. The evaluation results show that it outperforms most of the knowledge-based methods for WSD in the biomedical domain. However, the disambiguation algorithm does not consider the role of the links' weights in the graph.

Inspired by this work, we go further to make some improvements to it. This paper proposes a novel approach based on Weighted Co-occurrence bio-Term Graphs (WCOTG) for performing WSD in the biomedical domain. On the one hand, the weights of links in the created graph are taken into account in the disambiguation algorithms. Comparative evaluation of weighted bio-term graphs and non-weighted ones is also performed. On the other hand, the work by Duque et al. [5] makes use of the UMLS MetaMap[1] program to convert biomedical texts to concepts and build concept graphs, while in our case, a co-occurrence graph is built based on terms directly instead of concepts. This is mainly based on the following considerations. Mapping programs [22] have automatic disambiguation limitations. ''One of the most difficult problems MetaMap deals with is ambiguity''[4]. Using MetaMap, monosemous words will be attached to just one concept, whereas polysemous words may be attached to several concepts [5]. Although the MetaMap program offers a disambiguation server, it cannot eliminate

[1] https://metamap.nlm.nih.gov/Docs/FAQ/WSD.pdf

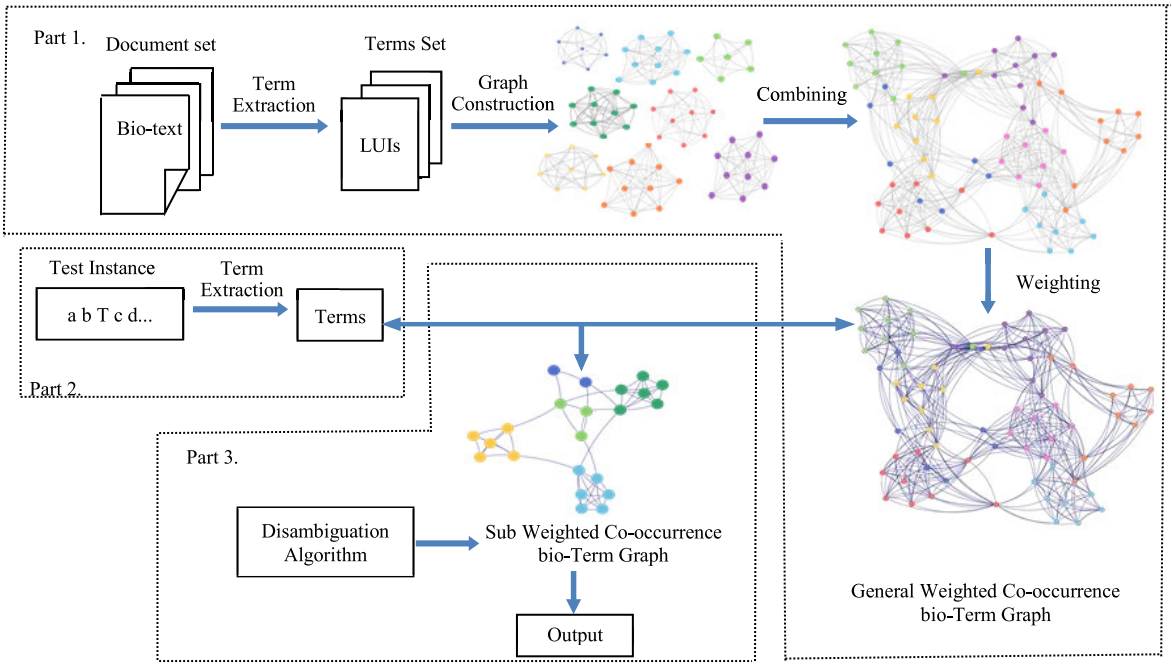

FIGURE 1. Disambiguation system based on weighted co-occurrence bio-term graph.

most of the ambiguity. And manually assistant mapping is difficult and time-consuming.

Furthermore, we use a transformer-based neural network model to calculate the weights between terms in the built graph. In recent years, pre-trained neural language models, such as BERT (Bidirectional Encoder Representation from Transformers) [17], [18], [23], XLNET [24], GlossBERT [25], BEM [26], and ALBERT [27], etc. have achieved state-of-the-art results in various natural language processing (NLP) tasks. In Word Sense Disambiguation (WSD) tasks, Loureiro and Jorge [28] used the BERT model to create sense embeddings for all senses in WordNet. Vial et al. [29] and Kumar et al. [30] used pre-trained BERT word vectors as input embeddings. In the biomedical domain, transformer-based neural network models are also successfully used to predict the DNA promoters [31] and identify DNA N6 [32]. In the clinical domain, Biseda et al. [33] predict ICD codes from hospital notes by utilizing the Clinical BERT variant. Mulyar et al. [34] developed Multitask-Clinical BERT for multitasking information extraction.

However, WSD in the biomedical field is an intermediary step, which requires not only natural language processing methods or models but also domain knowledge or corpus support. Therefore, neural network models in the general field have significant limitations in the field of biomedical disambiguation. The transformer-based neural network models in the general field are normally pre-trained on general English corpora (English Wikipedia and BooksCorpus) and designed for general language understanding. Therefore, they often perform poorly in biomedical natural language processing tasks, because biomedical texts contain a large number of domain-proper nouns and terms. Then some variant models of BERT in the field of biomedicine are proposed, and great success has been achieved in biomedical natural language processing tasks. They are all pre-trained on large domain corpora. For example, BioBERT [19] is initialized with weights from the original BERT and then is pre-trained on biomedical domain corpora (PubMed abstracts and PMC full-text articles) for 23 days on eight NVIDIA V100 GPUs [19]. Blue_BERT (NCBI BERT, Biomedical Language Understanding Evaluation) [20], another popular BERT-based model in the biomedical domain, is pre-trained on PubMed abstracts and MIMIC-III clinical notes. Based on the support of the domain knowledge source (Unified Medical Language System), this paper applies domain BERT to the constructed term graph and achieves good disambiguation results.

III. WSD BASED ON WEIGHTED CO-OCCURRENCE BIO-TERM GRAPH (WCOTG)

A. SYSTEM DESCRIPTION

Our system relies on a data model of weighted co-occurrence bio-term graph and exploits a modified Personalized PageRank (PPR) to generate disambiguation in the biomedical domain. The complete system is shown in Fig.1.

Part 1 illustrates the steps for building a weighted co-occurrence bio-term graph. Each bio-document in the original set is transformed into biomedical terms through automatic term extraction, and this work is based on a dictionary, e.g., the dictionary auto-generated from the file MRCONSO.RRF[2] in the Unified Medical Language

[2]https://www.ncbi.nlm.nih.gov/books/NBK9676/

System (UMLS) Knowledge Sources. Then each term set forms a fully connected graph. The terms extracted from the bio-document eventually become the nodes of the graph, and the co-occurrence relations between terms, i.e., appearing in the same document, become the edges (links) in the graph. Next, we combine all the above full connection graphs to get a general co-occurrence graph by removing duplicate nodes and links. At last, a measure of relation for each pair of nodes is applied as the weight for the corresponding edge in the graph. The weighting method is introduced in Section III-B.

In Part 2, for a test instance of the ambiguous target term T, biomedical terms are extracted from it. Referring to these terms, in Part 3, a sub-weighted bio-term graph is extracted from the general graph constructed above, in order to improve the efficiency of the subsequent disambiguation algorithm. In subgraph, the terms (nodes) come from the intersection of the terms in the general graph and the one extracted from the test instance. And the relationships between the nodes in the subgraph come from their relationships in the general graph. Through the construction of a subgraph, i.e., extracting a subgraph from the general graph, the nodes and edges that have nothing to do with the current disambiguation task are greatly removed, which means the size of disambiguation data is reduced effectually.

In Part 3, a disambiguation algorithm introduced in Section III-C is applied to output the most suitable sense of the ambiguous target term in the test instance.

B. WEIGHTING METHODS

In this article, two weighting methods are applied to weight the co-occurrence term graph.

