International Journal of Techno-informatics Engineering
Vol: 02; Issue: 01; pg: 001-007

Development in Named Entity Extraction and Its Significance for Improving Knowledge Graph Building

Chennakesh S[1], Dr. Kamlesh Kumar Pandey[2]
[1] Research Scholar, Capital University, Jhumri Telaiya, Jharkhand, India
[2] Professor, Capital University, Jhumri Telaiya, Jharkhand, India

ABSTRACT
Knowledge Graphs (KGs) have emerged as a powerful tool for representing semantic relationships between data entities, widely used in various applications such as search engines, information retrieval, and natural language processing. This study provides an extensive literature review of Named Entity Extraction (NEE), focusing on recent advances in Named Entity Recognition (NER), Named Entity Disambiguation (NED), and Named Entity Linking (NEL). The paper highlights how these processes facilitate the transformation of unstructured natural language data into structured knowledge suitable for KGs. Additionally, we explore the evolution of approaches from rule-based systems to machine learning and neural network-based methods, emphasizing their impact on accuracy and efficiency. The findings suggest that while NER, NED, and NEL are critical for semantic lifting, challenges remain, particularly in handling ambiguous entities and integrating off-the-shelf NER tools with domain-specific knowledge graphs.
Keywords: *Named Entity Extraction; Named Entity Recognition; Knowledge Graphs; Named Entity Disambiguation; Natural Language Processing (NLP)*

INTRODUCTION

Knowledge Graphs (KG) [1] were introduced for broader use by Google in 2012, primarily to enhance the interlinking of data to improve the accuracy and efficiency of search queries [2]. In the context of KGs, nodes represent concrete objects, abstract concepts, information resources, or data about these elements, while the edges define the semantic relationships between them [3]. KGs have emerged as a widely adopted framework for representing complex information in a structured, computer-processable manner. Their development is deeply rooted in the vision of Tim Berners-Lee's semantic web—an augmentation of the original web of human-readable documents into a web of machine-processable data. This semantic web, made possible by technologies like RDF, RDFS, and OWL, offers a foundation for more intelligent, interconnected data systems. However, in practice, KGs tend to be less rigid than the ontology-driven framework envisioned for the semantic web [4].

Simultaneously, a vast portion of information available on the internet is expressed in natural language (NL), which remains challenging for computers to process directly. This disparity creates a pressing need for Natural Language Processing (NLP) techniques and other information extraction tools. One of the core challenges in transforming unstructured natural language text into structured data is identifying the entities embedded within the text. These entities can range from specific individuals, organizations, or locations to abstract concepts like events, theories, or processes. By encoding entities as nodes and relationships as edges, Knowledge Graphs offer a natural and intuitive way to represent NL text in a format that computers can process efficiently.

This paper reviews recent advances in a crucial area of research: Named Entity Extraction (NEE), which is central to the task of transforming NL texts into structured knowledge representations. NEE encompasses three primary sub-tasks: Named Entity Recognition (NER), which focuses on identifying mentions of named entities in text; Named Entity Disambiguation (NED), which resolves the ambiguity of these entities by identifying the correct reference; and Named Entity Linking (NEL), which associates recognized entities with specific objects in a knowledge base [5].

Over the past decade, there has been significant progress in each of these subfields. NER systems have evolved from rule-based approaches to sophisticated deep learning models, while NED and NEL have similarly benefited from advancements

Email ✉: chennakesh@aol.com

in machine learning and attention-based mechanisms. However, challenges persist in creating systems that can seamlessly integrate NEE with Knowledge Graphs, especially in domain-specific contexts where data sparsity or ambiguity is an issue.

Named Entity Extraction (NEE), known as Named Entity Recognition (NER), Disambiguation (NED), and Linking (NEL) has significantly advanced over the years, but still encounters several difficulties. Recent methods, especially deep learning based models provide enhanced accuracy and efficiency. But, there is still a long way to manage domain specific knowledge graphs especially data sparse or ambiguous domain. The effort for elaborate configuration often makes off-the-shelf NER tools to be integrated with the customized knowledge graphs quite less scalable in real-time large-scale system. Finally, there is also room for further research in developing models which are adaptive and not language or domain specific and can generalize well across languages/environments. The gap indicates that future research should instead work on automatically reifying any sufficiently powerful, universal model so that it scales across different domains without much manual effort.

