C-STSS: A Context -based Short Text Semantic Similarity approach applied to biomedical named entity linking*

Asma Djellal[1,2,*], Maya Souilah Benabdelhafid[1,2]

[1]Ecole Supérieure de Comptabilité et de Finance, ESCF Constantine, Algeria
[2]Lire laboratory, Abdelhamide Mehri Constantine 2 University, Constantine, Algeria

Abstract

This research paper delves into Human-Computer Interaction by investigating Knowledge Graph-based Question Answering systems in the biomedical domain. The study leverages Knowledge Graphs as potent tools to enhance Named Entity Linking in short texts, where limited context poses challenges. Conventional linking methods struggle with single Named Entity linking due to poor context and name variation issues, affecting their efficiency. To address these challenges, several scholars are working on designing Knowledge Graph-based Question Answering Systems with a focus on the name variation problem by relying on Named Entity morphological forms but they are rarely considering their semantic similarities. This paper introduces a Context-based Short Text Semantic Similarity approach for Named Entity Linking in the biomedical domain. The proposed approach improves the performance of Question Answering systems by utilizing contextual semantic similarities in short texts and combining knowledge-based and corpus-based methods for fine-grained meaning comparison, which allow addressing sparseness and vocabulary mismatches, showcasing the paper's uniqueness.

Keywords

Question Answering Systems, Natural Language Processing, Biomedical Named Entity Linking, Contextual Semantic Similarities, Short Text.

1. Introduction

In the ever-evolving landscape of Natural Language Processing (NLP), the challenge of deciphering the nuances of short texts, particularly within specialized domains like biomedicine, has emerged as a critical area of research. Short texts, encompassing brief queries and questions, lack the extensive context often found in longer texts, posing formidable obstacles for accurate Named Entity Linking (NEL), which is a key part for developing Question Answering Systems (QAS) [1]. The core difficulty lies in disambiguating Named Entities (NE) [2], especially those sharing similar surface forms [3], and capturing subtle semantic differences essential for accurate NEL.

To address these challenges, this paper pioneers a novel approach that considers the fine-grained meaning comparison by integrating knowledge-based and corpus-based methods [2]. Corpus-based methods leverage contextual information from textual data to compute general semantic relatedness between words. Meanwhile, knowledge-based methods draw upon the wealth of semantic information stored in resources like Knowledge Graphs (KG). By integrating these approaches, the study aims to overcome the sparseness and vocabulary mismatches inherent in short texts.

This paper introduces the Context-based Short Text Semantic Similarity (C-STSS) approach, a sophisticated framework that aims to bridge the gap between the limited context of short biomedical texts and the rich semantic knowledge encompassed within specialized domains. By dissecting semantic similarities and leveraging domain-specific knowledge, C-STSS provides nuanced analysis, facilitating accurate NEL even in the face of sparse and mismatched vocabulary. This innovative approach holds the promise of revolutionizing NEL within the constraints of short texts, opening new avenues for exploration at the intersection of NLP and Human-Computer Interaction (HCI).

The remainder of this paper is organized as follows. Section 2 outlines some preliminaries related to the research work. Section 3 reviews some related works and analyses drawbacks of recent biomedical NEL systems. Section 4 constitutes the bulk of the paper and presents C-STSS, our proposed approach for dealing with NEL problem in short biomedical text. Section 5 concludes the paper and suggests directions for future works.

2. Preliminaries

Recently, with the advent of Linked Open Data such as DBpedia [4], Freebase [5], and Wikidata [6], KG gain significant momentum in developing accurate Knowledge Graphs-based Question Answering Systems (KGQAS) [7]. KGQAS leverage open KG to extract precise answers from user NL questions and offer flexibility, allowing schema evolution over time. KGQAS focus on understanding

6th International Hybrid Conference On Informatics And Applied Mathematics, December 6-7, 2023 Guelma, Algeria
*Corresponding author.
adjellal@escf-constantine.dz (A. Djellal);
mbenabdelhafid@escf-constantine.dz (M. S. Benabdelhafid)

user queries and extracting answers by matching and reasoning in KG. For instance, to answer the question "Who is Apple CEO? (see Figure 1), these systems tackle challenges like:

1. **Named Entity Recognition (NER)**, identifies fragments mentioning NE in text. In the above question the mention "Apple" is identified as a NE
2. **Named Entity Disambiguation (NED)**, seeks for each NE its corresponding meaning over a given KG, e.g. Wikidata. In our case, "Apple" can be linked to several Wikidata entries with different QID e.g. Q89 (apple, the fruit) or Q312 (Apple Inc., the company).
3. **Named Entity Linking (NEL)**, links each NE to its exact meaning over a KG, e.g., IRIs in Wikidata, based on the surrounding context. According to the question, "Apple" has to be disambiguated as "Apple Inc., the company" with the ID Q312. Therefore, NEL task has to link it to its IRI https://www.wikidata.org/wiki/Q312.

