Expert Systems With Applications 238 (2024) 121759

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

SRSCL: A strong-relatedness-sequence-based fine-grained collective entity linking method for heterogeneous information networks

Lizheng Zu[4], Lin Lin[*,2], Jie Liu[*,1], Song Fu[3], Changsheng Tong[5], Hao Guo[6]

School of Mechatronics Engineering, Harbin Institute of Technology, Harbin, China

ARTICLE INFO

Keywords:
Collective entity linking
Heterogeneous information network
Overall relatedness
Strong relatedness sequence
Knowledge representation

ABSTRACT

The development of efficient methods for mining information from heterogeneous information networks (HINs) has become essential for improving the accuracy of collective entity linking in the absence of third-party knowledge bases. Currently, there remain three major challenges in the latest research: (1) The objective function of collective entity linking does not fully embrace the concept of “collective linking”. (2) The objective function employs the mean value rather than the maximum value of the entity relatedness as a link parameter, while discounting the importance of the strong logical associations between the text and language for meaning recognition. (3) The objective function utilizes only one type of 2-hop path to contribute to entity relatedness, thereby disregarding other types of 2-hop path that exist in actual HINs. To address the aforementioned issues, this paper proposes a strong-relatedness-sequence-based fine-grained collective entity linking method (SRSCL). The SRSCL is capable of capturing the contextual information of the entity in the HINs, thereby providing improved accuracy in entity linking. Specifically, SRSCL constructs a knowledge representation learning model and proposes an overall semantic similarity model for entity mentions and candidate entities to solve the objective function and thereby reflect the idea of “collective linking”. Additionally, a strong-relatedness-sequence-based overall relatedness measurement model is proposed for candidate entities to emphasize the strong logical associations between them. Furthermore, SRSCL defines three types of 2-hop path and evaluates the importance of each path to accurately measure the relatedness of entities. Finally, the experimental results demonstrate that the proposed SRSCL is more effective in capturing the overall relatedness of entities than the latest model. Particularly, when the number of entity mentions contained in one sliding window is greater than 6, the proposed SRSCL improves the precision, recall and F1 score by more than 10% compared with the latest model.

1. Introduction

A heterogeneous information network (Sun et al., 2009) (HIN) is a structured text knowledge representation consisting of a series of nodes and edges between nodes, which contains multiple node types and relation types. For instance, YAGO can be regarded as a HIN (Huang et al., 2016). As illustrated in Fig. 1, YAGO contains numerous node types and relations. With its simple and efficient knowledge representation and powerful semantic reasoning capability, HINs have become a prevalent method for knowledge storage. Consequently, the development of natural language processing technologies related to HINs has accelerated, such as entity linking (Oliveira et al., 2021; Ravi et al., 2021), named entity recognition (Nasar et al., 2021; Song et al., 2021) and relation extraction (Geng et al., 2020; Li & Tian, 2020).

The task of entity linking for heterogeneous information networks (HINs) is a critical problem in the application of such networks. This task, which involves mapping entity mentions in a text to the corresponding knowledge base (represented as a HIN in this paper), requires

[*] Corresponding authors at: Harbin Institute of Technology, 92 Xidazhi Road, Harbin, Heilongjiang 150001, China.
E-mail addresses: waiwaiyl@163.com (L. Lin), 624003414@qq.com (J. Liu).
[1] ORCID: 0000-0002-1874-1507.
[2] ORCID: 0000-0001-9525-1168.
[3] ORCID: 0000-0002-6007-1818.
[4] ORCID: 0000-0001-9465-5064.
[5] ORCID: 0000-0001-8095-5453.
[6] ORCID: 0000-0001-8725-0560.

https://doi.org/10.1016/j.eswa.2023.121759
Received 29 December 2022; Received in revised form 19 September 2023; Accepted 19 September 2023
Available online 21 September 2023

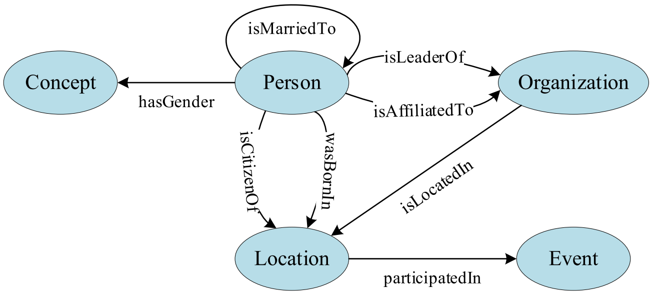


Fig. 1. A simple framework of YAGO.

finding the unique entity in the HIN that corresponds to the entity mention in order to complete the mapping process. An accurate entity linking method can effectively match the knowledge in the knowledge base that is consistent with the query target, recommending the query target and the associated knowledge to the user and thus achieving the goal of knowledge association. Moreover, such a method can improve the efficiency of knowledge retrieval (Akabe et al., 2021; Wu et al., 2022), thus providing better user experience. Furthermore, it can also facilitate the development of a variety of downstream applications, including knowledge recommendation (Xie et al., 2021; Ye et al., 2021) and knowledge reasoning (Z. Li et al., 2021).

Entity linking is divided into two parts, including: independent entity linking and collective entity linking (Shen, Wang, et al., 2014). Typically, independent entity linking (Onoe & Durrett, 2020; Wang et al., 2015) relies solely on the information of the current entity mention. Thus, independent entity linking methods typically necessitate no training and are computationally simpler than other approaches. Yet, without considering sufficient contextual information from a window or sentence, the accuracy of independent entity linking is substantially reduced, particularly when the context of the entity mention is sparse or contains considerable noise. For example, Daiber et al. developed an open source entity linking system Spotlight, allowing users to configure the system according to their specific needs (Daiber et al., 2013). Seufert et al. proposed a KORE method to estimate the semantic similarity between entities based on key phrase overlap (Hoffart et al., 2012).

Generally, collective entity linking (Liu et al., 2019; Sevgili et al., 2022) considers the relatedness between multiple entity mentions in a window or a sentence. In collective entity linking, not only the features of the current entity mention are considered, but also the relatedness to other entity mentions needs to be considered. As a result, entity mentions corroborate each other in the entity linking process. Thus, collective entity linking is completed, as shown in in Fig. 2. Nevertheless, due to the need to calculate the semantic relations between the entity mentions, collective entity linking methods often have high linking accuracy but a high computational complexity. For example, Chong et al. (Chong et al., 2017) considered that events or geographic points of interests often lead to related entities being mentioned in space and time, and used tweets that are spatially and temporally close for collective entity linking. Xia et al. (Xia et al., 2020) proposed a collective entity linking algorithm based on topic models and graphs, which combines the entity context and semantic relations between entities.

Extracting knowledge from heterogeneous information networks has become a difficult issue for improving the accuracy of entity linking. Methods based on probability models for entity linking in heterogeneous information networks is currently a hot research topic. Ganea et al. (Ganea et al., 2016) proposed a probabilistic approach PBoH that makes use of an effective graphical model to perform collective entity disambiguation. However, this method is based on certain empirical assumptions, and it ignores the relational information in the knowledge base, so it may be difficult to apply to HINs with limited information. To address the limited information of HINs, Shen et al. (Shen, Han, et al., 2014) established a general SHINE entity linking framework, which is a probability model to link named entities in web text with HINs. To further extract information form HINs, Wang et al. (Shen et al., 2017) added a knowledge population algorithm (Shen, Han, et al., 2014) to SHINE and proposed a general unsupervised framework SHINE+. While these methods provide efficient solutions, they rely on meta-paths to extract information on entity types and relation types from HINs, thereby ignoring the information of the entities and relations themselves. To address this issue, Li et al. (J. Li et al., 2021) combined the “global precedence” cognitive mechanism of the human brain with entity linking for the first time and proposed a coarse-to-fine collective entity linking method (CFEL) for heterogeneous information networks, which makes full use of the information of the entities and relations themselves.