1) TAKING THE CO-OCCURRENCE PROBABILITY OF TERMS AS WEIGHT

Duque [35] introduced a method to calculate the p-value for the co-occurrence of each pair of CUIs (CUI-Concept Unique Identifiers) in the corpus. Inspired by this, we use a similar idea to measure the co-occurrence probability between biomedical terms.

Suppose there are two terms T_1 and T_2 appear in the n_1 and n_2 number of documents respectively. The number of how many ways those terms could co-occur in exactly k documents can be given by the multinomial coefficient[3]:

$$\binom{N}{k, n_1 - k, n_2 - k} \tag{1}$$

where, $N = n_1 + n_2 - k$, the total number of documents is n ($n \geq N$). The probability of the terms exactly co-occurring k times, that is, co-occurring in k documents, by pure chance is given by:

$$p(k) = \binom{N}{n_1}^{-1} \binom{N}{n_2}^{-1} \binom{N}{k, n_1 - k, n_2 - k} \tag{2}$$

[3]https://brilliant.org/wiki/multinomial-coefficients/

Equation (2) can be written as follows.

$$p(k) = \prod_{j=0}^{n_2-k-1} \left(1 - \frac{n_1}{N-j}\right) \times \prod_{j=0}^{k-1} \frac{(n_1 - j)(n_2 - j)}{(N - n_2 + k - j)(k - j)} \tag{3}$$

In our work, $p(k)$ is taken as the p-value of each pair of co-occurrence terms. And the weight of the edge between two nodes V_i and V_j in the co-occurrence graph is set as:

$$w_{ij} = log(\frac{p_{ij}}{p_{min}}) \tag{4}$$

where, the p_{ij} is the p-value of the link (V_i, V_j), and the p_{min} is the minimum value of all the p-values of each pair of co-occurrence terms.

2) TAKING THE RELATEDNESS VALUE BETWEEN TERM VECTORS AS WEIGHT

Another weighting method proposed is that the relatedness value between biomedical terms is taken as the weight for the corresponding edge in the graph. In this paper, we use a transformer-based neural network model to calculate the relatedness of terms in the graph.

A common way to use BERT and its variants is to combine them with the fine-tuning process for a specific NLP task. However, in our work, BioBERT and Blue_BERT models are used alone to generate contextualized word embeddings for each term in the graph without additional training data. Before applying BERT-based models, two special tokens are inserted into a term sequence, the classifier token [CLS] at the beginning, and the ending token [SEP] at the end. Then we get the tokened term sequences. For example, in the built graph, there is a term ''regulation of carbohydrate metabolism'' (LUI: L2321589) linked to the ambiguous target term ''Digestive''. The term sequence for it is made up as follows.

[CLS] regulation of carbohydrate metabolism [SEP]

Given such term sequences, a BERT-based model is applied to them separately. Then for each sequence, we get a vector, i.e., term embedding combined with the embedding of the tokens and the average embedding of each word in the sequence.

Next, semantic relevance measurement is performed based on cosine similarity theory, which is to evaluate the similarity or relatedness of two vectors by calculating the cosine value of the angle between them. The greater the cosine value, the greater the similarity between the two vectors, and the more related or similar the semantics of the terms corresponding to the vectors.

In this way, the weight of the edge between two nodes V_i and V_j in the co-occurrence graph is set as:

$$w_{ij} = cos(EV_i, EV_j) \tag{5}$$

where, the EV_i and EV_j are the BERT-based vectors, i.e., term embeddings, OF THE corresponding terms.

C. DISAMBIGUATION ALGORITHM

The disambiguation algorithm that we have selected is the Personalised PageRank algorithm [36], [37] which has been successfully applied to WSD tasks. Personalized PageRank (PPR) evaluates the importance of the vertices of a graph depending on its link structure. It is normally based on a directed graph, represented by its incoming and outgoing edges.

Formally, let G be a directed graph with N vertices (V_1, $V_2 \ldots, V_N$). The PPR value of vertex V_i in the graph can be calculated through the following formula:

$$P = cMP + (1 - c)\, v \tag{6}$$

where c is called a damping factor, a scalar value between 0 and 1. It is usually set in the range [0.85, 0.95], and the value of 0.85 for it is reported to be the best [38]. P is the vector that contains the PPR values for each node. M is the N × N transition probability matrix corresponding to the directed graph G, built in this way:

$$M_{ij} = \begin{cases} \dfrac{1}{outdegree\ (i)} & if\ V_i \to V_j \\ 0 & otherwise \end{cases} \tag{7}$$

If there is a link from V_i to V_j then the matrix entry M_{ji} has the value $1/d_i$ where d_i is the out-degree of vertex V_i, and all other entries have the value 0. And $\boldsymbol{v}$ is an $N \times 1$ vector.[4]

For a given vertex V_i, the PPR value can be calculated by a rewritten Equation of (8):

$$p\,(V_i) = c \sum\nolimits_{V_{j \in In(V_i)}} \frac{1}{d_j} p\left(V_j\right) + (1 - c)\, v_i \tag{8}$$

where In(V_i) is the set of vertices pointing to the vertex V_i, d_j the out-degree of vertex V_j, and v_i the element of the vector $\boldsymbol{v}$.

The PPR algorithm is normally applied over the directed graph, to use the PPR algorithm for WSD over our built term graph, undirected graph, in which the vertices represent terms and the edges represent relations between terms, we modify (8) to the following Formula:

$$p\,(V_i) = \mathrm{c} \sum\nolimits_{V_{j \in Link(V_i)}} \frac{1}{d_j} p\left(V_j\right) + (1 - c)\, v_i \tag{9}$$

Furthermore, considering the weight on edges, (9) is modified as follows:

$$p\,(V_i) = \mathrm{c} \sum\nolimits_{V_{j \in Link(V_i)}} \frac{1}{d_j} p\left(V_j\right) w_{ij} + (1 - c)\, v_i \tag{10}$$

where w_{ij} defined by (4) or (5) is taken as the weight on the edge between the vertices V_i and V_j. The Link(V_i) in (9) and (10) is the set of vertices linking to the vertex V_i, and d_j is the degree of vertex V_j.

Suppose there are total n vertices in the graph, the personalization vector $\boldsymbol{v}$ used in PPR is initialized with the value as follows:

$$\boldsymbol{v} = \left(\frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \ldots\ldots, \frac{1}{n}\right) \tag{11}$$

[4]Note that, v_i in lower case refers to one element in vector $\mathbf{v}$, while V_i in capital letter refers to a vertex in the graph.

The modified PPR (Personalized PageRank) is calculated by applying an iterative algorithm that computes (10) repeatedly until a pre-specified number of iterations have been executed. After that, each node in the graph will be assigned a PPR value for ranking. And the candidate term that has the highest PPR value will be finally chosen as a disambiguation result.

IV. DATASETS

This section introduces the datasets for the evaluation and presents in detail the process of building a corpus that is used for constructing the co-occurrence term graph.

A. NLM-WSD AND MSH-WSD TEST COLLECTIONS AND CORPUS FOR BUILDING THE CO-OCCURRENCE TERM GRAPH

Two medical-text test collections of the National Library of Medicine's Word Sense Disambiguation (NLM-WSD)[5] are used to evaluate the performance of the proposed method. One collection is named "NLM-WSD Test Collection". It consists of 50 highly frequent ambiguous terms. Each of the 50 ambiguous cases has 100 ambiguous instances randomly selected from the 1998 MEDLINE citations. The average number of senses per ambiguous term is 2.3. Each instance of the ambiguous word considered for disambiguation is determined manually with a correct sense and annotated with a sense number by human evaluators. Some instances are marked as None by evaluators, for there are no appropriate senses in UMLS Metathesaurus for them. We have discarded those instances with the None category, so finally, the test collection contains 3, 983 instances and 49 ambiguous terms out of the 50s since the instances of the ambiguous term "association" have been assigned entirely to None. Another test dataset is named "MSH-WSD Test Collection". It consists of 203 highly frequent ambiguous terms and contains both biomedical terms and abbreviations. 37, 888 ambiguity instances are randomly selected from the 2010 MEDLINE citations. Each instance of the ambiguous word has a correct sense annotated by human evaluators. Each ambiguous case contains approximately 187 instances and has 2.08 possible senses. There is no None category marked in this test collection.