It is important to note that this paper aligns with the prevailing research trend, employing the term NEL to encompass both Disambiguation and Linking tasks, a convention adopted by several state-of-the-art approaches. Throughout the remainder of this paper, the primary focus revolves specifically around NEL task, rather than the complete QAS. For in-depth technical insights into NER task, interested readers are referred to the comprehensive surveys [8, 9]. The NEL process generally involves two steps:

- **Retrieving Candidates Entities:** The first step entails retrieving a set of candidate entities from the KG that the recognized NE may refer to. Various techniques are employed, including name dictionary-based methods [10], surface form expansion [11], and semantic relationships [12]. These methods rely primarily on string comparisons between the NE and the candidates, generating a set of potential entities. For example, "Apple", might be mapped to candidates like Q89 and Q312 in Wikidata (see Figure 1).
- **Selecting the Correct Candidate:** Given that a NE can often refer to a large number of candidate entities [13], the challenge lies in selecting the most relevant one. This step requires ranking the candidate entities based on the surrounding context and selecting the highly scored candidate that best fits the meaning of the given NE. For instance, if "Apple" refers to both the fruit and the company, according to the context, the correct candidate "Apple Co" needs to be selected.

3. Related Work

In recent years, the focus of NLP research has extended from the general language domain to the biomedical field, driven by Biomedical NLP (BioNLP) shared tasks and the increasing application of BioNLP tools in areas like clinical research and quality improvement [14, 15, 16, 17]. More particularly, Biomedical QA (BioQA) have been introduced for enabling innovative applications to effectively perceive, access, and understand complex biomedical knowledge [18]. On one hand, we can find for instance cTAKES [19], TaggerOne [20], and QuickUMLS [21] that are commonly used as rule-based knowledge-intensive concept normalization tools. These solutions use rules to generate lexical variants for each noun phrase and then perform dictionary queries for each variant. Although they provide robust performance, they implicitly assume the availability of concept aliases in the target language and focus on normalizing mentions and recognizing NE without effectively linking them [22].

Despite the developments, BioQA systems are still immature and rarely used in real-life settings. Current research often emphasizes morphological and string similarities of NE, neglecting their semantic similarities. NEL approaches are being introduced to maps various expressions, terms, or abbreviations to their corresponding common semantic representation or concept identifier in a given terminology or vocabulary. Biomedical language models are being explored to improve entity-linking strategies and to achieve automatic term mapping and some effective approaches to English corpora have been proposed. For instance, in [23], authors have proposed a collective inference approach, which leverages semantic information and structures in ontology to solve the NEL problem for biomedical literature. Also, in [24], scholars have proposed a graph-based linking approach which starts by constructing graphs for mentions, KG, and candidates and then exploits the information entropy and similarity algorithm to perform NEL. Like our approach, these contributions are dependent on the context and KG. In addition, scholars in [25] have proposed LATTE, a LATent Type Entity linking model, leveraging latent semantic information to improve entity linking, while authors in [26] have used semantic type information for improved entity disambiguation.

Different from the above works where no evaluation benchmark has been developed to evaluate how well language models represent biomedical concepts according to their corresponding context, authors in [27] propose a novel dataset, BioWiC, to evaluate the ability of language models to encode biomedical terms in context. Another research direction is to use for example BERT-based retrieve and re-rank models [28]. For instance, in [29], scholars have improved biomedical pretrained language models with knowledge.