The aim of this study is to provide a comprehensive review of the recent developments in Named Entity Extraction, focusing on how these advancements have contributed to improving Knowledge Graph construction. By reviewing the evolution of NER, NED, and NEL techniques, this paper aims to identify current challenges and opportunities for future research, especially in leveraging these methods to bridge the gap between unstructured natural language text and structured data representations in Knowledge Graphs.

NATURAL-LANGUAGE PROCESSING (NLP)

Natural-language processing (NLP) attempts to enable computers to process human language in meaningful ways [6]. For example, Chowdhary, and Chowdhary [6] describes NLP as ‘‘an area of research and application that explores how computers can be used to understand and manipulate natural-language text or speech to do useful things’’. NLP is commonly used to derive semantics from text or speech and to encode it in a structured format that is suitable for semantic search and other types of computer processing. Many well-established techniques and tools are already available for this purpose, such as GATE and other NLP pipelines; IBM Watson1 and other NL analysis and lifting services; NLTK, Stanford CoreNLP, DBpedia Spotlight, and other NL programming APIs; OpenIE2, MinIE, and other information extraction tools [7].

KNOWLEDGE GRAPHS

A knowledge graph (KG) represents semantic data as triples (i.e., as ordered sets of terms) composed as (s,p,o): subjects, predicatep, objecto, that can be either IRIs (Internationalized Resource Identifier) i∈I, blank nodes b∈B or literals l∈L, so that s∈I∪B,p∈I and o∈I∪B∪L [8].

The IRIs used as subjects, predicates, and objects can be taken from well-defined vocabularies or ontologies in the Linked Open Data (LOD) cloud, that attempt to define their meaning as precisely as possible. The literal values used as objects can be represented using well-defined data types, such as the ones defined by XML Schema Definition (XSD). Through the use of terms with well-defined meanings for types, instances, and relations, the knowledge graph aims to describe the semantics of real-world entities and their relations precisely and to link the descriptions to further information in semantic LOD repositories [9].

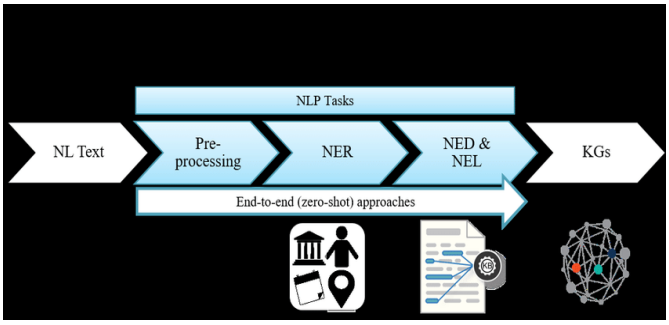


Figure 1. Natural-language processing (NLP) tasks (Bizer, et al 2023) [9]

REPRESENTING NAMED ENTITIES IN KGS

A named entity is an individual such as a person, organization, location, or event. A mention is a piece of text that refers to an entity. As already mentioned, extracting named entities from texts and representing them as nodes in KGs involves three main tasks: Named Entity Recognition (NER) attempts to find every segment in a text that mentions a named entity. Named Entity Disambiguation (NED) attempts to determine which named entity a mention refers to; for example, the mention ‘‘Trump’’ can refer to either a person, a corporation or a building. Named Entity Linking (NEL) attempts to provide a standard IRI for each disambiguated entity; for example, Trumpthe-president can be linked to the IRI that represents him in1.

NED and NEL are closely entwined, because an ambiguous entity must be disambiguated before it can be linked and because an IRI is a good way to represent the result of disambiguation. Therefore, we will often discuss NED and NEL together. Figure 1.2 shows the resulting sequence of tasks from NL to KGs. The figure also shows the more recent end-to-end (also called zero-shot) approaches, which lump together all three tasks, typically using deep neural networks. These approaches usually rely on standard pre-processing steps, which we will review first, before we go on to the other tasks [10].

PRE-PROCESSING

The most frequently used pre-processing techniques in NL lifting are tokenization and part-of-speech (POS) tagging [11]. Other common techniques include: stop-word removal, normalization, sentence splitting, lemmatization, chunking and dependency parsing, and structural parsing. Although pre-processing is used by default, some studies do not describe in detail how they are performed. A possible reason is that most studies, especially on NER, use standard datasets that have already been pre-processed. For example, CoNLL2003, the most popular dataset for NER, has already applied tokenization, POS tagging, and a chunking to the raw data.