In this work, we present an unsupervised system that does not need any annotations, however, we need to acquire data to build the co-occurrence term graph. For the evaluation of the NLM-WSD and MSH-WSD test collection, we performed a search on Medline for each ambiguous target term. And after downloading the file in PubMed format from Medline, the results are restricted to 800 (Or the maximum number that has been searched) PubMed format files per target term with two senses, 1200 for the target term with 3 senses, and so on. And then we extracted abstracts only from the downloaded PubMed format files. After checking and de-duplication with

[5]The NLM-WSD and MSH-WSD test datasets are available from the National Library of Medicine: https://lhncbc.nlm.nih.gov/ii/areas/word-sense-disambiguation.html

TABLE 1. A resume of the acronym datasets.

	WCOTG/ UWCOTG				BCG		
Test datasets	T100	T150	T200	T300	A100	A200	A300
Ambiguous terms	18	14	16	14	18	16	14
Instances	1800	1923	3105	4200	1800	3199	4199
The total number of abstracts for graph building	21075	16781	17836	15064	50143	50143	50143
Min/Max senses per term	2	3	2/4	2/4	2/4	2/4	2/4
Average senses	2	3	2.92	2.85	2.61	2.5	2.57

each test instance in the test set, the total number of abstracts in the corpus for the NLM-WSD test is 35809. This number may not be the best choice for the disambiguation system, but it is very close to the number of abstracts (35282) in the corpus of the work [5], which is helpful to objectively compare the efficiency of the two methods on the same corpus scale. And for the MSH-WSD test, the total number of abstracts in the corpus is 155233.

B. ACRONYM CORPUS

We also validate the proposed method on the acronym test dataset. Stevenson and Guo [38] provided an acronym corpus that has been successfully used in WSD research [5], [39], [40]. However, we failed to successfully download this corpus, so we built new acronym corpora containing 22875 abstracts downloaded from Medline. Each of these abstracts contains an ambiguous acronym from the set of 18 acronyms originally developed in [41] and is widely used in previous studies. Each acronym, used as an ambiguous target term in evaluation, consists of 3 letters, and it is associated with between 2 and 4 extended forms which are considered candidate senses.

We create 18 corpora, and each corpus is corresponding to a target acronym. The specific creation process is similar to the process of creating the corpus for building graphs introduced in Section IV-A. The difference is that extended forms, i.e., candidate senses, are gotten by querying each item in the MRCONSO.RRF file of UMLS Metathesaurus, instead of from the NLM-WSD or MSH-WSD Test Collection. For example, as to the acronym "ANA", two extended forms "American Nurses Association" and "Antinuclear Antibody" have the same CUI as it in the MRCONSO.RRF file. So they are extracted from the MRCONSO.RRF file as the candidate senses for the acronym "ANA". And then a search query is made in Medline based on each candidate sense. We download 600 (Or the maximum number that has been searched) PubMed format files for each candidate sense and convert them to 600 files that only contain abstracts. Then we split these abstract files into two parts. One is used to form the test datasets, and the other is used as the corpus for graph building. In partitioning, the data used to create the graph will never be included in any test datasets. Most of the test datasets are evenly distributed, i.e., there is nearly the same number of test instances corresponding to each candidate sense, except the test dataset for the acronyms "CMV", "DIP", "LAM" and "MAC", due to an insufficient number of instances in the original corpus.

Four test datasets are created and referred to as T100, T150, T200, and T300. Table 1 provides a resume of the acronym datasets used for the evaluation. It can be seen, the number of senses per term in T100 and T200 is fixed. That is, the first two datasets T100 and T150 both have fixed numbers of senses for each ambiguous word, 2 for T100 and 3 for T150 respectively. All 18 acronyms are present in T100. For an insufficient number of their extended forms, the acronyms "ANA", "BPD", "EMG", and "RSV" are not present in the data set T150 whose sense length is 3. The last two datasets T200 and T300 are mixed-length datasets, and each acronym has between 2 and 4 extended forms. These two datasets correspond to the datasets A200 and A300 respectively, that is, they have the same acronyms.

Table 1 compares the databases of the T series and A series in terms of sense length (average senses), the number of abstracts for graph building, and Min/Max senses per term, etc. We can observe from Table 1, that the number of abstracts for building graphs is much smaller than that of the Bio-Concept-Graph method (BCG) [5] introduced in Section II. And the average number of possible senses in the T serials of acronym datasets is higher than that in the A serials of datasets, except T100.

In our experiment on acronym disambiguation, it is found that the accuracy of disambiguation on the acronyms with two senses is much higher than that with three senses. In our opinion, the test dataset with a fixed number of senses for each ambiguous word is of great significance to the comparative experiments. So T100 and T150 test datasets, as well as T200 and T300, for the acronym WSD have been made freely available for research and may be obtained from *www.ketrc.com* and *www.condillac.org*.

V. EVALUATION (RESULTS AND DISCUSSION)

In this section, the experimental result obtained by this approach is described and compared with other state-of-the-art systems. Further experiments are also performed and analyzed.

The performance in all experiments is measured by accuracy, precision, F1, and recall. Specifically, accuracy is the percentage of instances correctly disambiguated, that is the number of correctly disambiguated instances divided by the total number of instances in the test collection. Macro-P, macro-R, and macro-F1, as shown in (12) to (16), are applied in our experiments, for some ambiguities have multiple senses.

$$Macro\_P = \frac{1}{n}\sum_{i=1}^{n} P_i \tag{12}$$

$$Macro\_R = \frac{1}{n}\sum_{i=1}^{n} R_i \tag{13}$$

TABLE 2. Performance of WSD on the NLM-WSD test collection of 49 ambiguities.

WCOTG	Macro_Acc.	Macro_Prec.	Macro_F1	Macro_Rec.
BioBERT_L	0.8307	0.8291	0.8291	0.8307
BlueBERT_B	0.8325	0.8217	0.8310	0.8324
COP	0.8589	0.8435	0.8523	0.8589

TABLE 3. Comparative results (accuracy) for state-of-the-art methods and the ones reported in this work (WCOTG, UWCOTG).

	Accuracy
MRD	0.6389
AEC	0.6836
WCOTG	
COP	0.8589
BlueBERT_B	0.8325
BioBERT_L	0.8307
UWCOTG	0.8270
BCG	0.7516
MFS	0.8471

$$Macro\_F = \frac{2 * Macro\_P * Macro\_R}{Macro\_P + Macro\_R} \tag{14}$$

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{15}$$

$$R_i = \frac{TP_i}{TP_i + FN_i} \tag{16}$$

For the modified Personalized PageRank (PPR) algorithm used for disambiguation, the damping factor is set as 0.85 and the number of iterations is set as 100. For the BERT-based models used for calculating the term embeddings, the version of BioBERT-Large v1.1 and BlueBERT-Base are chosen and shortened to "BioBERT_L" and "BlueBERT_B" respectively in our experiments.

A. EVALUATION ON THE NLM-WSD TEST COLLECTION OF 49 AMBIGUITIES

In order to verify the effect of edge weights on system accuracy, we carry out a comparative experiment on the weighted bio-term graph and the unweighted bio-term graph respectively. And then we compare them with several other state-of-the-art knowledge-based methods, which all have evaluations over the NLM-WSD Test Collection of 49 ambiguities. This can lead to a more objective comparison.