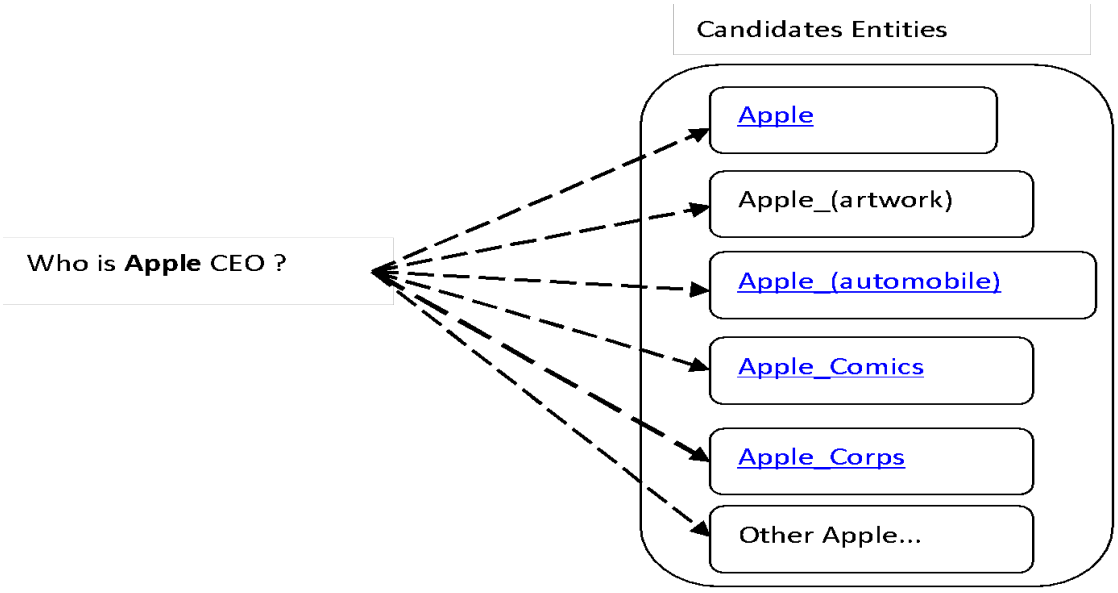

Figure 1: Illustrative example: A part of KG dedicated to the NE Apple

Let us notice that a particularly challenging is the task of NEL in short texts, such as questions, where limited contextual information hampers conventional linking methods. Addressing these challenges, this paper introduces a C-STSS approach, designed to enhance the performance of biomedical NEL systems dealing with short texts through contextual semantic similarities.

4. C-STSS Approach for Biomedical NEL

C-STSS approach involves four main sub-processes (see Figure 2). First, the Pre-process verifies and prepares the input question and recognizes the involved NE. Then, the Expansion generates the NE context by expanding the input question. Thereafter, Candidates Generation retrieves all NE candidates from DBpedia. Finally, the Ranking sub-process uses Semantic similarities to score candidates based on the generated context, and then links the NE to the highest scored candidate. This process frames NEL task as a ranking problem and will be detailed further in the following sections.

4.1. Pre-Process

The pre-processing step is vital as it significantly influences the outcome of the linking process, ensuring that the input question is refined and suitable for subsequent analysis. In the Pre-Process stage, the input question Q is subjected to critical transformations. After verifying the question's structure for grammatical or spelling errors, cleaning and normalization are performed to remove unnecessary or noisy words. This involves employing NL techniques such as tokenization [2] and stop-word removal [30], focusing on retaining only nouns, verbs, and adjectives. Finally, cTAKES [18], an open-source NLP tool, is utilized in order to recognize the involved NE.

It is noteworthy that due to the brevity of questions, words from the entity mention are included in the context window, especially if the entity consists of two or more words. For instance, in the case of NE "Malignant tumor" contextual words like "Malignant" and "tumor" are extracted as they contain meaningful common nouns.

In a biomedical scenario, a sample question Q could be: "How can Cancer be prevented and detected". Having this question as input, the pre-process generates as output the set of recognized NE and a set of words W.

Input: "How can Cancer be prevented and detected?"

Output:

- A set of words = Cancer, prevented, detected
- A set of = Cancer

4.2. Expansion

The Expansion module aims to enhance the contextual semantic similarity measurement particularly in short texts. In such case, traditional entity-entity relatedness