At present, three main challenges remain for probability-based collective entity linking methods, which impede the effective extraction of information from heterogeneous information networks and further hinder the accuracy of entity linking: (1) The objective function of collective entity linking is not solved in the full meaning of “collective linking”, which may lead to the failure of collective entity linking. (2) The link parameter is based on the mean of the relatedness between candidate entities rather than the maximum, which does not highlight the important role of the strong logical relationship between text and language for the meaning, leading to difficulty in accurately measuring the strong logical associations between multiple candidate entities. (3) The contribution of entity relatedness by other types of 2-hop path is neglected in the objective function, only considering one type of 2-hop path, resulting in inaccurate measurement of entity relatedness. Thus, this paper aims to propose a collective entity linking method to address the above issues.

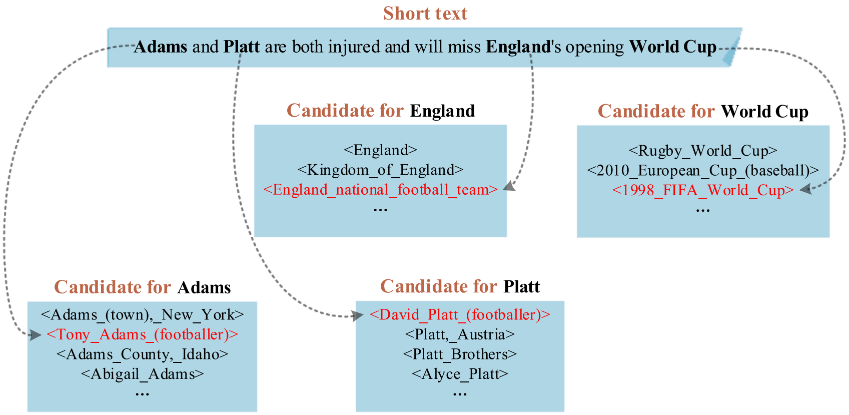


Fig. 2. An example of collective entity linking.

2. Motivation

This section begins by taking the objective function solution of the latest collective entity link method as an example to analyze three key issues of the current collective entity linking methods for heterogeneous information networks in depth. Lastly, it highlights the contributions and innovations of this paper to collective entity linking.

The state-of-the-art collective entity linking method for heterogeneous information network CFEL (J. Li et al., 2021) is employed to solve the collective entity linking objective function by taking into account the relatedness of entities and the surface similarity between entity mention and candidate entity, as expressed in Eq. (1).

$$\begin{aligned}\arg\max_{e_1,\cdots,e_q} P\left(m_1,\cdots,m_q;e_1,\cdots,e_q\right) &= \arg\max_{e_1,\cdots,e_q} P\left(e_1,\cdots,e_q\right)\times P\left(m_1,\cdots,m_q\middle|e_1,\cdots,e_q\right)\\ &= \arg\max_{e_1,\cdots,e_q} P\left(e_1,\cdots,e_q\right)\times\prod_{i=1}^{q}P\left(m_i\middle|e_i\right)\\ &= \arg\max_{e_1,\cdots,e_q}\sum_{i=1}^{q}\sum_{j=1\&j\neq i}^{q}R\left(e_i,e_j\right)\times\prod_{i=1}^{q}S\left(m_i\middle|e_i\right)\end{aligned} \tag{1}$$

where m_i is the ith entity mention in the text and e_i is a candidate entity of the entity m_i. M is the entity mention set in the text and $q = |M|$ denotes the size of the entity mention set. $R\left(e_i, e_j\right)$ is a function that measures the relatedness between entity e_i and entity e_j. $S(m_i|e_i\,)$ represents the surface similarity between the entity m_i and entity e_i.

To accurately and clearly describe the collective entity linking objective function, this paper denotes a tuple of all the entity mentions in a text as an entity mention group, and a tuple composed of one candidate entity for each entity mention as a candidate entity group. For example, in Eq. (1), $\left(m_1,\cdots,m_q\right)$ is an entity mention group and $\left(e_1,\cdots,e_q\right)$ is a candidate entity group of $\left(m_1,\cdots,m_q\right)$.

Eq. (1) indicates that the CFEL model leverages the relatedness of entities to estimate the co-occurrence probability of entities, and further computes the conditional probability between the entity mention group and the candidate entity group through the surface similarity between the entity mention and the candidate entity. Evidence has shown that the CFEL model achieves excellent entity linking performance in collective entity linking. However, there are three main challenges need to be addressed in CFEL.

(1) CFEL does not employ full meaning of “collective linking” concept in solving the objective function of collective entity linking. When CFEL estimates the conditional probability of the entity mention group and the candidate entity group $P\left(m_1,\cdots,m_q\middle|e_1,\cdots,e_q\right)$, it assumes that each entity mention independently selects the candidate entity, i.e., $P\left(m_1,\cdots,m_q\middle|e_1,\cdots,e_q\right) = \prod_{i=1}^{q}P(m_i|e_i\,)$. This calculation runs counter to the idea of “collective linking”, which essentially assumes that entity mentions $m_1,\cdots,m_q$ and $e_1,\cdots,e_q$ are independent, with no relatedness between them. However, given the logical nature of language, it is generally accepted that the semantic similarity of entity mentions in a text is usually strong. Consequently, the impact of the overall semantic similarity of entity mention group and the overall semantic similarity of candidate entity group should be considered when computing the conditional probability.

(2) CFEL employs the mean rather than the maximum value of the entity relatedness as a link parameter for solving the objective function, thus discounting the importance of the strong logical associations between the text and language for meaning recognition. Specifically, the CFEL adopts the cumulative relatedness of arbitrary candidate entity pairs to represent the overall logical associations of the entity mention group (denoted as overall relatedness), and then estimates $P\left(e_1,\cdots,e_q\right)$, i.e., $P\left(e_1,\cdots,e_q\right) = \sum_{i=1}^{q}\sum_{j=1\&j\neq i}^{q}R\left(e_i,e_j\right)$. Due to the introduction of entity pairs that are unrelated or weakly related during the computing process, the contribution of strongly-related entity pairs to the overall relatedness is weakened. It has been shown that the cumulative relatedness to represent the overall relatedness of entity mention group can easily lead to the failure of collective entity linking. For example, there is a real case in YAGO. (MANCHESTER, England, Glamorgan, Robert Croft) is an entity mention group and (<Manchester>, <England>, <Glamorgan_County_Cricket_Club>, <Robert_Croft>) and (<Manchester>, <England>, <Glamorgan_County_Cricket_Club>, <Robert_Frost>) are two candidate entity groups of the entity mention group. Moreover, the first candidate entity group is the correct candidate entity group. However, the cumulative relatedness of the first/second candidate entity group is 3.37/3.43. Obviously, the entity linking is failed because of using the cumulative relatedness to measure the overall relatedness of the candidate entity group.

(3) CFEL only adopts one type of 2-hop path (i.e., directed 2-hop) to solve the objective function when computing the relatedness of entities. This neglects the contribution of other types of 2-hop path in heterogeneous information networks to the relatedness of entities, limiting the accuracy of entity relatedness measurement. In fact, there are various types of 2-hop path between two entities, such as path in which two entities point to a common entity, path in which two entities are pointed to by a common entity, and so on. For example, Fig. 3 shows that there are three types of 2-hop paths between entities “Ming Yao” and “Li Ye”. “Ming Yao $\overset{\text{isFriendOf}}{\rightarrow}$ Na Xie $\overset{\text{isFriendOf}}{\rightarrow}$ Li Ye” and “Li Ye $\overset{\text{isFriendOf}}{\rightarrow}$ Guanxi Chen $\overset{\text{isFriendOf}}{\rightarrow}$ Ming Yao” are two directed 2-hop paths. The former is a directed 2-hop path from “Ming Yao” to “Li Ye”, while the latter is a directed 2-hop path from “Li Ye” to “Ming Yao”. “Ming Yao $\overset{\text{isSonOf}}{\rightarrow}$ Zhiyuan Yao$\longleftarrow$ isDaughter - in - lawofLi Ye” is a path that “Li Ye” and “Ming Yao” point to the common entity “Zhiyuan Yao” and “Ming Yao$\leftarrow^{\text{isDaughterof}}$ Qinlei Yao $\overset{\text{isDaughterof}}{\rightarrow}$ Li Ye” is a path that “Li Ye” and “Ming Yao” are pointed by the common entity “Qinlei Yao”. In fact, all the 2-hop paths contribute to the relatedness of two entities. “Qinlei Yao”, daughter of “Ming Yao” and “Li Ye”, plays a vital role in measuring the relatedness of “Ming Yao” and “Li Ye”. However, this aspect is not taken into consideration in CFEL, making it challenging to accurately quantify the relatedness of two entities with the directed 2-hop path.