The "COP" (co-occurrence probability) refers to the proposed method that the probability of the co-occurrence terms is taken as the weight for the WCOTG, while the next two "BioBERT_L" and "BlueBERT_B" refer to that in which semantic-relatedness between two terms (BERT-based vectors) is used as the weight in the WCOTG. The MRD (machine-readable dictionary) [1] is an unsupervised vector approach. It chooses the concept whose feature vector is the closest to the instance vector for the target ambiguous word. The AEC (Automatic Extracted Corpus) [1] method trains a Naïve Bayes classifier for WSD on the automatically retrieved citations from PubMed. The

TABLE 4. Word-by-word comparative results over the NLM-WSD test collection.

NLM-WSD Words	MRD	AEC	BCG (Base-line)	MFS (Baseline)	WCOTG			UW-TCG
					Bio-BERT_L	Blue-BERT_B	COP	
Adjust-ment	0.2308	0.6237	0.6882	0.6667	0.6989	0.6989	**0.7609**	0.7283
Blood pressure	0.4343	0.3700	**0.5000**	**0.5300**	0.4400	0.4800	0.4800	0.4200
Cold	0.6044	0.3895	0.7579	**0.9053**	0.6882	0.6882	0.7789	0.7053
Condition	0.3370	0.7065	**0.9783**	**0.9783**	0.7391	0.7500	0.6630	0.7283
Culture	0.8200	0.6000	0.9500	0.8900	**0.9900**	**0.9900**	**0.9900**	0.9900
Degree	0.4923	0.8923	**0.9692**	**0.9692**	0.9231	0.9538	0.8923	0.8769
Depression	0.9941	0.9529	**1.0000**	**1.0000**	0.7882	0.8118	0.9529	0.9529
Deter-mination	0.9936	0.1392	0.9494	**1.0000**	**1.0000**	**1.0000**	**1.0000**	1.0000
Discharge	**0.9861**	0.7067	0.8400	**0.9867**	0.8533	0.8267	0.8933	0.8133
Energy	0.4536	0.4000	0.8200	**0.9900**	0.7600	0.7700	0.8500	0.7500
Evaluation	0.5800	0.5000	0.5000	0.5000	0.7000	0.6900	**0.7500**	0.6800
Extraction	0.2907	0.7471	0.8621	0.9432	0.9195	**0.9655**	0.9425	0.9310
Failure	0.5862	**0.8621**	0.1379	**0.8621**	0.4828	0.5172	0.4828	0.4828
Fat	0.9718	0.8356	0.0274	**0.9726**	0.7945	0.6986	0.8767	0.7945
Fit	0.8387	0.8889	**1.0000**	**1.0000**	0.8889	0.8889	**1.0000**	0.8889
Fluid	0.6082	0.4800	0.8600	**1.0000**	0.8700	0.8800	0.9900	0.9100
Frequency	0.9362	0.6064	**1.0000**	**1.0000**	0.9894	0.9894	0.8617	0.7021
Ganglion	0.9565	0.8600	0.9300	0.9300	0.9700	**0.9800**	0.9600	0.9800
Glucose	0.2755	0.7800	**0.9100**	**0.9100**	0.9000	0.9000	0.9000	0.9000
Growth	0.6700	0.3700	0.6200	0.6300	0.7600	0.7500	**0.8000**	0.7200
Immunosu p-pression	0.4896	0.5700	**0.7300**	0.5800	0.6300	0.6600	0.7000	0.6300
Implanta-tion	0.8316	**0.9490**	0.8673	0.8265	0.7959	0.8163	0.8673	0.8469
Inhibition	0.9697	0.8384	**0.9899**	**0.9899**	**0.9899**	**0.9899**	**0.9899**	0.9899
Japanese	0.9211	0.6329	0.9241	**0.9367**	0.8608	0.8481	0.8608	0.8481
Lead	0.3793	0.8276	0.9310	0.9310	0.9310	0.9310	**0.9655**	0.9310
Man	0.3187	0.6522	0.6413	0.6304	0.5978	0.5543	**0.6630**	0.6630
Mole	0.8916	0.4405	**0.9881**	**0.9881**	0.7738	0.7024	0.8095	0.7857
Mosaic	0.5795	**0.8144**	0.4639	0.5360	0.7526	0.7732	0.7629	0.7629
Nutrition	0.3933	0.3708	0.2697	**0.5056**	0.3820	0.3933	0.4719	0.4270
Pathology	0.3939	0.6061	0.1717	**0.8586**	0.8182	0.8182	0.8384	0.8283
Pressure	0.9836	0.5208	0.9688	**1.0000**	**1.0000**	**1.0000**	0.9913	0.9914
Radiation	0.6979	0.7449	0.6224	0.6122	0.7347	**0.7551**	**0.7551**	0.7245
Reduction	0.8182	0.9091	0.7273	0.8182	**1.0000**	**1.0000**	0.9091	1.0000
Repair	0.8358	0.8529	0.8971	0.7647	0.8824	0.8971	**0.9265**	0.8824
Resistance	0.3333	**1.0000**	0.0000	**1.0000**	**1.0000**	**1.0000**	**1.0000**	1.0000
Scale	0.0615	0.7231	0.9846	**1.0000**	0.9692	0.9692	0.9846	0.9692
Secretion	0.3535	0.4600	**0.9900**	**0.9900**	0.7900	0.8400	0.9100	0.8300
Sensitivity	0.8431	0.7255	**0.9608**	**0.9608**	0.8627	0.8824	0.8627	0.8627
Sex	0.5455	0.6000	**0.8900**	0.8000	0.8600	0.8600	0.8600	0.8500
Single	0.0400	0.8900	0.9600	**0.9900**	0.9800	0.9800	0.9700	0.9800
Strains	0.9780	0.9570	0.9785	0.9892	0.9892	0.9892	**1.0000**	0.9785
Support	0.3000	**1.0000**	0.2000	0.8000	0.9000	0.9000	0.8700	0.7600
Surgery	0.9394	0.1900	0.7800	**0.9800**	0.7100	0.7100	0.8500	0.7600
Transient	**0.9900**	0.9100	**0.9900**	**0.9900**	**0.9900**	**0.9900**	**0.9900**	0.9900
Transport	0.9780	**1.0000**	0.9894	0.9894	0.9787	0.9787	0.9787	0.9787
Ultrasound	0.6667	0.7400	0.8400	0.8400	0.8500	0.8100	**0.8900**	0.8100
Variation	0.7600	0.6900	0.8100	0.8000	0.8700	0.8700	**0.8800**	0.8700
Weight	0.4717	0.6604	0.3208	0.5472	**0.9057**	0.8679	0.8679	0.9057
White	0.4831	0.5111	0.6444	0.5444	0.7444	0.7778	**0.8333**	0.7111
Accuracy	**0.6389**	**0.6836**	**0.7516**	**0.8471**	**0.8307**	**0.8325**	**0.8589**	**0.8270**

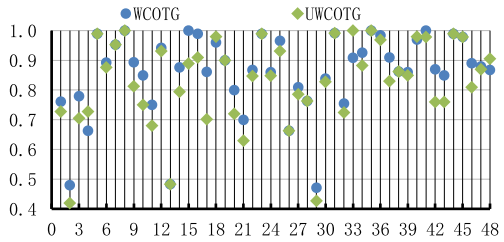

FIGURE 2. Evaluation of the accuracy between system WCOTG and UWCOTG.

Bio-Concept-Graph method (BCG) [5] introduced a technique based on co-occurrence concept graphs for performing WSD in the biomedical domain. Wherein, the UMLS Metamap was used to map terms from the created knowledge base onto biomedical concepts for graph building, and the PPR algorithm was used for WSD then. Although it achieved a good result, it did not account for the links' weights in the graph. In the evaluation, the MFS (Most Frequency Sense) that is standard in WSD evaluation can be considered as a supervised baseline. And the unsupervised Bio-Concept-Graph (BCG) method also on co-occurrence graph ideas is used as another baseline.

Table 2 shows the performance of WSD on the NLM-WSD test collection of 49 ambiguities with three modes for calculating the weights in the graph. And Table 3 shows the comparative results with other knowledge-based methods.