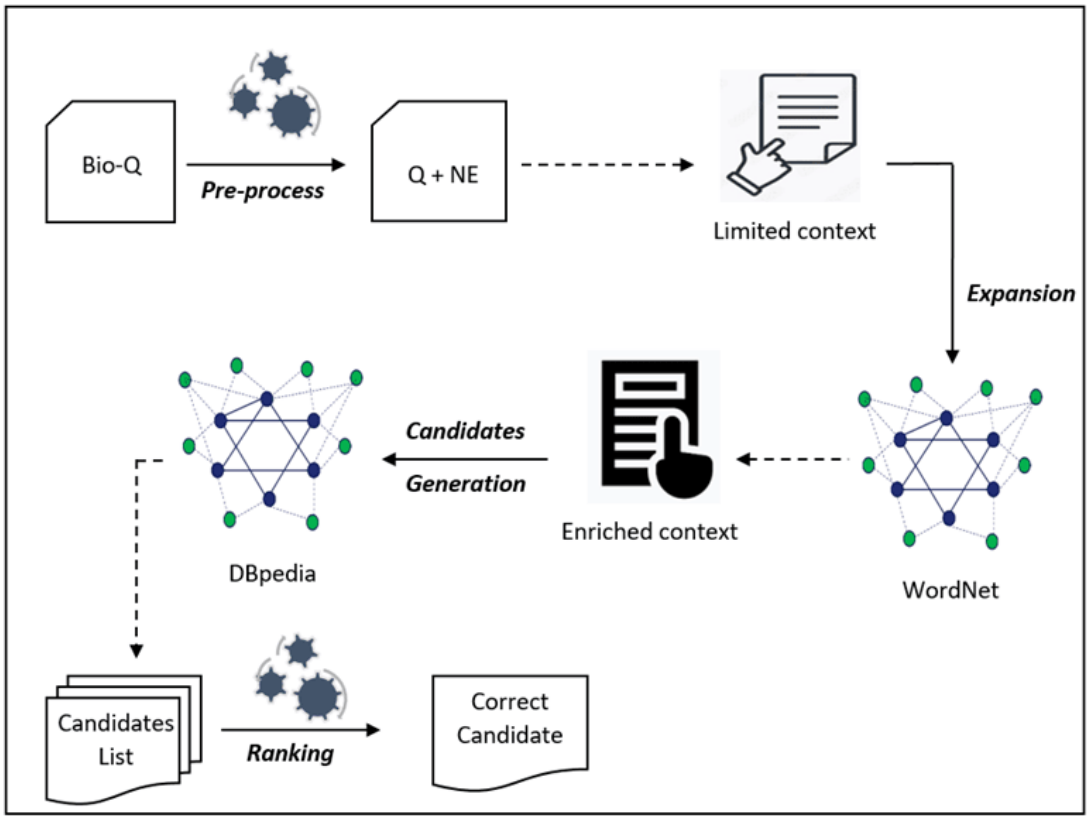

Figure 2: C-STSS system

approaches become ineffective due to the lack of context, and vocabulary mismatch further complicates the measurement of similarity between candidate descriptions and context. To overcome these challenges, the Expansion module enrich and expand the question with semantically related words. Initially, a stemming algorithm is used to reduce each word $w_i \in W$ to its root or stem in order to ensure a consistent comparison [31]. Then, it enriches the stemmed words by incorporating their synonyms using WordNet [32] as a background KG. Consequently, this module enables a more comprehensive analysis of the semantic similarities between the recognized NE and its candidates by allowing:

- **Lexical comparison:** Family words sharing the same stem, e.g., "prevention" and "prevented", could be compared. These words, although slightly varied, are semantically related.
- **Semantic comparison:** Words with different lexical forms but similar meanings, namely synonyms e.g., "prevented" and "avoided" bridge the vocabulary gap in short texts. Synonyms, despite their different surface form, are strongly semantically related.

At the end of the Expansion, the context window will be enriched with additional related words. Following our biomedical scenario, the set of words W is enriched resulting to the context C_W as represented in Table 1.

Input: A set of words = Cancer, prevented, detected

Output: A set of contextual words

4.3. Candidate Generation

The Candidate Generation module focuses on retrieving potential candidate entities to which the NE can refer to within DBpedia, a central KG comprising over 228 million entities from Wikipedia and Wikidata. The process begins by a simple string comparison to identify candidates whose names match the NE. However, dealing with name variations is a considerable challenge in

Table 1
Expansion of the question contextual window

Word W	Contextual words C_W	
	Stem	Synonyms
Cancer	Cancer	Malignant tumor, malignant neoplasm, metastatic tumor
Prevented	Prevent	Avoid, counter, forestall, foreclose, preclude, forbid...
Detected	Detect	Observe, find, discover, notice...

Table 2
Examples of biomedical NE and their candidates

Technique	NE	Candidate C_{NE}
Exact string match	Tumor	Tumor
Abbreviations / Acronyms	HIV	Human Immunodeficiency Virus
	DNC, D & C, or D and C	Dilation and curettage
Numbers	Vitamin B II	Vitamin B2
	three vessel disease	3 vessel disease
	triple vessel disease	3 vessel disease
Adjectives	Blood and urine tests	Blood tests, Urine tests
	Blood/urine tests	
	Blood or urine tests	
	have smooth, distinct borders	
Tokenization	High-dose vaccine	High dose vaccine

the biomedical field [33]. This variation is so extensive that a single entity can have multiple names, for instance, "decreases in hemoglobin" could refer to at least four different entities in MedDRA , which all look alike: "changes in hemoglobin", "increase in hematocrit", "hemoglobin decreased", and "decreases in platelets". Addressing the challenge of name variation, Candidate Generation employs several techniques:

- **Exact String Match:** Candidates sharing the exact string name with the NE are considered.
- **Abbreviations/Acronyms:** Biomedical dictionaries are utilized to handle abbreviations and acronyms common in the biomedical domain.
- **Numbers:** Variations in writing numbers (Arabic, Roman, or English spelled) are normalized for consistency.
- **Adjectives:** Multiple adjectives associated with a single noun employing composites like "and," "/", or "or" are separated and considered individually.
- **Tokenization:** Biomedical terms composed of multiple tokens connected by hyphens require dehyphenation for proper token sequence generation.

These techniques are elaborated in Table 2, providing an example for each case.

Let us notice that, exact string matches can be retrieved using DBpedia's disambiguation pages. If multiple DBpedia entries share the same name, a disambiguation page is created to differentiate them. For that, we generate a SPARQL query, specifying the NamedEntity (disambiguation) notion and the property wikiPageDisambiguates, to retrieve all links listed on this page and add them to the set of candidates.

From the previous Biomedical scenario, we retrieve the set of candidates: $C_{NE} = Dina_{cancer}, Chinese_{Astronomy}...$ having exact string match with $NE = Cancer$ by executing the SPARQL query presented in the following listing over DBpedia. The result is shown in Figure 3.