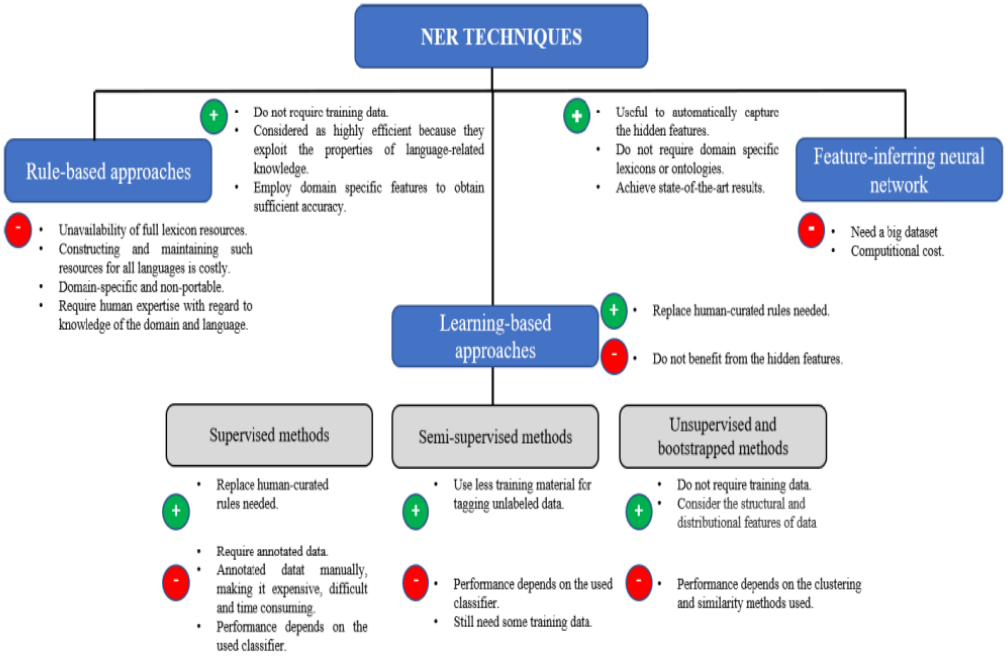


Figure 2. Named-entity recognition (NER) approaches Shelar et al (2020) [10].

The appropriate choice of pre-processing technique depends on the lifting technique to be used. For example, removing stop words might be good for a Bag-of-Words (BOW) based approach or for a model that does not consider word order, but deep-learning approaches might leverage stop words to disambiguate entities that have different meanings. Recent contributions indicate that robust NED and NEL systems require accurate tuning of several prior steps, especially tokenization and semantic similarity. Recently, deep neural networks, in particular the end-to-end approaches, have reduced

[1] Wikidata: https://www.wikidata.org/entity/Q22686

the need for pre-processing steps. Using deep neural networks for pre-processing tasks such as tokenization has also produced promising results [12].

Data quality plays a key role in selecting the most suitable pre-processing technique too. For example, most gold-standard datasets do not require the same pre-processing as raw-web or real-time streaming data, from which cleaning and normalization are needed to remove unnecessary or noisy terms (like emojis, currency symbols, hashtags, and so forth).

NAMED ENTITY RECOGNITION

NER was first introduced in . According to Yadav, and Bethard (2019) [13], the purpose of NER is to identify named entities contained in a text like persons, locations, organizations, time, clinical procedures, money, biological proteins, etc

NER is a rapidly evolving research field. Most of the proposed approaches have been domain-specific, limiting themselves to for example news, reviews, etc. We divide them into three main categories following [14]

The first, and earliest, category is the rule- or knowledge-based approaches. Most studies in this category are based on hand-crafted rules. The advantage of such methods is that they do not require annotated training data since they rely on lexical resources. Another advantage is that the precision of handcrafted methods can become high because of the lexicons and domain-specific knowledge. The disadvantage is that this also makes them domain dependent that lexicon resources may be unavailable, and that constructing and maintaining such resources for many languages is costly [14].