Table 3 shows that on the NLM-WSD Test Collection of 49 ambiguities the proposed WCOTG (Weighted Co-occurrence bio-Term Graph), all 3 kinds of weighting, outperforms MRD and AEC. It is also better than the baseline Bio-Concept-Graph (BCG) method based on similar co-occurrence ideas. The average result is near the baseline MFS. From the table, we can see that the accuracy of the method UWCOTG (Non-Weighted Co-occurrence bio-Term Graph) is lower than that of the baseline MFS. But it achieves better accuracy over the baseline Bio-Concept-Graph (BCG) method. The former is based on the co-occurrence term graph and the latter is on the co-occurrence concept graph, both unweighted.

Table 4 shows a word-by-word comparison between the WCOTG and others listed in Table 3. As we can observe in the table, the COP-based WCOTG offers a large number of best results, the best result in 17 out of 49 cases. Relatively, the baseline MFS offers the best result in 27 cases, the BlueBERT_B-based WCOTG 10, the BioBERT_L-based WCOTG 8, the MRD 2, the AEC 6, and the BCG 14.

The COP-based WCOTG outperforms the baseline MFS in 25 results (including 5 equal cases), and both the BlueBERT_B-based WCOTG and BioBERT_L-based WCOTG in 23 results (including 7 equal cases) respectively. The COP-based WCOTG outperforms the baseline BCG in 37 results (including 5 equal cases), the BlueBERT_B-based WCOTG in 28 results (including 4 equal cases), and the BioBERT_L-based WCOTG in 29 results (including 3 equal cases).

TABLE 5. Comparative results (overall accuracy) of WSD on the MSH-WSD dataset - 203 ambiguities.

	Accuracy
UWCOTG	0.7728
WCOTG	
COP	0.8054
BioBERT_L	0.8097
BlueBERT_B	0.7845
BERT-based method (Mao, Y. et al., 2020) [6]	
BERT_L	0.7530
BlueBERT_L	0.7960
BlueBERT_LE	0.8050
BioWordVec	0.7840

TABLE 6. Performance of WSD on the MSH-WSD test collection.

WCOTG	Macro_Acc.	Macro_Prec.	Macro_F1	Macro_Rec.
BlueBERT_B	0.7845	0.7833	0.7834	0.7861
BioBERT_L	0.8097	0.8011	0.8010	0.8097
COP	0.8054	0.8021	0.8033	0.8106

Fig.2 shows a comparison between the UWCOTG and WCOTG (COP-based). From Fig.2 and Table 4, it can be seen that the WCOTG performs better than the UWCOTG. Specifically, the WSD of the COP-based performs better in 31 out of 49 cases than that of the UWCOTG, while the BioBERT_B-based WCOTG better in 22 results, and the BioBERT_L-based WCOTG better in 17 results. The UWCOTG offers the best result in 14 out of 49 cases in the absence of the WCOTG.

B. EVALUATION ON THE MSH-WSD TEST COLLECTION OF 203 AMBIGUITIES

We also evaluate our method on the MSH-WSD Test Collection (203 ambiguous words) and compare experimental results with the work of Mao and Wah [6], which applies BERT variant models for WSD in the unsupervised biomedical domain and using the same test dataset.

As we can observe in Table 5, over the same test collection, without extra training, the BioBERT_L method outperforms the best BlueBERT_LE-based method that enhanced the BlueBERT_Large model with additional training by the concept definitions in the UMLS [6]. While the COP-based WCOTG obtain similar disambiguation results to it. We can also see that the accuracy of WCOTG (weighted method) based on two kinds of weights currently tried both outperform the UWCOTG (unweighted method). Table 6 shows the performance (Macro_Acc., Macro_Prec., Macro_F1., and Macro_Rec.) of the proposed WCOTG method for WSD on the MSH-WSD test collection of 203 ambiguities with three modes for calculating the weights in the graph.

TABLE 7. Comparative results (accuracy) over the T and A serial of test datasets.

	Term-based		Concept-based
	Weighted	No-weight	No-weight
Accuracy	WCOTG	UWCOTG	BCG
T100 /A100	**0.9939**	0.9867	0.8278
T200 /A200	**0.9784**	0.8681	0.8006
T300 /A300	**0.8983**	0.8062	0.8257
Average	**0.9569**	0.8870	0.8180

TABLE 8. Word-by-word comparative results obtained by the WCOTG and BCG over the T and A serial of test datasets.

Acronyms	T100	A100	T200	A200	T300	A300
	WCOTG	BCG	WCOTG	BCG	WCOTG	BCG
ANA	**0.9900**	0.7800				
APC	**0.9800**	**0.9800**	**1.0000**	0.9650	0.9467	**0.9600**
BPD	**1.0000**	0.9700	**0.9900**	0.9800	**0.9867**	0.9767
BSA	**1.0000**	0.9400	**0.9500**	0.9100	**1.0000**	0.9100
CAT	**0.9700**	0.9500	**1.0000**	0.9500	0.7967	**0.9367**
CML	**1.0000**	0.9200	**0.9650**	0.9300	0.8000	**0.9300**
CMV	**1.0000**	0.9800	**1.0000**	0.9850	0.9067	**0.9900**
DIP	**1.0000**	0.9600	**0.9800**	0.9600		
EMG	**1.0000**	0.1200	**0.9950**	0.1150	**0.9933**	0.1167
FDP	**0.9900**	0.9500				
LAM	**0.9900**	0.9600	**0.9700**	0.9650	**0.9633**	0.9533
MAC	**0.9800**	0.6400	**0.9650**	0.6550	**0.9767**	0.64333
MCP	**1.0000**	0.6000	**0.9950**	0.6150	**0.9700**	0.6267
PCA	**1.0000**	0.9700	**1.0000**	0.9749	**0.9833**	0.9766
PCP	**1.0000**	0.9900	**1.0000**	0.5800	**0.8100**	0.5767
PEG	**1.0000**	**1.0000**	0.8650	**1.0000**	0.6967	**0.9967**
PVC	**0.9900**	0.2300	**0.9850**	0.2500		
RSV	**1.0000**	0.9600	**0.9950**	0.9750	0.7467	**0.9667**
Accuracy all	**0.9939**	0.8278	**0.9784**	0.8006	**0.8983**	0.8257
TN_ graph	21075	50143	17836	50143	15064	50143

TABLE 9. Comparative performance obtained by UWCOTG and WCOTG.

	T100 (2 senses per term)		T150 (3 senses per term)		T200 (2-4 senses per term)		T300 (2-4 senses per term)	
	UWCOTG	WCOTG	UWCOTG	WCOTG	UWCOTG	WCOTG	UWCOTG	WCOTG
Macro_Acc	0.9867	**0.9939**	0.8372	**0.9602**	0.8681	**0.9784**	0.8062	**0.8983**
Macro_Pre	0.9869	**0.9932**	0.8492	**0.9599**	0.8332	**0.9666**	0.7861	**0.8842**
Macro_F1	0.9866	**0.9933**	0.8398	**0.9579**	0.8403	**0.9682**	0.7924	**0.8915**
Macro_Rec	0.9864	**0.9994**	0.8470	**0.9564**	0.8724	**0.9698**	0.8172	**0.9012**

C. EVALUATION ON ACRONYM TEST DATASETS

On the built acronym test datasets, we verify the difference in disambiguation accuracy between the term-graph method and concept-graph one both based on the co-occurrence graph ideas. Table 7 shows the experimental results of the WCOTG and UWCOTG methods based on the co-occurrence term graph on the T serial test datasets against the Bio-Concept-Graph (BCG) [5] method based on the co-occurrence concept graph on the parallel A serials test datasets. It can be seen from Table 7 that the WCOTG (BioBert-based) outperforms others over all test datasets listed in the table.