To address the three key challenges of CFEL and improve the accuracy of collective entity linking, a novel collective entity linking method for HINs is proposed in this paper, i.e., a strong-relatedness-sequence-based fine-grained collective entity linking method (SRSCL). The innovations of our approach are as follows.

- To make the collective entity linking solution more consistent with the idea of “collective linking”, this paper develops a knowledge representation learning model to extract the semantic information of entities, and innovatively proposes an overall semantic similarity model for candidate entity group to accurately estimate the conditional probability between entity mention group and candidate entity group $P\left(m_1,\cdots,m_q\middle|e_1,\cdots,e_q\,\right)$, which is discussed in Section 3.2.4. Experimental results verify the effectiveness of considering entity semantic information.
- To emphasize the contribution of strong logical associations of entities to the overall relatedness of candidate entity group, SRSCL introduces the concept of relatedness graph and relatedness sequence, and proposes a strong-relatedness-sequence-based overall relatedness measurement model to capture the overall relatedness of the candidate entity group (as discussed in Section 3.2.2). The experimental results validate the effectiveness of the proposed strong

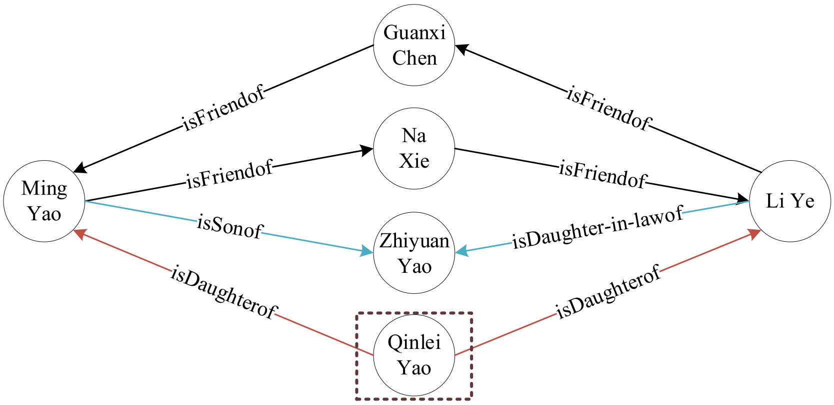


Fig. 3. Relationship between Yao Ming and Ye Li.

relatedness sequence in capturing the overall relatedness of the candidate entity group.

- To make full use of the graph structure information in heterogeneous information networks, this paper defines three types of 2-hop path to measure entity pair relatedness. The first is the path that two entities point to a common entity, the second is the path that two entities are pointed by a common entity, and the third is a directed 2-hop path between two entities. Based on the three types of 2-hop path and their respective contributions to entity relatedness, an entity pair relatedness measurement model is proposed to accurately measure

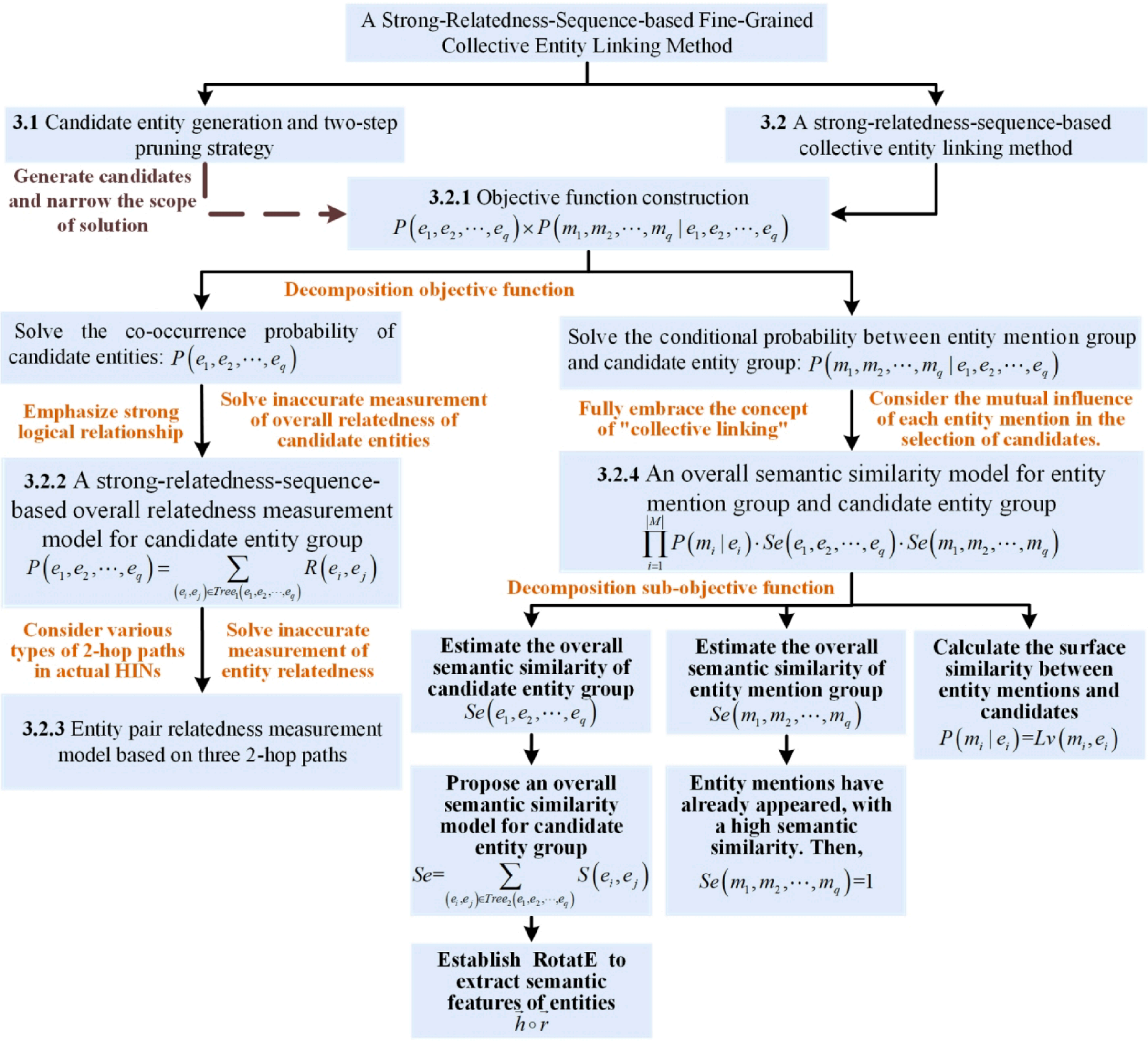

A Strong-Relatedness-Sequence-based Fine-Grained Collective Entity Linking Method
3.1 Candidate entity generation and two-step pruning strategy
3.2 A strong-relatedness-sequence-based collective entity linking method
Generate candidates and narrow the scope of solution
3.2.1 Objective function construction
$P(e_1,e_2,\cdots,e_q)\times P(m_1,m_2,\cdots,m_q|e_1,e_2,\cdots,e_q)$
Decomposition objective function
Solve the co-occurrence probability of candidate entities: $P(e_1,e_2,\cdots,e_q)$
Solve the conditional probability between entity mention group and candidate entity group: $P(m_1,m_2,\cdots,m_q|e_1,e_2,\cdots,e_q)$
Emphasize strong logical relationship
Solve inaccurate measurement of overall relatedness of candidate entities
Fully embrace the concept of "collective linking"
Consider the mutual influence of each entity mention in the selection of candidates.
3.2.2 A strong-relatedness-sequence-based overall relatedness measurement model for candidate entity group
$P(e_1,e_2,\cdots,e_q)=\sum_{(e_i,e_j)\in Tree_1(e_1,e_2,\cdots,e_q)} R(e_i,e_j)$
3.2.4 An overall semantic similarity model for entity mention group and candidate entity group
$\prod_{i=1}^{|M|} P(m_i|e_i)\cdot Se(e_1,e_2,\cdots,e_q)\cdot Se(m_1,m_2,\cdots,m_q)$
Decomposition sub-objective function
Consider various types of 2-hop paths in actual HINs
Solve inaccurate measurement of entity relatedness
3.2.3 Entity pair relatedness measurement model based on three 2-hop paths
Estimate the overall semantic similarity of candidate entity group $Se(e_1,e_2,\cdots,e_q)$
Estimate the overall semantic similarity of entity mention group $Se(m_1,m_2,\cdots,m_q)$
Calculate the surface similarity between entity mentions and candidates $P(m_i|e_i)=Lv(m_i,e_i)$
Propose an overall semantic similarity model for candidate entity group
$Se=\sum_{(e_i,e_j)\in Tree_2(e_1,e_2,\cdots,e_q)} S(e_i,e_j)$
Entity mentions have already appeared, with a high semantic similarity. Then,
$Se(m_1,m_2,\cdots,m_q)=1$
Establish RotatE to extract semantic features of entities
$\overline{h}\circ r$