4.4. Ranking

Ranking module holds immense importance in the NEL process as it discerns the most suitable candidate for the NE based on the question context. When provided with a context C_W and a set of candidates C_{NE}, this module uses a ranking algorithm to compute for each candidate different contextual semantic similarities.

These contextual semantic similarities refer to the measurement of how closely the candidate aligns with the context. To this end, the algorithm computes some contextual semantic similarities according to the equation (1). The candidate with the highest score will be identified as the correct meaning of the NE. It is essential to highlight that the similarity between each candidate and the context is measured over its description in DBpedia.

$$score(C_i) = Sum(sim_{f_j}(C_i, Context)) \quad (1)$$

SPARQL query

```
Select?Candidates
Where {
<http://dbpedia.org/resource/Cancer_(disambiguation)>
<http://dbpedia.org/ontology/wikiPageDisambiguates>
 ?Candidates.
}
```

Results:


```
{ "head": { "link": [], "vars": ["Candidates"] },
  "results": { "distinct": false, "ordered": true, "bindings": [
    { "Candidates": { "type": "uri", "value": "http://dbpedia.org/resource/Dinah_Cancer" }},
    { "Candidates": { "type": "uri", "value": "http://dbpedia.org/resource/Cancer_(Chinese_astronomy)" }},
    { "Candidates": { "type": "uri", "value": "http://dbpedia.org/resource/Cancer_(Showbread_album)" }},
    { "Candidates": { "type": "uri", "value": "http://dbpedia.org/resource/Cancer_(band)" }},
    { "Candidates": { "type": "uri", "value": "http://dbpedia.org/resource/Cancer_(journal)" }},
    { "Candidates": { "type": "uri", "value": "http://dbpedia.org/resource/Cancer_(mythology)" }},
    { "Candidates": { "type": "uri", "value": "http://dbpedia.org/resource/Cancer" }},
    { "Candidates": { "type": "uri", "value": "http://dbpedia.org/resource/Cancer_(Confession_album)" }},
    { "Candidates": { "type": "uri", "value": "http://dbpedia.org/resource/Cancer_(My_Chemical_Romance_song)" }},
    { "Candidates": { "type": "uri", "value": "http://dbpedia.org/resource/Cancer_(My_Disco_album)" }},
    { "Candidates": { "type": "uri", "value": "http://dbpedia.org/resource/Cancer_(Transformers)" }},
    { "Candidates": { "type": "uri", "value": "http://dbpedia.org/resource/Cancer_(astrology)" }},
    { "Candidates": { "type": "uri", "value": "http://dbpedia.org/resource/Cancer_(constellation)" }},
    { "Candidates": { "type": "uri", "value": "http://dbpedia.org/resource/Cancer_(film)" }},
    { "Candidates": { "type": "uri", "value": "http://dbpedia.org/resource/Cancer_(genus)" }},
    { "Candidates": { "type": "uri", "value": "http://dbpedia.org/resource/Cancer_(comics)" }},
    { "Candidates": { "type": "uri", "value": "http://dbpedia.org/resource/Luis_Cancer" }},
    { "Candidates": { "type": "uri", "value": "http://dbpedia.org/resource/Tropic_of_Cancer_(disambiguation)" }} ] } }
```


Figure 3: Retrieving candidates having exact string match with the NE

Here, sim_{f_j} is a semantic similarity function. For each $C_i \in C_{NE}$, the following semantic similarities are computed:

Textual similarity: Given the NE context and a candidate $C_i \in C_{NE}$, we create two vectors representing their textual content: the candidate description vector, noted as v_{C_i} and the contextual words vector, noted as v_w. It should be noted that, lemmatization is applied on candidates descriptions for omitting stop words, very frequent and very rare words. We employ a standard Vector Space Model, with a $tf - idf$ weighting scheme for representing both vectors: $v_w = (d_1, d_2, ..., d_m)$, each dimension d_i of v_w corresponds to the word weight and is defined as:

$$d_i = tf(w_i, C_i) \times idf(w_i) \quad (2)$$

Where $tf(w_i, C_i)$, is the Term-Frequency function and denotes the frequency of the contextual word w_i in the candidate description C_i. It assesses the significance of the contextual word within the candidate's description. While $idf(w_i)$, stands for Inverse Document Frequency, signifies the number of candidates whose descriptions incorporate the contextual word w_i. In order to account for words that are overly frequent in candidate descriptions, $idf(w_i)$ is employed to assign lower weights to these less distinguishing words.