Table 8 shows a word-by-word comparison result of the WCOTG and Bio-Concept-Graph (BCG) methods. As we can observe in the table, over the test dataset with 100 instances, the WCOTG offers all the best results (including 2 equal cases), and the average accuracy of WCOTG reaches 0.9939. Over the test dataset with 200 instances, the WCOTG offers almost all the best results, except in the case of the acronym "PEG". And over the test dataset with 300 instances, the WCOTG offers the best result in 8 out of 14 cases while the BCG offers 6, and the results are basically flat. However, the total number (TN_graph) of abstracts used for building graphs varies greatly, with a maximum number of 21075 for the WCOTG-based system and 50143 for the Bio-Concept-Graph based one. In other words, our method only uses almost half the composition level relative to the Bio-Concept-Graph based method.

TABLE 10. Word-by-word comparative results of the UWCOTG and WCOTG.

	T100 (2 senses per term)		T150 (3 senses per term)		T200 (2-4 senses per term)		T300 (2-4 senses per term)	
	UWCOTG	WCOTG	UWCOTG	WCOTG	UWCOTG	WCOTG	UWCOTG	WCOTG
ANA	0.9900	**0.9900**						
APC	0.9600	**0.9800**	0.7733	**0.9467**	0.8250	**1.0000**	0.7700	**0.9467**
BPD	**1.0000**	**1.0000**			0.9750	**0.9900**	0.9867	**0.9867**
BSA	0.9900	**1.0000**	0.7467	**0.9400**	0.7750	**0.9500**	0.8367	**1.0000**
CAT	**0.9700**	**0.9700**	0.9533	**0.9733**	0.9450	**1.0000**	0.7100	**0.7967**
CML	**1.0000**	**1.0000**	0.6800	**0.9733**	0.6900	**0.9650**	0.5567	**0.8000**
CMV	**1.0000**	**1.0000**	0.9346	**0.9439**	0.9700	**1.0000**	0.8867	**0.9067**
DIP	**1.0000**	**1.0000**	0.1028	**0.9533**	0.5100	**0.9800**		
EMG	**1.0000**	**1.0000**			**0.9950**	**0.9950**	0.9933	**0.9933**
FDP	**0.9900**	**0.9900**	0.9133	**0.9267**				
LAM	**0.9900**	**0.9900**	**0.9519**	**0.9519**	0.9600	**0.9700**	0.9333	**0.9633**
MAC	0.8800	**0.9800**	0.8381	**0.9333**	0.8400	**0.9650**	0.9633	**0.9767**
MCP	**1.0000**	**1.0000**	0.9533	**0.9667**	0.9650	**0.9950**	0.9300	**0.9700**
PCA	**1.0000**	**1.0000**	0.9533	**0.9600**	0.8050	**1.0000**	**0.9867**	0.9833
PCP	**1.0000**	**1.0000**	0.9667	**0.9733**	0.7750	**1.0000**	0.4467	**0.8100**
PEG	**1.0000**	**1.0000**	0.9800	**1.0000**	**0.8900**	0.8650	**0.7267**	0.6967
RSV	0.9900	**1.0000**			**0.9950**	**0.9950**	0.5600	**0.7467**
PVC	**1.0000**	0.9900	0.9733	**1.0000**	0.9750	**0.9850**		
Accuracy all	0.9867	**0.9939**	0.8372	**0.9602**	0.8681	**0.9784**	0.8062	**0.8983**

Table 9 shows the concrete performance of the UWCOTG and WCOTG methods - both based on bio-term graphs - over the T serial acronym test datasets, and Table 10 shows a word-by-word comparison of results. As we can see from Table 10, over the T100 dataset with 2 senses per acronym, the WCOTG offers the best result in 17 out of 18 cases, including 12 equal cases with the UWCOTG. Over the T150 dataset with 3 senses per acronym, the WCOTG offers all the best results, only including 1 equal case. The efficiency gap between the UWCOTG and WCOTG systems is only 0.7 ((0.9939-0.9867) * 100%) percentage points over the T100 with 2 senses per acronym. However, over the T150 with 3 senses per acronym, the efficiency gap between the two becomes 12.3 ((0.9602-0.8372)*100%) percentage points. The WCOTG outperforms the UWCOTG in 15 results over the T200 dataset, and 12 results over the T300 dataset. In all, the WCOTG system performs much better than the UWCOTG system on the test dataset with 2 more senses per acronym.

TABLE 11. Comparative results of the WCOTG and other systems based on BERT models.

	Accuracy	Test Set	Number	Domain
WCOTG		NLM-WSD dataset	3983	Biomedicine
BioBERT_L	0.8307			
BlueBERT_B	0.8325			
WCOTG		MSH-WSD dataset	37888	
BioBERT_L	0.8097			
BlueBERT_B	0.7845			
BERT-based method [6]				
BERT_L	0.7530			
BlueBERT_L	0.7960			
BlueBERT_LE	0.8050			
BioWordVec	0.7840			
WCOTG (BioBERT_B)	0.9939	T100/A100	1800	
	0.9784	T200/A200	3105	
	0.8983	T300/A300	4200	
GlossBert [25]	0.7610	SemEval 13	1644	General
	0.8040	SemEval 15	1022	
BEM [26]	0.7970	SemEval 13	1644	
	0.8170	SemEval 15	1022	
EWISE [30]	0.6940	SemEval 13	1644	
	0.7450	SemEval 15	1022	
MMS_{2348} (BERT) [28]	0.7500	SemEval 13	1644	
	0.7700	SemEval 15	1022	

D. COMPARATION OTHER WORKS BASED ON BERT MODELS

As illustrated above, our system uses the Unified Medical Language System (UMLS) Knowledge Sources in the biomedical field for building term graphs and has completed the experimental comparison on the NLM-WSD, MSH-WSD, and acronym test collections in the biomedical field.

In recent years, there have been some reports in the general domain that BERT models have been well applied in WSD tasks, but their test sets usually come from the general domain. In this part, we summarize and compare the efficiency of some BERT-based WSD systems both in the general domain and biomedical domains.

Table 11 shows the comparative results. Wherein, in the general domain, the GlossBert [25] fine-tunes the pre-trained BERT model on the training corpus. The BEM [26] presents a bi-encoder model built on top of BERT that is designed

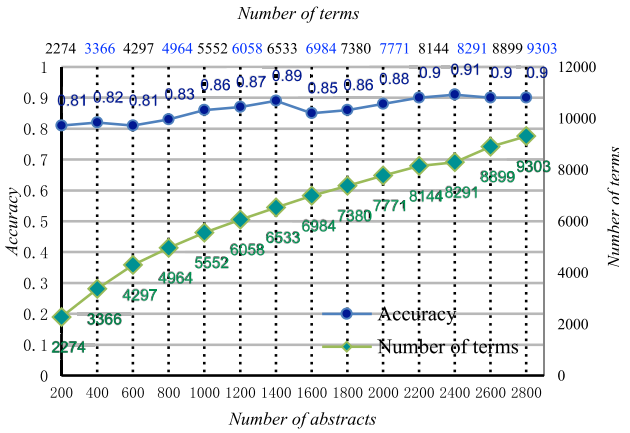

FIGURE 3. Four-axis line chart.

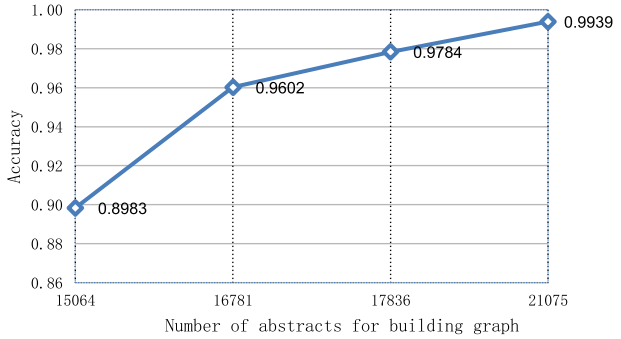

FIGURE 4. Two-axis line chart.

to improve performance on rare and zero-shot senses. The EWISE [30] uses pre-trained BERT word vectors as input embeddings. The MMS_{2348} (BERT) [28] uses the BERT model to create sense embeddings for all senses in WordNet in the WSD task.