Fig. 4. The structure diagram of this paper.

the relatedness of two entities (as discussed in Section 3.2.3). Experimental results verify the effectiveness of the three types of 2-hop path proposed in this paper in improving the accuracy of the entity linking.

The structure of this paper is as follows. Section 3 presents the proposed SRSCL method, comprising of two parts: (i) the candidate entity generation method, which aims to generate candidate entities and prune them to narrow down the scope of the collective entity linking problem (as discussed in Section 3.1), and (ii) a strong-relatedness-sequence-based collective entity linking method, which aims to address the three key issues for accurately solving the collective entity linking function (as discussed in Section 3.2). Section 4 introduces the experimental results and analysis. Section 5 gives the conclusion of this paper.

The structure diagram of this paper is shown in Fig. 4. Fig. 4 indicates that this paper gradually decomposes and analyzes the objective function from top to bottom. Initially, a candidate entity generation method is proposed to generate candidate entities and narrow down the solution scope. Subsequently, the objective function is decomposed into two subtasks: computing the co-occurrence probability of entities and the conditional probability between entity mention group and candidate entity group. For the former, the co-occurrence probability is calculated based on the strong relatedness sequence, which is obtained from the entity relatedness. For the latter, the conditional probability is further broken down into three sub-objectives to complete the sub-objective solution. Therefore, the collective entity linking objective function can be solved.

3. A strong-relatedness-sequence-based fine-grained collective entity linking method for heterogeneous information networks

In this section, we propose a strong-relatedness-sequence-based fine-grained collective entity-linking method (SRSCL) for heterogeneous information networks to tackle the three key issues of CFEL in collective entity linking. The proposed method consists of two parts, as follows.

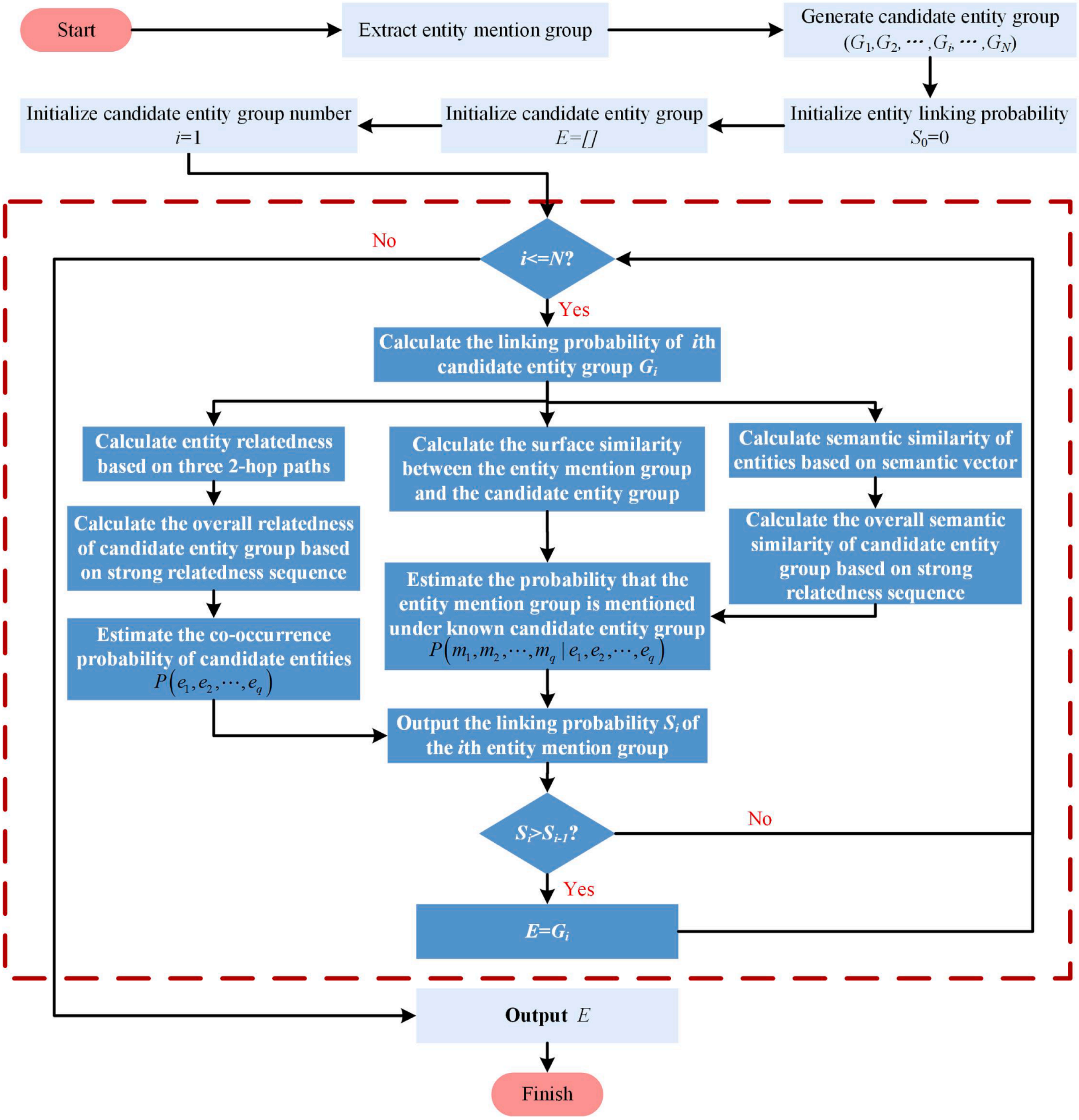

Start
Extract entity mention group
Generate candidate entity group $(G_1, G_2, \cdots, G_i, \cdots, G_N)$
Initialize entity linking probability $S_0=0$
Initialize candidate entity group $E=[]$
Initialize candidate entity group number $i=1$
$i<=N$?
No
Yes
Calculate the linking probability of ith candidate entity group G_i
Calculate entity relatedness based on three 2-hop paths
Calculate the overall relatedness of candidate entity group based on strong relatedness sequence
Estimate the co-occurrence probability of candidate entities $P(e_1, e_2, \cdots, e_q)$
Calculate the surface similarity between the entity mention group and the candidate entity group
Estimate the probability that the entity mention group is mentioned under known candidate entity group $P(m_1, m_2, \cdots, m_q \mid e_1, e_2, \cdots, e_q)$
Calculate semantic similarity of entities based on semantic vector
Calculate the overall semantic similarity of candidate entity group based on strong relatedness sequence
Output the linking probability S_i of the ith entity mention group
$S_i>S_{i-1}$?
No
Yes
$E=G_i$
Output E
Finish

Fig. 5. The collective entity linking implementation process of the SRSCL.

i) The first part is candidate entity generation and two-step pruning strategy, which is responsible for generating a set of possible candidate entities for each entity mention.

ii) The second part is a strong-relatedness-sequence-based collective entity linking method, which sorts candidate entities to identify the correct candidate entities.

The collective entity linking implementation process of the SRSCL is illustrated in Fig. 5.