The second category of NER systems is the learning based approaches [15]. These models are used to replace the human-curated rules needed by the first category. The methods in this category can be divided into three types: supervised, semi-supervised, and unsupervised. In supervised and semi-supervised methods, a machine-learning model is trained on input examples together with their targeted outputs. Support Vector Machines (SVM), Hidden Markov Models (HMM), Conditional Random Fields (CRF), and decision trees are common in this category [13]. NER accuracy is sometimes limited by the used classifier. For example, when HMM and SVM are employed, the dependencies among words are not considered. The unsupervised and bootstrapped methods are more automated, although they need a minimal training dataset (seeds). Although these methods do not require as much effort as the first category, seed data is still needed for training. Moreover, they do not benefit from the feature inference of the third category below [15].

The last, and most recent, category is the feature inferring neural-network approaches [16]. They rely on machine learning like the previous category, but they differ from the rule- and learning-based approaches by automatically inferring features through deep learning. Recent research reports that they thus outperform earlier methods. Unlike the above-mentioned approaches, they do not require seeds, ontologies, or domain specific lexicons and are thus more domain-independent. They also benefit from the precision of their inferred features. On the other hand, big datasets are needed to build robust models

Although numerous NER studies have been reported recently, few have been used for semantic lifting or received attention in KG research. There may be at least two reasons. The first reason is that NER is an initial step for many other tasks — such as sentiment analysis, concept extraction, event identification and so forth — that have received more attention than semantic lifting. The second reason is that semantic lifting research has focused on the KG-construction side, leaving the NER task to off-the-shelf APIs and tools. However, using standard systems suffers from configuration restrictions and makes the combination of or switching between different NER solutions more difficult [17].

NAMED ENTITY DISAMBIGUATION

Based on how they rank candidate entities, NED approaches can be classified into three categories:

Traditional NED approaches

Traditional NED studies typically use hand-designed features to calculate the similarity between the mention and its candidate entities [18]. These studies can be further subdivided into independent (or individual) and collective approaches.

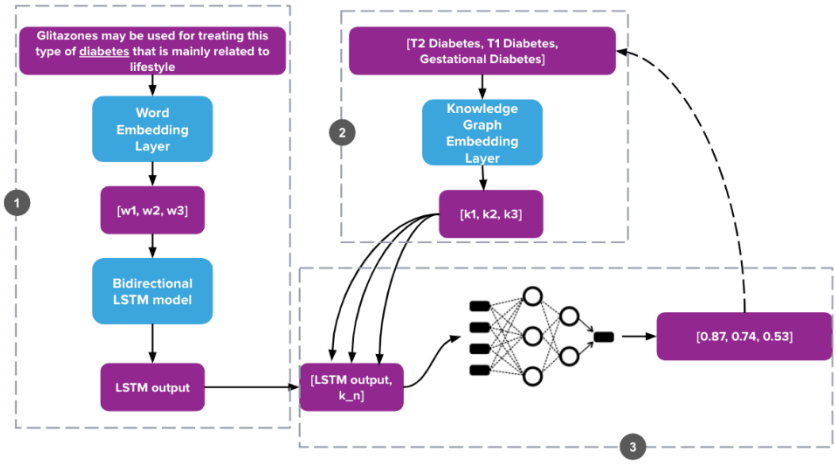
Independent approaches use semantic similarity techniques to rank candidate entities solely according to their lexical similarity and/or empirical co-occurrence with mentions. In these methods, each mention is independently disambiguated, and they consider entity disambiguation as a ranking problem that picks the entity with the highest confidence. The confidence value is scored by combining hand-designed features extracted from the mention's context (surrounding words) with textual descriptions of the candidate entity, for example using text about it from Wikipedia. Different similarity measures have been proposed, such as BOW-based cosine similarity. Because hand-designed features tend to contain little textual information, the accuracy of such approaches decreases in complex cases [19]. Moreover, these methods fail to catch interactions between mentions in the same document.