As shown in the table, the WCOTG has achieved good results in the collection of acronyms in the biomedical field. On 1800 acronym test cases, the WSD accuracy can reach 99%. Specific experimental results can refer to in Table 7 to Table 10. Over the same test collection MSH-WSD, without extra training, the WCOTG based on BioBERT_L performs slightly better than the best BlueBERT_LE-based method [6].

When measuring system efficiency, the size of the test set is an important factor. On the one hand, we can see from Table 11 that as the number of instances in the test set increases, the system efficiency under the same method decreases. The WCOTG based on the domain BERT model on the NLM-WSD set has better accuracy than on the MSH-WSD set. Disambiguation models in general domains also have the same characteristics. It can be seen from the table that the accuracy of the GlossBert, BEM, EWISE, and MMS methods on the SemEval 15 test set is generally higher than that on the SemEval 13. On the other hand, it can be seen from Table 11 that the accuracy of the WCOTG on NLM-WSD and NLM-WSD test sets, as well as the accuracy of the GlossBert and BEM in general fields, can all exceed

80%. However, the test cases used in the biomedical field are far greater than those in the general domain.

E. FURTHER EXPERIMENT-SCALE OF CORPUS FOR BUILDING GRAPH

A co-occurrence graph is created based on the terms extracted from the abstracts of biomedical documents. So a question arises: will the size of the abstract corpus affect the result of WSD?

Fig.3, a four-axis line chart based on the further experiment on the NLM-WSD Test Collection of 49 ambiguities, shows the evolution of the relations among the system accuracy, the average number of biomedical abstracts for each ambiguous target term, and the number of terms extracted from the abstracts and used for building the co-occurrence graph. We can see that, as the number of abstracts that are used to build a co-occurrence graph for an ambiguous term increases, the number of terms extracted from the abstracts increases subduedly on the whole. We also find that number of abstracts (around 800) selected for each target term in the previous experiment, introduced in Section IV-A, is not optimal, although that number is good for the comparison with a similar system of Bio-Concept-Graph (BCG) [5]. The size of the corpus will affect the efficiency of the system to a certain extent. The overall accuracy increases with the number of abstracts and terms used to build the graph, but the performance for each WCOTG reaches a plateau beginning from around the point of 2, 200 abstracts and 8, 000 terms.

Fig. 4, a two-axis line chart is built based on the analysis of the experiment on the built acronym corpus. The statistical data - the number of abstracts for building term graphs and system accuracy- is from Table 1 and Table 7. It can be found that the disambiguation accuracy is higher and higher with the increasing number of abstracts for building term graphs. As introduced in Section V-C and shown in Table 1 and Table 8, although the number of abstracts for building term graphs is much less than that used in [5] for building concept graphs, the system of WCOTG has achieved better results in WSD on acronym test set.

VI. CONCLUSION AND FUTURE WORK

This paper describes the application of an unsupervised method based on weighted co-occurrence term graph for performing WSD in the biomedical domain. We build graphs from terms extracted from biomedical abstracts directly, instead of concepts gotten through the mapping tools. This avoids the bottleneck of mapping terms to concepts when transforming polysemous terms in biomedical documents into concepts. The most important contributions of this paper are that we add weights to the links of the built bio-term graphs, and take them as important factors in the disambiguation algorithm. Furthermore, a sub-term graph with a relatively small amount of data is extracted from the general one, to reduce the operation scale of the disambiguation algorithm. And the PPR algorithm which is normally applied over the unweighted directed graph is modified for the WSD task over our built weighted undirected term graph.

The paper provides comparative evaluations of similar previous methods and several WSD methods over the NLM-WSD, MSH-WSD test datasets, and an acronym test dataset. The results show that the proposed unsupervised WCOTG method outperforms the concept-graph based one on a very similar corpus scale, and outperforms some unsupervised ones addressing the same problem. Comparative experiments of weighted and non-weighted bio-term graphs validate the positive effect of links' weight on disambiguation efficiency. Further statistical experiments on the relationship between system accuracy, the numbers of biomedical abstracts in the corpus, and the corresponding extracted terms suggest an excellent minimum corpus scale when resources are limited.

Although our method achieved good overall disambiguation efficiency on the three datasets tested, there were also some failed cases in the experiment. For example, the 79th test case of the word 'Japanese' in the MSH-WSD test set is as follows: ''A Japanese experience''. After removing the stop word and the word 'Japanese' that needs to be disambiguated, only one valid word 'experience' remains. This kind of too-short test case often leads to testing failure on this case. How to effectively disambiguate the semantics of words in ultra-short sentences is a problem that our future research will address.

Currently, in the proposed unsupervised WCOTG, cosine similarity theory is applied in measuring the semantic relatedness between terms when weighing edges in the built term graph. However, there are also many other association measures that are worth trying to generate weights on the constructed term graph. In the future, we plan to explore the use of new methods to calculate the weight of edges in the co-occurrence term graph and realize corresponding comparative experiments. In addition, we would like to construct a term graph based on the combination of UMLS Metathesaurus and self-created biomedical corpus and explore hybrid artificial intelligence approaches for WSD.

REFERENCES

[1] A. J. Jimeno-Yepes, B. T. McInnes, and A. R. Aronson, ''Exploiting MeSH indexing in medline to generate a data set for word sense disambiguation,'' *BMC Bioinf.*, vol. 12, no. 1, Dec. 2011.

[2] R. Navigli, ''Word sense disambiguation: A survey,'' *ACM Comput. Surv.*, vol. 41, no. 2, pp. 1–8, 2009.

[3] A. J. Jimeno-Yepes and A. R. Aronson, ''Knowledge-based biomedical word sense disambiguation: Comparison of approaches,'' *BMC Bioinf.*, vol. 11, no. 1, p. 569, Dec. 2010.

[4] M. Sarrouti and S. Ouatik El Alaoui, ''SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions,'' *Artif. Intell. Med.*, vol. 102, Jan. 2020, Art. no. 101767.

[5] A. Duque, M. Stevenson, J. Martinez-Romo, and L. Araujo, ''Co-occurrence graphs for word sense disambiguation in the biomedical domain,'' *Artif. Intell. Med.*, vol. 87, pp. 9–19, May 2018.

[6] Y. Mao and F. K. Wah, ''Use of word and graph embedding to measure semantic relatedness between unified medical language system concepts,'' *J. Amer. Med. Inform. Assoc.*, no. 10, p. 10, 2020.