3.1. Candidate entity generation and two-step pruning strategy

Candidate entity generation plays an important role in entity linking, which finds the candidate entities that may refer to the same reality as the entity mention from a given network. If the correct candidate entity is not in the entity mention's candidate entity set, an entity mention cannot be linked to the correct candidate entity. Therefore, the candidate entity generation should try to ensure that the correct entity of the entity mention is in the candidate entity set, i.e., ensuring the recall of the entity mention. In this paper, a candidate entity set E_i can be generated for the entity mention $m_i \in M$ through the following strategy.

- An entity in the HIN, which completely contained in the entity mention or completely contains the entity mention, is considered as a candidate entity of the entity mention.
- Except for unimportant words such as "of" and "I", an entity in the HIN that have common words with the entity mention is considered as a candidate entity of the entity mention.
- An entity in the HIN, which is the abbreviation of the entity mention or the full name of the entity mention, is considered as a candidate entity of the entity mention.

Nevertheless, some entity mentions have a large number of candidate entities, which greatly increases the computational complexity of collective entity linking. For example, the size of the candidate entity set of "World Cup" is 123. To reduce the complexity, it is necessary to prune the candidate entity set while ensuring that the correct candidate entity of the entity mention is still present in the new candidate entity set. The entity type can be easily obtained through entity type annotation. The entity mention type can also be obtained through language model and classification model (J. Li et al., 2021). Thus, this paper assumes that the entity mention type and the entity type in HIN are already known. To this end, a two-step pruning strategy is established to prune the candidate entity set to get the top-*k* candidate entities that are similar to the entity mention. First, the candidate entity set is pruned based on the known entity mention type. Second, the candidate entity set is pruned based on the similarity between the candidate entity and the entity mention.

- Remove entities in the candidate entity set that are of a different type than the entity mention.
- If the size of the candidate entity set exceeds k, preference is given to the entity with the abbreviation of the entity mention or the full name of the entity mention. By doing so, the top-k most similar candidate entities to the entity mention can be obtained.

3.2. A strong-relatedness-sequence-based collective entity linking method

In this section, a strong-relatedness-sequence-based collective entity linking method is proposed to solve the three key issues in the objective function of collective entity linking. Specifically, Section 3.2.2 proposes a strong-relatedness-sequence-based overall relatedness measurement model for candidate entity group to highlight the strong logical associations between natural language and accurately estimate the co-occurrence probability. Section 3.2.3 proposes an entity pair relatedness measurement model based on three types of 2-hop path to effectively utilize various types of 2-hop path in HINs, allowing for the estimation of entity relatedness with greater accuracy. Moreover, Section 3.2.4 introduces an overall semantic similarity model for entity mention group and candidate entity group to ensure that the objective function solution does not deviate from the idea of "collective linking". In this way, the conditional probability between entity mention group and candidate entity group can be estimated with greater precision. Finally, based on the three models described above, three representations of the SRSCL are proposed, as detailed in Section 3.2.4.

3.2.1. Objective function construction for collective entity linking in heterogeneous information networks

To accurately and clearly describe the collective entity linking task, this section first defines the concepts of entity mention group and candidate entity group. Subsequently, an objective function for collective entity linking towards heterogeneous information networks is constructed.

Definition 1. (*(entity mention group, candidate entity group)*) Given an entity mention set $M = \{m_1, m_2, \cdots, m_q\}$, and a set of candidate entities $E_i = \{e_{i_1}, e_{i_2}, \cdots, e_{i_k}, \cdots, e_{i_{ni}}\}$ (generated by Section 3.1) for entity mention m_i, where e_{i_k} denotes the *k*th candidate entity for m_i and *ni* denotes the number of candidate entities for m_i. In the entity linking process, a tuple of all the entity mentions in M is referred to as an entity mention group, and a tuple composed of one candidate entity for each entity mention as a candidate entity group, i.e., $(e_{1_{k1}}, e_{2_{k2}}, \cdots, e_{i_{ki}}, \cdots, e_{q_{kq}}) \in E_1 \times E_2 \times \cdots \times E_i \times \cdots \times E_q$ is an candidate entity group of the entity mention group $(m_1, m_2, \cdots, m_q)$.

Definition 2. (*(collective entity linking for HINs)*) Given an entity mention set $M = \{m_1, m_2, \cdots, m_q\}$, and a set of candidate entities $E = \{E_1, E_2, \cdots, E_q\}$, where E_i denotes the a set of candidate entities for m_i. The goal of collective entity linking is to identify a candidate entity group $(e_1, e_2, \cdots, e_q)$ from HINs that denotes the same real-world fact as the mention group $(m_1, m_2, \cdots, m_q)$.

Objective function construction for collective entity linking: given an entity mention set $M = \{m_1, m_2, \cdots, m_q\}$ of a text or a sliding window, an entity mention type set $t = \{t_1, t_2, \cdots, t_q\}$ and a candidate entity set $E^T = \{E_1^T, E_2^T, \cdots, E_q^T\}$, where t_i is the type of entity mention m_i, $E_i^T = \{(e_{i1}, t_{i1}), (e_{i2}, t_{i2}), \cdots, (e_{in}, t_{in})\}$ denotes the candidate entity set of entity mention m_i, and (e_{i1}, t_{i1}) denotes that t_{i1} is the type of candidate entity e_{i1}. After pruning, the pruned candidate entity set is $E = \{E_1, E_2, \cdots, E_q\}$, where $E_i = \{e_{i1}, e_{i2}, \cdots, e_{in}\}$. Then objective function for collective entity linking in heterogeneous information networks is given in Eq. (2).

$$\begin{aligned} & \arg \max_{e_1 \in E_1^T, \cdots, e_q \in E_q^T} P(M, T; E^T) \\ & = \arg \max_{e_1 \in E_1, \cdots, e_q \in E_q} P\big((m_1, m_2, \cdots, m_q); (e_1, e_2, \cdots, e_q)\big) \\ & = \arg \max_{e_1 \in E_1, \cdots, e_q \in E_q} P(e_1, e_2, \cdots, e_q) \times P\big((m_1, m_2, \cdots, m_q) | (e_1, e_2, \cdots, e_q)\big) \end{aligned} \tag{2}$$

where $(m_1, m_2, \cdots, m_q)$ is an entity mention group of M and m_i is the *i*th entity mention. $(e_1, e_2, \cdots, e_q)$ is a candidate entity group of $(m_1, m_2, \cdots, m_q)$, and e_i is an element of E_i. $P\big((m_1, m_2, \cdots, m_q); (e_1, e_2, \cdots, e_q)\big)$ denotes the probability that the entity mention group $(m_1, m_2, \cdots, m_q)$ refers to the candidate entity group $(e_1, e_2, \cdots, e_q)$. $P(e_1, e_2, \cdots, e_q)$ indicates the probability that $(e_1, e_2, \cdots, e_q)$ co-occurrence in the same text. Additionally, $P\big((m_1, m_2, \cdots, m_q) | (e_1, e_2, \cdots, e_q)\big)$ refers to the conditional probability of $(m_1, m_2, \cdots, m_q)$ when $(e_1, e_2, \cdots, e_q)$ is known. The calculation of $P\big((m_1, m_2, \cdots, m_q) | (e_1, e_2, \cdots, e_q)\big)$ is described in detail as below.

3.2.2. *A strong-relatedness-sequence-based overall relatedness measurement model for candidate entity group*

In this section, a strong-relatedness-sequence-based overall relatedness measurement model is proposed to emphasize the strong logical associations between text and language, as well as to accurately estimate the co-occurrence probability of candidate entity group. This model addresses the shortcomings of traditional methods which rely on accumulative the relatedness of every two candidate entities to measure overall relatedness. By doing so, the model highlights the contribution of entities with strong relatedness to the overall relatedness of the candidate entity group.

According to the logical characteristics of human language, entity mentions in the same text are highly interrelated. Thus, the entity mention group of a text can be considered as a whole. Because heterogeneous information networks are constructed by extracting large amounts of textual information, the candidate entity group corresponding to entity mention group can also be regarded as a whole. Generally, if the overall relatedness of the candidate entity group is stronger, the co-occurrence probability is greater. However, it is difficult to effectively measure the overall relatedness of candidate entity group.