Hence, in order to compute textual similarity between the two vectors $v_w = (d_1, d_2, .., d_n)$ and $v_{C_i} = (d'_1, d'_2, .., d'_n)$, the cosine method is applied. This method calculates the cosine of the angle between these two vectors. It is defined as :

$$Sim_{text}(C_i, Context) = cos(v_{C_i}, v_w)$$
$$= \sum_i d_i d'_i / (\sqrt{\sum_i d_i^2} \sqrt{\sum_i {d_i}'^2}) \quad (3)$$

The primary challenge with using cosine similarity in advanced models lies in vocabulary mismatch. Cosine similarity essentially measures the correlation between the words of two textual vectors [2]. Consequently, this method fails to measure similarity when the vectors do not share identical words. Even if there are semantically related words, they are not taken into account. To face this drawback, we opt for knowledge-based methods to expand the input question with all words with semantic relevance when generating its context. This will successfully overcome issues such sparseness and vocabulary

mismatch while assessing textual similarity.

Candidate Popularity: Measuring the popularity of entities is a crucial factor in determining their relevance to a given NE. According to [13] a simple linking method based solely on candidate popularity can achieve 71% accuracy. It is essential to note that certain candidates are exceptionally rare compared to others. For instance, consider the $NE = Cancer$; while "Cancer (film)" might be a rare occurrence, "Cancer (astrology)" might be more common, with "Cancer (disease)" being the most popular entity. This observation can be formalized by analyzing candidates' incoming and outgoing links within DBpedia. The candidate popularity function, denoted as $R(C_i)$, is defined as follows:

$$Sim_{pop}(C_i) = R(C_i) = N_L(C_i) / \sum_{C_j \in C_{NE}} N_L(C_j) \quad (4)$$

Here, $N_L(C_i)$ represents the number of links pointing to the candidate C_i in DBpedia.

Word co-occurrence: In the state-of-the-art systems, the co-occurrence feature traditionally signifies the simultaneous appearance of a set of NE within the same text, allowing them to be collectively linked. Regrettably, this approach faces limitations when applied to short texts, where the presence of multiple NE is rare. Despite that, we adapted the co-occurrence concept to measure the contextual relevance between the NE and a given candidate. In our methodology, this feature is redefined as:

"The appearing of several contextual words within a given candidate description"

Obviously, the more different contextual words found within the candidate description, the closer it aligns with the NE context. To quantify this similarity, given the NE context and a candidate C_i, we examine two sets of words: the set of contextual words denoted as C_w and the set of candidate description words denoted as D_{C_i}. The word co-occurrence similarity function is defined as follows:

$$Sim_{Wco}(C_i, Context) = co{-}occurence(D_{C_i}, C_w) = \sum w_i^{D_{C_i}} / |C_w| \quad (5)$$

Here, $\sum w_i^{D_{C_i}}$ signifies the count of contextual words contained within D_{C_i}. This refined definition offers a nuanced understanding of word co-occurrence, enhancing the precision of context relevance measurements.

The details provided above are condensed into the subsequent algorithm, outlining our C-STSS approach. It encapsulates the intricacies of our C-STSS approach for biomedical NEL. Given an input question, C-STSS process employs the NER function to recognize the involved NE, generates the context using the Context function, and retrieves all potential candidates over DBpedia by employing Candidates function. These candidates are selected based on the five cases explained earlier. C-STSS algorithm incorporates furthermore functions in order to identify the more relevant candidate: Lemmatization is applied to omit stop words, very frequent and very rare words above context and candidate to enhance clarity. Words function retrieves feature words for context and candidate, shaping the subsequent analysis. Frequency function uses a $tf - idf$ weighting scheme for representing context and candidate vectors, ensuring a robust representation of the textual data.

Algorithm 1 C-STSS approach of biomedical NEL

Require: Question Q
Ensure: $C \in C_{NE}$ having the highest score

$Context \leftarrow context(Q)$
$NE \leftarrow NER(Q)$
$C_{NE} \leftarrow Candidates(NE)$
$SelectedC \leftarrow \emptyset$
$MaxScore \leftarrow 0$
$MinScore \leftarrow 0$
$context \leftarrow lemmatization(context)$
$v_w \leftarrow frequency(context)$
$C_w \leftarrow Words(context)$
$TotLinks \leftarrow 0$
for each $C \in C_{NE}$ **do**
    $TotLinks \leftarrow TotLinks + Links(C)$
end for
for each $C \in C_{NE}$ **do**
    $C \leftarrow lemmatization(C)$
    $v_C \leftarrow frequency(C)$
    $Sim_{txt} \leftarrow cos(v_w, v_C)$
    $L_C \leftarrow Links(C)$
    $Sim_{Pop} \leftarrow L_C / TotLinks$
    $D_w \leftarrow Words(C)$
    $N_w \leftarrow 0$
    for each $w \in C_w$ **do**
        if $w \in D_C$ **then**
            $N_w \leftarrow N_{w+1}$
        end if
    end for
    $Sim_{Wco} \leftarrow N_w / |C_w|$
    $ScoreS \leftarrow \sum(Sim_{txt}, Sim_{Pop}, Sim_{Wco})$
    $ScoreV \leftarrow min(1/n \sum_{i=1}^{3} (sim_{f_{j_i}} - \overline{sim_{f_j}})^2)$
    if $(ScoreS > MaxScore$ and $ScoreV < MinScore)$ **then**
        $MaxScore \leftarrow ScoreS$
        $MinScore \leftarrow ScoreV$
        $SelectedC \leftarrow C$
    end if
end for
Return $(SelectedC)$