Collective approaches also rely on semantic similarity, but they take into account that what is mentioned in the same (part of a) text tends to be about the same topic and that co-occurring entities should therefore often be semantically related. Most commonly used collective approaches thus establish the associations between candidate entities and build the mention-entity pairs using a probabilistic graph approach. AGDISTIS, Babelify and TagME use graph connectedness to exploit semantic relations between disambiguation candidates described in a knowledge base (KB). In fact, feature representations of entities and mentions are the key factors for most of these NED approaches. Most of them rely on BOW, which, in general, has some shortcomings, such as ignoring word meanings and expensive computation [19].. The methods thus do not exploit latent features in mention contexts and candidate descriptions. A well-matched entity group can be identified using either random walks, Pagerank, or dense sub-graph computations. Although the collective approaches perform more robustly than the independent ones, computation costs grow rapidly as the numbers of mentions and the lengths of documents increase. Phan et al (2019) [20], proposes a simplified collective pair-linking approach, which resolves the candidate entities pair-by-pair in order to decrease computational cost and complexity. In hand-crafted feature-based approaches, it is difficult for NED systems to fully leverage the inherent semantics of mention contexts and entity descriptions, which is pivotal for NED accuracy. This is due to the limitations of hand-designed features, which capture lexical information based on the surface form of the text.

NEURAL-NETWORK (NN) APPROACHES

In recent years, neural-network (NN) based methods have become more common and achieved competitive results. One family of approaches map words into continuous vector spaces using word2vec or similar models that comprise more semantic information than traditional BOW representations. Another family uses deep NNs to learn the latent semantic features automatically. NED accuracy is thereby enhanced along with the model's generalization ability.

The majority of the earlier NN-based approaches grant all the words in the mention context equal importance, which is adequate for many practical cases [21]. More recently, attention mechanisms have been introduced to assign graded importance. However as mentioned in Nie et al (2019) [21], most of these methods only apply attention to mention contexts and omit the entity side. Also, they only apply attention to a single aspect of knowledge, which may not be sufficient in complex circumstances, for example with high noise contexts or less popular entities.



sFigure need explanation (NED task) (Nie, et al 2018) [21]

DISCUSSION

The review of NEE approaches demonstrates significant progress, particularly with the advent of neural network-based methods. Rule-based systems, while precise in certain domains, rely heavily on domain-specific lexicons and handcrafted rules, limiting their applicability to new or unfamiliar contexts. Machine learning, and more recently, deep learning models, have provided alternatives that can automatically infer features and adapt to various domains. These advances have reduced the dependency on manual rule creation, enabling wider applicability across different fields. However, challenges remain, especially in Named Entity Disambiguation (NED), where computational costs become significant in collective approaches. Although these methods enhance robustness by considering entity interactions within documents, they are resource-intensive and may not scale well in real-time applications.

Attention mechanisms have been incorporated into NED to improve accuracy by assigning importance to relevant contextual elements. However, optimization challenges still persist, especially in handling noisy contexts, low-resource languages, or less common entities. Recent trends toward end-to-end models, which combine NER, NED, and NEL in a single framework, offer promise by streamlining the process and reducing the need for extensive pre-processing. These models, often based on deep neural networks, present opportunities for more efficient semantic lifting. Nonetheless, domain-specific adaptation and the integration of these models into real-world systems remain ongoing areas of research. The future of NEE will likely focus on creating adaptable models that can seamlessly integrate with Knowledge Graphs across various domains and contexts.

CONCLUSION

Named Entity Extraction (NEE) has evolved significantly from early rule-based systems to sophisticated deep learning techniques, leading to enhanced accuracy, efficiency, and adaptability in constructing Knowledge Graphs. While current systems show remarkable improvements in generalizability and precision, there remain unresolved challenges. The high computational cost associated with collective disambiguation approaches and the complexity of handling ambiguous or noisy data present hurdles to large-scale deployment, particularly in real-time applications. Moreover, domain-specific variations and low-resource settings demand the development of more flexible models capable of dynamically adjusting to different contexts without significant manual intervention.

The future scope of NEE research involves optimizing the balance between computational efficiency and accuracy. Efforts should focus on creating lightweight models that can operate in real-time, making them suitable for applications requiring fast processing, such as search engines, conversational agents, and recommendation systems. Additionally, integrating more sophisticated attention mechanisms and multi-modal learning approaches could further enhance the handling of complex or ambiguous entities. There is also potential for expanding the applicability of NEE to underrepresented languages and domains, which would increase the versatility of Knowledge Graph systems. In parallel, future work should explore methods for improving semantic lifting techniques, ensuring that the extracted data is not only precise but also rich in contextual meaning. Finally, there is scope for developing more comprehensive, domain-agnostic NEE systems that seamlessly combine entity recognition, disambiguation, and linking in a unified, scalable framework, catering to diverse data sources and applications.