To address the above problems, this paper proposes the concept of relatedness graph and strong relatedness sequence. The strong relatedness sequence is designed to weaken the influence of unrelated entity pairs or weakly related entities on the overall relatedness of the candidate entity group. Therefore, the contribution of strong related entity pairs to the overall relatedness of candidate entity group is highlighted.

Definition 3. (*(relatedness graph, relatedness sequence and strong relatedness sequence)*) Given a candidate entity group $(e_1, e_2, \cdots, e_q)$ and the relatedness between each two candidate entities, a q-order relatedness graph is a q-order undirected complete graph (Biggs, 2002), in which the edge between two entities is represented by their relatedness. A relatedness sequence is defined as a connected spanning subgraph of the relatedness graph, and its edge set contains $q-1$ edges. The relatedness sequence with the highest cumulative relatedness in the relatedness graph is termed the strong relatedness sequence.

Fig. 6 depicts a 4-order relatedness graph and its two relatedness sequences (i.e. the red solid line and the blue solid line). And the red relatedness sequence has the highest relatedness among all the relatedness sequences in the 4-order relatedness graph, with a relatedness of 2.4. It is noteworthy that the *strong relatedness sequence* does not necessarily have to be a path, as demonstrated by the blue solid line in Fig. 6, which is a tree.

Based on the strong relatedness sequence, this paper proposes a strong-relatedness-sequence-based overall relatedness measurement model. Moreover, this model is effective in estimating the co-occurrence probability of the candidate entity group, which is given in Eq. (3).

$$P(e_1, e_2, \cdots, e_q) = \sum_{(e_i, e_j) \in Tree_1(e_1, e_2, \cdots, e_q)} R(e_i, e_j) \tag{3}$$

where $Tree_1(e_1, e_2, \cdots, e_q)$ denotes the strong relatedness sequence of entity $e_1, e_2, \cdots, e_q$ and $R(e_i, e_j)$ denotes the relatedness between entity e_i and entity e_j.

As shown in Fig. 7, a variety of relatedness sequences are present in the 4-order relatedness graph. Enumerating all relatedness sequences is difficult, therefore, three relatedness sequences are listed in Fig. 7. The subgraph (a), (b) and (c) represent the first, second and third relatedness sequences, respectively. And (d) represents other relatedness sequences. The goal is to identify the strong relatedness sequence. To this end, the Maximum Weight Spanning Tree Algorithm can be employed, in accordance with the principles of graph theory, to obtain the strong relatedness sequence from the relatedness graph.

To verify the efficacy of the strong-relatedness-sequence-based overall relatedness measurement model, we compare it with the traditional method in CFEL for calculating the overall relatedness of the candidate entity group. Specifically, we measure the performance of both methods in terms of their ability to capture the overall relatedness. For example, the two methods are applied to the real case of YAGO introduced in Section 2. By the proposed model in this paper, the overall relatedness of the candidate entity group (<Manchester>, <England>, <Glamorgan_County_Cricket_Club>, <Robert_Croft>) is 1.39, and the overall relatedness of the candidate entity group (<Manchester>, <England>, <Glamorgan_County_Cricket_Club>, <Robert_Frost>) is 1.34. The calculation results demonstrate that the strong-relatedness-sequence-based overall relatedness measurement model can effectively complete the entity linking. Compared to the method used in the CFEL, the strong relatedness sequence can more accurately measure the overall relatedness of the candidate entity group, leading to improved accuracy of the collective entity linking.

3.2.3. *Entity pair relatedness measurement model based on three 2-hop path*

In this section, we address the problem of accurately quantifying entity relatedness using only one types of 2-hop path in the CFEL. To this end, we analyze heterogeneous information networks and propose a novel entity pair relatedness measurement model based on three 2-hop path. Specifically, we define three types of 2-hop path and measure the contribution of each type of 2-hop path to the entity relatedness.

It is necessary to obtain the relatedness of each two entities in advance when calculating the co-occurrence probability of a candidate entity group. Previous studies have verified that the 1-hop and 2-hop path of entities can effectively capture the relatedness information between entities. However, longer paths will bring more noise and reduce the relatedness information between entities (J. Li et al., 2021). Meanwhile, CFEL only adopts one type of 2-hop path (i.e., directed 2-hop path) to measure the relatedness of two entities. In fact, there are several types of 2-hop path connecting two entities, such as path in which the two entities point to a common entity. Importantly, these various 2-hop path have different contributions to accurately measure the relatedness of two entities. For example, “Qinlei Yao” is the daughter of “Ming Yao” and “Li Ye”, “Ming Yao” is closely connected with “Li Ye” through “Qinlei Yao”. CFEL ignores the contribution of “Qinlei Yao” when calculating the relatedness between “Ming Yao” and “Li Ye”, which hinders its accuracy in measuring the relatedness between entity pair.

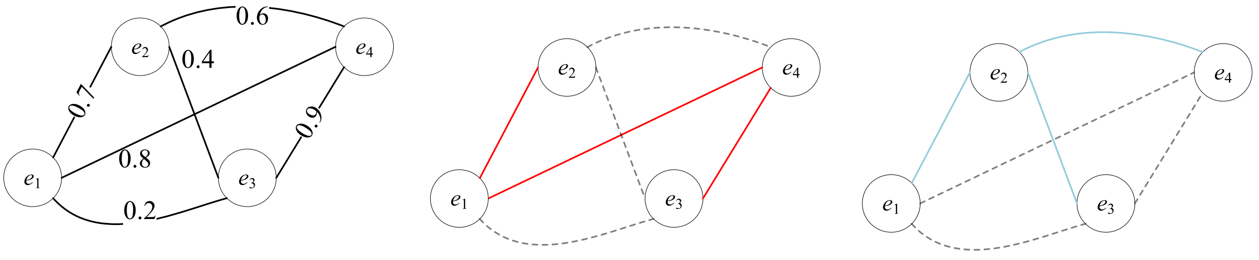
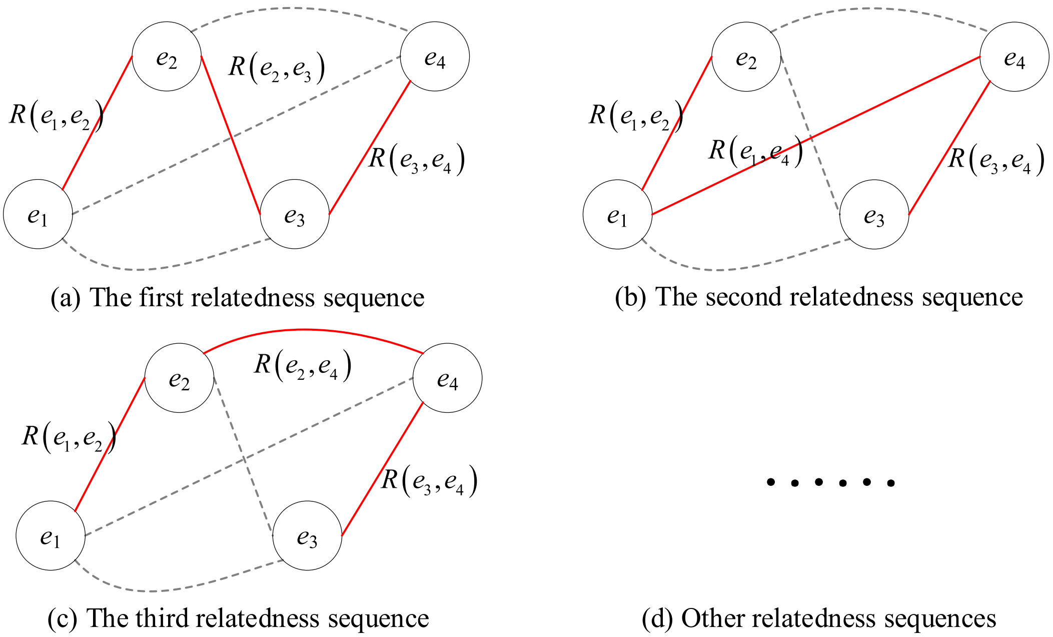


Fig. 6. Examples of relatedness graph and its two relatedness sequences.



(a) The first relatedness sequence

(b) The second relatedness sequence

(c) The third relatedness sequence

(d) Other relatedness sequences

Fig. 7. Relatedness sequences in a 4-order relatedness graph.