To assess the similarity between words in the context and those in the candidate descriptions, three distinct semantic similarity metrics are calculated and combined to score each candidate:

- Sim_{txt} Represents the cosine similarity between the context and the candidate description,
- Sim_{Pop} Represents the candidate popularity, providing valuable insight into its relevance.
- $Sim_{W_{co}}$ Evaluates the extent to which the candidate's description covers the context, offering a holistic perspective.

This similarity computation is iteratively applied to all candidate entities in order to scoring them. The candidate with the highest score and the lowest standard deviation is returned as the correct one.

4.5. Discussion

While various scholars focus on addressing the name variation problem in BioQA by considering morphological forms of biomedical NE, few incorporate semantic similarities. C-STSS approach combines NE morphological forms and contextual semantic similarities. To further enhance its efficacy, our approach integrates knowledge-based methods with corpus-based ones, alleviating issues related to sparseness and vocabulary mismatch. This fusion of techniques forms the core innovation of this research.

To conclude, it is now well established that biomedical text requires methods targeted for the domain. Developments in Deep Learning and a series of successful shared challenges have contributed to a steady progress in techniques for Bio-NLP text. Contributing to this ongoing progress and particularly focusing on computational methods, our future issue will aim to create and encourage research in novel approaches for analyzing biomedical text. More particularly, on transformer-based models that seem to be the future of NLP as explained in recent surveys [34, 35, 36, 37, 38].

5. Conclusion

In recent years, KG have undergone substantial growth in both theoretical frameworks and practical applications. Despite these advancements, KGQAS encounter persistent challenges. They face limitations due to historical precedents and excessive human intervention, necessitating innovative solutions.

Within the intricate domain of biomedicine, additional complexities emerge. Indeed, NEL in the medical domain is a newer problem. This paper presents a Context-based Short Text Semantic Similarity approach, designed to enhance biomedical NEL systems by exploiting contextual semantic similarities in order to face short texts limited context.

C-STSS approach not only consider morphological forms of biomedical NE but also delves into contextual semantic similarities. In addition, it frames the NEL task as a ranking problem, employing multiple semantic measures to score each candidate based on the context derived from expanding the input question.

Currently, we further probe our proposed algorithm focuses on biomedical NEL for short texts, characterized by their succinct and constrained contexts. We are actively refining our algorithm through rigorous testing and optimization exploiting DBpedia as a background KG. The initial implementation has been executed and is awaiting thorough evaluation. Looking ahead, our future endeavors involve the exploration of Deep Learning techniques [30] to further enhance the proposed algorithm. Additionally, we plan to delve into the exploration of diverse KG, broadening the scope of our research.

References

[1] E. Dimitrakis, K. Sgontzos, Y. Tzitzikas, A survey on question answering systems over linked data and documents, Journal of intelligent information systems 55 (2020) 233–259.

[2] G. Zhu, C. A. Iglesias, Exploiting semantic similarity for named entity disambiguation in knowledge graphs, Expert Systems with Applications 101 (2018) 8–24.

[3] R. Navigli, Word sense disambiguation: A survey, ACM computing surveys (CSUR) 41 (2009) 1–69.

[4] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, in: international semantic web conference, Springer, 2007, pp. 722–735.

[5] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 2008, pp. 1247–1250.

[6] D. Vrandečić, Wikidata: A new platform for collaborative data collection, in: Proceedings of the 21st international conference on world wide web, 2012, pp. 1063–1064.

[7] M. R. A. H. Rony, D. Chaudhuri, R. Usbeck, J. Lehmann, Tree-kgqa: an unsupervised approach for question answering over knowledge graphs, IEEE Access 10 (2022) 50467–50478.

[8] T. Al-Moslmi, M. G. Ocaña, A. L. Opdahl, C. Veres, Named entity extraction for knowledge graphs: A literature overview, IEEE Access 8 (2020) 32862–32881.

[9] V. Yadav, S. Bethard, A survey on recent advances in named entity recognition from deep learning models, arXiv preprint arXiv:1910.11470 (2019).

[10] Z. Yang, H. Lin, Y. Li, Exploiting the performance of dictionary-based bio-entity name recognition in biomedical literature, Computational biology and chemistry 32 (2008) 287–291.

[11] A. Reshamwala, D. Mishra, P. Pawar, Review on natural language processing, IRACST Engineering Science and Technology: An International Journal (ESTIJ) 3 (2013) 113–116.

[12] R. Meymandpour, J. G. Davis, A semantic similarity measure for linked data: An information content-based approach, Knowledge-Based Systems 109 (2016) 276–293.