[7] M. Bevilacqua and R. Navigli, "Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Dec. 2020, pp. 2854–2864.
[8] A. Jimeno Yepes, "Word embeddings and recurrent neural networks based on long-short term memory nodes in supervised biomedical word sense disambiguation," *J. Biomed. Informat.*, vol. 73, pp. 137–147, Sep. 2017.
[9] Z. Li, F. Yang, and Y. Luo, "Context embedding based on BI-LSTM in semi-supervised biomedical word sense disambiguation," *IEEE Access*, vol. 7, pp. 72928–72935, 2019.
[10] A. Raganato, C. Delli Bovi, and R. Navigli, "Neural sequence learning models for word sense disambiguation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1–8.
[11] E. Barba, T. Pasini, and R. Navigli, "ESC: Redesigning WSD with extractive sense comprehension," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 4661–4672.
[12] A. Pesaranghader, S. Matwin, M. Sokolova, and A. Pesaranghader, "DeepBioWSD: Effective deep neural word sense disambiguation of biomedical text data," *J. Amer. Med. Inform. Assoc.*, vol. 26, no. 5, pp. 438–446, May 2019.
[13] T. Pasini, "The knowledge acquisition bottleneck problem in multilingual word sense disambiguation," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020., pp. 1–8.
[14] D. Yuan, J. Richardson, R. Doherty, C. Evans, and E. Altendorf, "Semi-supervised word sense disambiguation with neural models," 2016, *arXiv:1603.07012*.
[15] J. Lau, P. Cook, D. Mccarthy, D. J. Newman, and T. Baldwin, "Word sense induction for novel sense detection," in *Proc. 13th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2012, pp. 765–876.
[16] S. Henry, C. Cuffy, and B. Mcinnes, "Evaluating feature extraction methods for knowledge-based biomedical word sense disambiguation," in *Proc. BioNLP*, 2017, pp. 272–281. [Online]. Available: https://www.researchgate.net/publication/318739041_Evaluating_Feature_Extraction_Methods_for_Knowledge-based_Biomedical_Word_Sense_Disambiguation, doi: 10.18653/v1/W17-2334.
[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[18] D. Loureiro, K. Rezaee, M. T. Pilehvar, and J. Camacho-Collados, "Analysis and evaluation of language models for word sense disambiguation," *Comput. Linguistics*, vol. 47, no. 2, pp. 387–443, 2020.
[19] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.
[20] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets," 2019, *arXiv:1906.05474*.
[21] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, "Publicly available clinical BERT embeddings," 2019, *arXiv:1904.03323*.
[22] A. Jimeno Yepes and R. Berlanga, "Knowledge based word-concept model estimation and refinement for biomedical text mining," *J. Biomed. Informat.*, vol. 53, pp. 300–307, Feb. 2015.
[23] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT sentence embedding," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 878–891.
[24] J. Philip Wahle, T. Ruas, N. Meuschke, and B. Gipp, "Incorporating word sense disambiguation in neural language models," 2021, *arXiv:2106.07967*.
[25] L. Huang, C. Sun, X. Qiu, and X. Huang, "GlossBERT: BERT for word sense disambiguation with gloss knowledge," 2019, *arXiv:1908.07245*.
[26] T. Blevins and L. Zettlemoyer, "Moving down the long tail of word sense disambiguation with gloss-informed biencoders," 2020, *arXiv:2005.02590*.
[27] P.-H. Chi, P.-H. Chung, T.-H. Wu, C.-C. Hsieh, Y.-H. Chen, S.-W. Li, and H.-Y. Lee, "Audio Albert: A lite BERT for self-supervised learning of audio representation," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 156–160.
[28] D. Loureiro and A. Jorge, "Language modelling makes sense: Propagating representations through wordnet for full-coverage word sense disambiguation," 2019, *arXiv:1906.10007*.
[29] L. Vial, B. Lecouteux, and D. Schwab, "Sense vocabulary compression through the semantic knowledge of wordnet for neural word sense disambiguation," 2019, *arXiv:1905.05677*.
[30] S. Kumar, S. Jat, K. Saxena, and P. Talukdar, "Zero-shot word sense disambiguation using sense definition embeddings," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 5670–5681.
[31] N. Q. K. Le, Q.-T. Ho, V.-N. Nguyen, and J.-S. Chang, "BERT-promoter: An improved sequence-based predictor of DNA promoter using BERT pre-trained model and SHAP feature selection," *Comput. Biol. Chem.*, vol. 99, Aug. 2022, Art. no. 107732.
[32] E. Qthd, "Deep transformers and convolutional neural network in identifying DNA N6-methyladenine sites in cross-species genomes," *Methods*, vol. 204, pp. 199–206, Dec. 2021.
[33] B. Biseda, G. Desai, H. Lin, and A. Philip, "Prediction of ICD codes with clinical BERT embeddings and text augmentation with label balancing using MIMIC-III," 2020, *arXiv:2008.10492*.
[34] M. Andriy, U. Ozlem, and M. I. Bridget, "MT-clinical BERT: Scaling clinical information extraction with multitask learning," *J. Amer. Med. Inform. Assoc.*, vol. 28, no. 10, pp. 2108–2115, Oct. 2021. [Online]. Available: https://academic.oup.com/jamia/article-abstract/28/10/2108/6333354?redirectedFrom=fulltext&login=false, doi: 10.1093/jamia/ocab126.
[35] A. Duque, L. Araujo, and J. MartinEZ-ROMO, "CO-graph: A new graph-based technique for cross-lingual word sense disambiguation," *Natural Lang. Eng.*, vol. 21, no. 5, pp. 743–772, Nov. 2015.
[36] Y. Wang, M. Wang, and H. Fujita, "Word sense disambiguation: A comprehensive knowledge exploitation framework," *Knowl.-Based Syst.*, vol. 190, Feb. 2020, Art. no. 105030.
[37] Y. Gutiérrez, S. Vázquez, and A. Montoyo, "Spreading semantic information by word sense disambiguation," *Knowl.-Based Syst.*, vol. 132, pp. 47–61, Sep. 2017.
[38] M. Stevenson and Y. Guo, "Disambiguation of ambiguous biomedical terms using examples generated from the UMLS metathesaurus," *J. Biomed. Informat.*, vol. 43, no. 5, pp. 762–773, Oct. 2010.
[39] B. T. Mcinnes, T. Pedersen, L. Ying, S. V. Pakhomov, and G. B. Melton, "Using second-order vectors in a knowledge-based method for acronym disambiguation," in *Proc. 15th Conf. Comput. Natural Lang. Learn.*, Jun. 2011, pp. 145–153.
[40] S. Henry, C. Cuffy, and B. Mcinnes, "Evaluating feature extraction methods for knowledge-based biomedical word sense disambiguation," in *Proc. BioNLP*, 2017, pp. 345–456.
[41] H. Liu, Y. A. Lussier, and C. Friedman, "Disambiguating ambiguous biomedical terms in biomedical narrative text: An unsupervised method," *J. Biomed. Informat.*, vol. 34, no. 4, pp. 249–261, Aug. 2001.

ZHENLING ZHANG received the Ph.D. degree in computer science and technology from Grenoble Alpes University, France, in 2019. She is currently an Associate Professor with the Computer Science and Technology School, Liaocheng University, Shandong, China. She is also a Research Member of the Condillac Research Group, Université Savoie Mont Blanc, France, and the Knowledge Engineering and Terminology Research Center (KETRC), Liaocheng University. Her research interests include natural language processing, computing terminology, and artificial intelligence.

YANGLI JIA (Member, IEEE) received the Ph.D. degree in computer science and technology from Beihang University, Beijing, China, in 2010. He is currently a Professor with the Computer Science and Technology School, Liaocheng University, Shandong. He is also the Vice Director of the Knowledge Engineering and Terminology Research Center (KETRC), Liaocheng University, and a Research Member of the Condillac Research Group, Université Savoie Mont Blanc, France. He is an Expert of ISO TC37. His research interests include natural language processing, computing terminology, and terminology standardization. He is a Senior Member of the China Computer Federation.

XIANGLIANG ZHANG received the master's degree in computer science and technology from Liaocheng University, Shandong, China, in 2022. His research interests include natural language processing, computing terminology, and artificial intelligence.

MARIA PAPADOPOULOU is currently a Digital Humanist, a Lecturer with the School of Philosophy, National Kapodistrian University of Athens, Greece, and a Researcher with the Computer Science, Systems, Information and Knowledge Processing Laboratory, Condillac Research Group on Terminology and Ontology, Université Savoie Mont Blanc, France. She is also with the Knowledge Engineering and Terminology Research Centre (KETRC), Liaocheng University, China. Her research has received funding from world class funding bodies, such as Princeton University and the European Commission. She has published around 50 academic articles in internationally acclaimed academic journals and edited books and has delivered invited lectures in more than 20 countries around the globe. Her research interests include the intersection of digital humanities with semantic web standards, ontologies, knowledge graphs, and linked open data technologies. Her research aims to model information and knowledge in regard to texts and artifacts from digital cultural heritage collections, in a scalable way.

CHRISTOPHE ROCHE (Member, IEEE) is currently a Full Professor in artificial intelligence with Université Savoie Mont Blanc, France, in charge of the Condillac Research Group in "Terminology and Ontology." He is also the Head of the Knowledge Engineering and Terminology Research Center (KETRC), Liaocheng University, China. He is an ISO Expert on Terminology (Project Leader of the ISO 1087-1 Standard) and the Chairperson of the French Committee of Standardization on Terminology (AFNOR, X03A Commission). His research interests include symbolic artificial intelligence (knowledge representation and ontology) and terminology.