In this paper, we divide 2-hop path into three types. The first is the path that two entities point to a common entity, the second is the path that two entities are pointed to by a common entity, and the last is the path that a directed 2-hop path exists between two entities (i.e., used in CFEL, which is marked with a wireframe in Fig. 8). Fig. 8 shows three possible types of 2-hop path between entity e_i and entity e_j. Actually, there are three types of 2-hop path between two entities in YAGO. For example, there are three types of 2-hop path between the entity <*Norway*> and the entity <*Spain*>. Specifically, <*Norway*>→<*Midajah*>←<*Spain*> is the first 2-hop path, <*Norway*>←<*Germany*>→<*Spain*> is the second 2-hop path and <*Norway*>→<*France*>→<*Spain*> is the third 2-hop path. The three types of path all contribute to the relatedness of <*Norway*> and <*Spain*>. CFEL only adopts the third 2-hop to measure the relatedness between <*Norway*> and <*Spain*>, ignoring the contribution of <*Midajah*> and <*Germany*> to the relatedness of <*Norway*> and <*Spain*>. Therefore, it is difficult for CFEL to accurately measure the relatedness of two entities.

The longer the path between two entities is, the smaller the relatedness of two entities is. Conversely, the more 2-hop path between two entities are, the greater the relatedness of two entities is. To accurately measure the relatedness of two entities, three types of 2-hop path and 1-hop paths (i.e., two candidate entities are directly connected by an edge) are adopted. Thus, an entity pair relatedness measurement model based on three 2-hop path is proposed in this paper, as given in Eq. (4).

$$R(e_i,e_j) = \begin{cases} 1 & if\ (e_i,r,e_j)or(e_j,r,\ e_i) \\ R_2(e_i,e_j) & otherwise \end{cases} \tag{4}$$

$$R_2(e_i,e_j) = \frac{1}{l_{e_i,e_j}}\cdot\left(\frac{\beta\cdot\left(\left|B_{e_i,e_j}\right| + \left|T_{e_i,e_j}\right|\right) + (1-\beta)\cdot\left|N_{e_i,e_j}\right|}{\left|N_{e_i}\right| + \left|T_{e_i,e_j}\right|} + \frac{\beta\cdot\left(\left|B_{e_j,e_i}\right| + \left|T_{e_j,e_i}\right|\right) + (1-\beta)\cdot\left|N_{e_j,e_i}\right|}{\left|N_{e_j}\right| + \left|T_{e_j,e_i}\right|}\right) \tag{5}$$

where $R(e_i,e_j)$ is the relatedness measure function of two entities and $R_2(e_i,e_j)$ is the relatedness measure function based on 2-hop path. l_{e_i,e_j} is the path length from e_i to e_j (here $l_{e_i,e_j} = 2$). $|N_{e_i}|$ denotes the number of entities pointed to by e_i and $|N_{e_j}|$ denotes the number of entities pointed

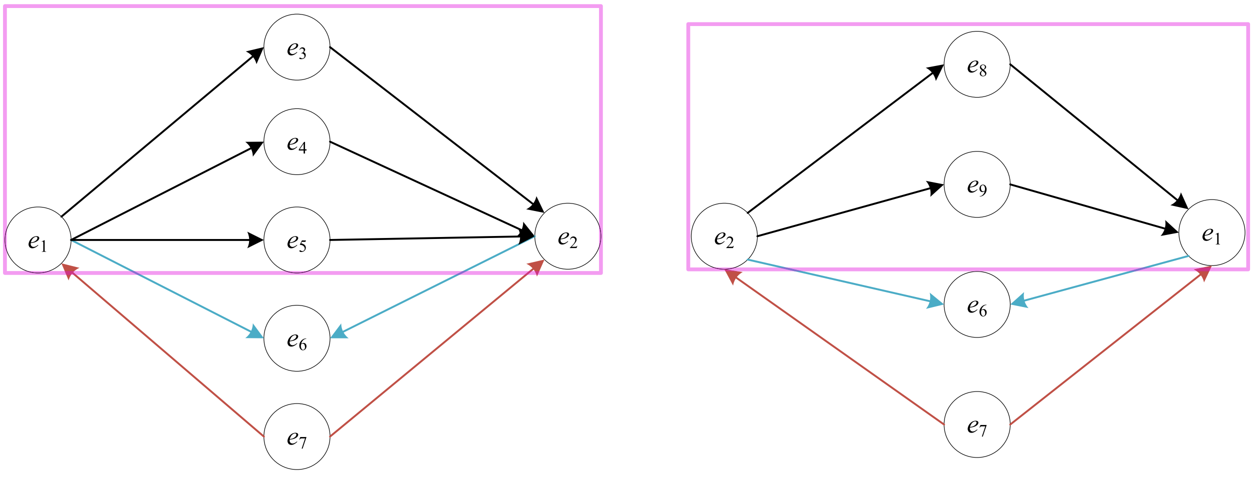


Fig. 8. Three types of 2-hop path between two entities.

to by e_j. $|B_{e_i,e_j}|$ and $|T_{e_i,e_j}|$ denote the number of the first type of 2-hop path and the second type of 2-hop path respectively, where $|B_{e_i,e_j}| = |B_{e_j,e_i}|$, $|T_{e_i,e_j}| = |T_{e_j,e_i}|$. $|N_{e_i,e_j}|$ denotes the number of the third type of 2-hop path from e_i to e_j and $|N_{e_j,e_i}|$ denotes the number of the third type of 2-hop path from e_j to e_i. β is an adjustable hyperparameter that quantifies the relative importance of the directed 2-hop path with respect to the other two types of 2-hop path.

If the relatedness between all entities is 0 through the entity pair relatedness measurement model (i.e., $P(e_1, e_2, \cdots, e_q) = 0$), which results in the value of the object Eq. (2) being 0 and renders it difficult to distinguish between multiple different candidate entity groups. To solve the above problem, a hyperparameter, γ, greater than 0 is added to Eq. (4). Then Eq. (4) can be expressed as Eq. (6).

$$R(e_i, e_j) = \begin{cases} 1 & if\ (e_i, r, e_j) or (e_j, r,\ e_i) \\ R_2(e_i, e_j) + \gamma & otherwise \end{cases} \tag{6}$$

3.2.4. An overall semantic similarity model for entity mention group and candidate entity group

To maintain the “collective linking” idea in the collective entity linking process, this section establishes a RotatE model to extract semantic information from HINs, and proposes an overall semantic similarity measurement model to accurately estimate the conditional probability (i.e., $P((m_1, \cdots, m_q) | (e_1, \cdots, e_q))$) of the entity mention group and the candidate entity group.

CFEL assumes that an entity mention independently selects an entity in HINs when calculating $P((m_1, \cdots, m_q) | (e_1, \cdots, e_q))$, i.e., $P((m_1, m_2, \cdots, m_q) | (e_1, e_2, \cdots, e_q)) = \prod_{i=1}^{|M|} P(m_i|e_i)$. However, this assumption is not necessarily valid, since according to the logic of human language, entity mentions in a sentence are semantically related to each other. In fact, if the overall semantic similarity of the entity mention group and the overall semantic similarity of the candidate mention group are greater, the conditional probability $P((m_1, \cdots, m_q) | (e_1, \cdots, e_q))$ is greater. Therefore, Eq. (7) is proposed to effective estimate the conditional probability.

$$P((m_1, \cdots, m_q) | (e_1, \cdots, e_q)) = \prod_{i=1}^{|M|} P(m_i|e_i) \cdot Se(e_1, e_2, \cdots, e_q) \cdot Se(m_1, m_2, \cdots, m_q) \tag{7}$$

where $|M|$ denotes the size of entity mention set M and $Se(*)$ is an overall semantic similarity function for the candidate entity group $(e_1, e_2, \cdots, e_q)$ and the entity mention group $(m_1, m_2, \cdots, m_q)$. Furthermore, $P(m_i|e_i)$ indicates the probability that the entity mention m_i selects the candidate entity e_i.