[13] W. Shen, J. Wang, J. Han, Entity linking with a knowledge base: Issues, techniques, and solutions, IEEE Transactions on Knowledge and Data Engineering 27 (2014) 443–460.

[14] G. Frisoni, G. Moro, A. Carbonaro, A survey on event extraction for natural language understanding: Riding the biomedical literature wave, IEEE Access 9 (2021) 160721–160757.

[15] T. A. Koleck, C. Dreisbach, P. E. Bourne, S. Bakken, Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review, Journal of the American Medical Informatics Association 26 (2019) 364–379.

[16] I. J. B. Young, S. Luz, N. Lone, A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis, International journal of medical informatics 132 (2019) 103971.

[17] E. French, B. T. McInnes, An overview of biomedical entity linking throughout the years, Journal of biomedical informatics 137 (2023) 104252.

[18] Q. Jin, Z. Yuan, G. Xiong, Q. Yu, H. Ying, C. Tan, M. Chen, S. Huang, X. Liu, S. Yu, Biomedical question answering: a survey of approaches and challenges, ACM Computing Surveys (CSUR) 55 (2022) 1–36.

[19] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, C. G. Chute, Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications, Journal of the American Medical Informatics Association 17 (2010) 507–513.

[20] R. Leaman, Z. Lu, Taggerone: joint named entity recognition and normalization with semi-markov models, Bioinformatics 32 (2016) 2839–2846.

[21] L. Soldaini, N. Goharian, Quickumls: a fast, unsupervised approach for medical concept extraction, in: MedIR workshop, sigir, 2016, pp. 1–4.

[22] R. Bhowmik, K. Stratos, G. de Melo, Fast and effective biomedical entity linking using a dual encoder, arXiv preprint arXiv:2103.05028 (2021).

[23] J. G. Zheng, D. Howsmon, B. Zhang, J. Hahn, D. McGuinness, J. Hendler, H. Ji, Entity linking for biomedical literature, BMC medical informatics and decision making 15 (2015) 1–9.

[24] H. Wang, J. G. Zheng, X. Ma, P. Fox, H. Ji, Language and domain independent entity linking with quantified collective validation, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 695–704.

[25] M. Zhu, B. Celikkaya, P. Bhatia, C. K. Reddy, Latte: Latent type modeling for biomedical entity linking, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 9757–9764.

[26] S. Vashishth, D. Newman-Griffis, R. Joshi, R. Dutt, C. P. Rosé, Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets, Journal of biomedical informatics 121 (2021) 103880.

[27] H. Rouhizadeh, I. Nikishina, A. Yazdani, A. Bornet, B. Zhang, J. Ehrsam, C. Gaudet-Blavignac, N. Naderi, D. Teodoro, Biowic: An evaluation benchmark for biomedical concept representation, bioRxiv (2023) 2023–11.

[28] Y. He, Z. Zhu, Y. Zhang, Q. Chen, J. Caverlee, Infusing disease knowledge into bert for health question answering, medical inference and disease name recognition, arXiv preprint arXiv:2010.03746 (2020).

[29] Z. Yuan, Y. Liu, C. Tan, S. Huang, F. Huang, Improving biomedical pretrained language models with knowledge, arXiv preprint arXiv:2104.10344 (2021).

[30] Z. Xu, X. Luo, S. Zhang, X. Wei, L. Mei, C. Hu, Mining temporal explicit and implicit semantic relations between entities using web search engines, Future Generation Computer Systems 37 (2014) 468–477.

[31] C. Ramasubramanian, R. Ramya, Effective preprocessing activities in text mining using improved porter's stemming algorithm, International Journal of Advanced Research in Computer and Communication Engineering 2 (2013) 4536–4538.

[32] C. Fellbaum, WordNet: An electronic lexical database, MIT press, 1998.

[33] L. Chen, G. Varoquaux, F. M. Suchanek, A lightweight neural model for biomedical entity linking, in: Proceedings of the AAAI conference on artificial intelligence, volume 35, 2021, pp. 12657–12665.

[34] L. Cai, J. Li, H. Lv, W. Liu, H. Niu, Z. Wang, Incorporating domain knowledge for biomedical text analysis into deep learning: A survey, Journal of Biomedical Informatics (2023) 104418.

[35] K. S. Kalyan, A. Rajasekharan, S. Sangeetha, Ammu: a survey of transformer-based biomedical pretrained language models, Journal of biomedical

informatics 126 (2022) 103982.
[36] S. Islam, H. Elmekki, A. Elsebai, J. Bentahar, N. Drawel, G. Rjoub, W. Pedrycz, A comprehensive survey on applications of transformers for deep learning tasks, Expert Systems with Applications (2023) 122666.
[37] K. Hall, V. Chang, C. Jayne, A review on natural language processing models for covid-19 research, Healthcare Analytics (2022) 100078.
[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).