REFERENCES

1. Yan J, Wang C, Cheng W, Gao M, Zhou A. A retrospective of knowledge graphs. Frontiers of Computer Science. 2018 Feb; 12:55-74.
2. Kellou-Menouer K, Kardoulakis N, Troullinou G, Kedad Z, Plexousakis D, Kondylakis H. A survey on semantic schema discovery. The VLDB Journal. 2022 Jul; 31(4):675-710.
3. Zaman G, Mahdin H, Hussain K, Abawajy J, Mostafa SA. An ontological framework for information extraction from diverse scientific sources. IEEE access. 2021 Mar 2; 9:42111-24.
4. Hofer M, Obraczka D, Saeedi A, Köpcke H, Rahm E. Construction of Knowledge Graphs: Current State and Challenges. Information. 2024 Aug 22; 15(8):509.
5. Mithun AM, Bakar ZA. Empowering information retrieval in semantic web. International Journal of Computer Network and Information Security. 2020;12(2):41-8.
6. Chowdhary K, Chowdhary KR. Natural language processing. Fundamentals of artificial intelligence. 2020:603-49.
7. Dessì D, Osborne F, Recupero DR, Buscaldi D, Motta E. Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain. Future Generation Computer Systems. 2021 Mar 1; 116:253-64.
8. Färber M, Bartscherer F, Menne C, Rettinger A. Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. Semantic Web. 2018 Jan 1;9(1):77-129.
9. Bizer C, Heath T, Berners-Lee T. Linked data-the story so far. InLinking the World's Information: Essays on Tim Berners-Lee's Invention of the World Wide Web 2023 Jul 7 (pp. 115-143).
10. Shelar H, Kaur G, Heda N, Agrawal P. Named entity recognition approaches and their comparison for custom ner model. Science & Technology Libraries. 2020 Jul 2;39(3):324-37.
11. Fossati M, Dorigatti E, Giuliano C. N-ary relation extraction for simultaneous T-Box and A-Box knowledge base augmentation. Semantic Web. 2018 Jan 1;9(4):413-39.
12. Boroş T, Dumitrescu ŞD, Burtica R. NLP-Cube: End-to-end raw text processing with neural networks. InProceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies 2018 Oct (pp. 171-179).

13. Yadav V, Bethard S. A survey on recent advances in named entity recognition from deep learning models. arXiv preprint arXiv:1910.11470. 2019 Oct 25.
14. Rustad S, Shidik GF, Noersasongko E. Systematic Literature Review on Named Entity Recognition: Approach, Method, and Application. Statistics, Optimization & Information Computing. 2024 Feb 28;12(4):907-42.
15. Li J, Sun A, Han J, Li C. A survey on deep learning for named entity recognition. IEEE transactions on knowledge and data engineering. 2020 Mar 17;34(1):50-70.
16. Baevski A, Edunov S, Liu Y, Zettlemoyer L, Auli M. Cloze-driven pretraining of self-attention networks. arXiv preprint arXiv:1903.07785. 2019 Mar 19.
17. Plu J, Rizzo G, Troncy R. Enhancing entity linking by combining NER models. InSemantic Web Challenges: Third SemWebEval Challenge at ESWC 2016, Heraklion, Crete, Greece, May 29-June 2, 2016, Revised Selected Papers 3 2016 (pp. 17-32). Springer International Publishing.
18. Wang C, Sun X, Yu H, Zhang W. Entity disambiguation leveraging multi-perspective attention. IEEE Access. 2019 Aug 7; 7:113963-74.
19. Conover M, Hayes M, Blackburn S, Skomoroch P, Shah S. Pangloss: Fast entity linking in noisy text environments. InProceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2018 Jul 19 (pp. 168-176).
20. Phan MC, Sun A, Tay Y, Han J, Li C. Pair-linking for collective entity disambiguation: Two could be better than all. IEEE Transactions on Knowledge and Data Engineering. 2018 Jul 19;31(7):1383-96.
21. Nie F, Cao Y, Wang J, Lin CY, Pan R. Mention and entity description co-attention for entity disambiguation. InProceedings of the AAAI Conference on Artificial Intelligence 2018 Apr 26 (Vol. 32, No. 1).