In entity linking, entity mention group is generally extracted from known texts or sentences. And its semantic similarity is strong. Therefore, the overall semantic similarity of the entity mention group is regarded as 1. To emphasize the overall semantic similarity of candidate entity group, the strong relatedness sequence is used to estimate the overall semantic similarity of the candidate entity group. Further, Eq. (8) is obtained through the simplification of Eq. (7).

$$\begin{aligned} P((m_1, \cdots, m_q) | (e_1, \cdots, e_q)) &= \prod_{i=1}^{|M|} P(m_i|e_i) \cdot Se(e_1, e_2, \cdots, e_q) \\ &= \prod_{i=1}^{|M|} P(m_i|e_i) \cdot Tree_2(e_1, e_2, \cdots, e_q) \\ &= \prod_{i=1}^{|M|} P(m_i|e_i) \cdot \sum_{(e_i,e_j) \in Tree_2(e_1,e_2,\cdots,e_q)} S(e_i, e_j) \end{aligned} \tag{8}$$

where $S(e_i, e_j)$ denotes the semantic similarity between entity e_i and entity e_j, and $Tree_2(e_1, e_2, \cdots, e_q)$ represents the overall semantic similarity of the strong relatedness sequence obtained by $S(e_i, e_j)$. The methods used to calculate $P(m_i|e_i)$ and $S(e_i, e_j)$ are described in detail below.

Theoretically, $P(m_i|e_i)$ can be calculated through word frequency. However, obtaining all web documents related to m_i and e_i is impractical. On the one hand, the information of the entity e_i and entity mention m_i are constantly changing on the web. On the other hand, it is difficult to query the information of some entities and entity mentions on the web. Thus, the method based on word frequency cannot be directly used to calculate $P(m_i|e_i)$. Actually, the higher the surface similarity between entity mention m_i and candidate entity e_i is, the higher $P(m_i|e_i)$ is. To this end, the Levenshtein distance is adopted to calculate the similarity between the entity mention m_i and the candidate entity e_i, as given in Eq. (9).

$$\begin{aligned} &\prod_{i=1}^{|M|} P(m_i|e_i) = \prod_{i=1}^{|M|} Lv(m_i, e_i) \\ &Lv(m_i, e_i) = 1 - \frac{lev(m_i, e_i)}{\max(|m_i|, |e_i|)} \end{aligned} \tag{9}$$

where $|m_i|$ and $|e_i|$ denote the character length of m_i and e_i respectively. $\mathrm{lev}(m_i, e_i)$ denotes the Levenshtein distance between the entity mention m_i and the candidate entity e_i, and $Lv(m_i, e_i)$ denotes the similarity between the entity mention m_i and the candidate entity e_i.

Next, the method to effectively obtain the semantic similarity (i.e., $S(e_i, e_j)$) between entity e_i and entity e_j is analyzes in detail.

Currently, knowledge representation model has been greatly developed, which can effectively learn the context information of entities and represent entities with semantic vectors. To effectively calculate the semantic similarity between candidate entities, an advanced knowledge representation model, RotatE (Sun et al., 2019), is adopted to learn the embedding representation of entities and relations in HINs. The rotation operation of the RotatE is illustrated in Fig. 9. Experimental results show that RotatE is effective in solving the complex relationship problems of knowledge graphs, including symmetric/antisymmetric, inversion and composition. Compared with traditional models such as TransE (Bordes et al., 2013), TransH (Wang et al., 2014), ComplEx (Trouillon et al., 2016), and others, RotatE has superior performance in link prediction. Therefore, RotatE can capture the semantic features of entities and relations in HINs more fully, that is, the entity embedding vectors obtained using RotatE can represent the entities. (Fig. 10).

According to the embedding vectors of entities, the semantic similarity between any two entities are quantified using cosine similarity or Euclidean distance, as expressed in Eq. (10). For example, under RotatE, the embedding vector of the entity “<1._FC_Köln>” is [0.012527352198958397,…,-0.0019661542028188705] + [-0.0177767332643 2705,…,0.03048780746757984]i and the embedding vector of the entity “<1._FC_Saarbrücken>” is [0.013352576643228531,…,0.0019227 27096825838] + [-0.017329551279544 83,…,0.026032166555523872]i. Then, the cosine similarity between the two entities is 0.215.

$$S(e_i, e_j) = Sim(\overrightarrow{e_i}, \overrightarrow{e_j}) \tag{10}$$

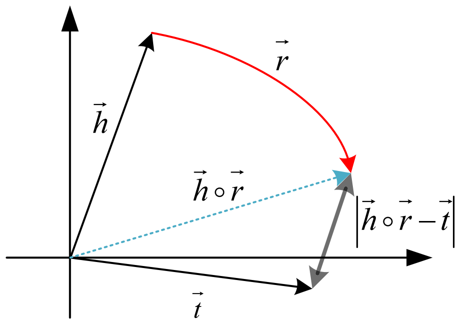


Fig. 9. The framework of the RotatE.

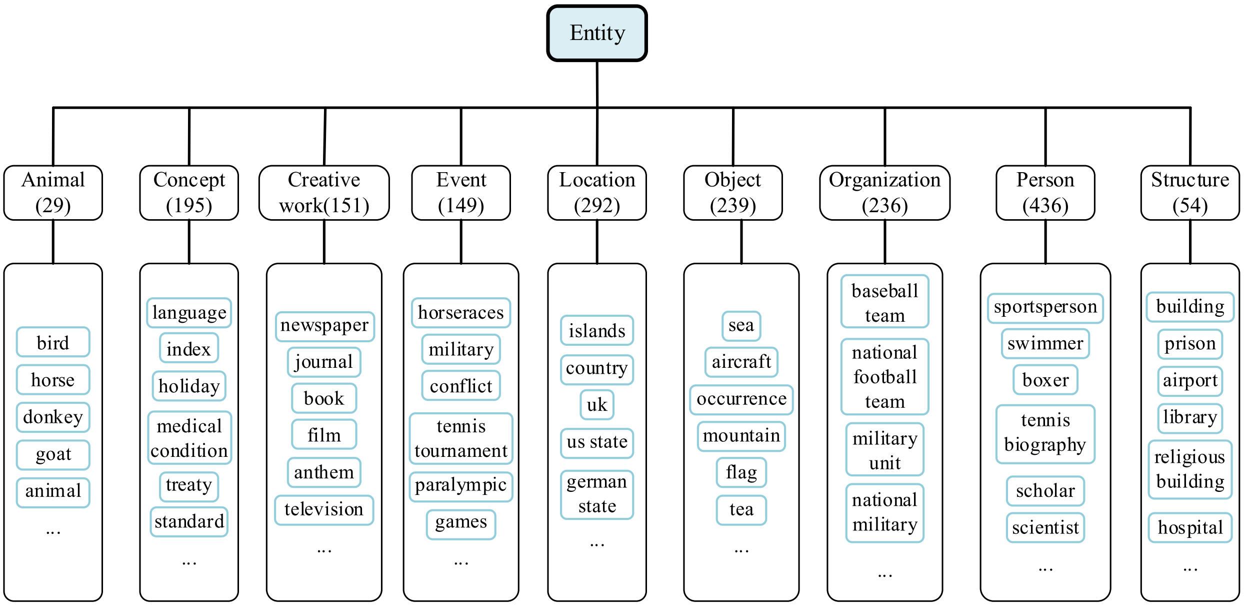


Fig. 10. Subcategories division in YAGO.

where $\overrightarrow{e_i}$ denotes the embedding vector of the entity e_i, and $Sim\left(\overrightarrow{e_i},\overrightarrow{e_j}\right)$ is the similarity function for calculating semantic similarity between the entity $\overrightarrow{e_i}$ and the entity $\overrightarrow{e_j}$, such as cosine similarity.

3.2.5. Three representations of strong-relatedness-sequence-based fine-grained collective entity linking method

Actually, the acquisition of entity relatedness and semantic information are both based on the structure of HINs. Consequently, it is generally difficult to distinguish between semantic similarity and relatedness. This means that Eq. (10) and Eq. (4) can be interchanged to some extent. Based on the entity pair relatedness measurement model and the semantic similarity measurement model, three representations of strong-relatedness-sequence-based fine-grained collective entity linking methods can be obtained, as given in Eqs. (11)–(13).

$$\begin{aligned}&\arg\max_{e_1\in E_1,\cdots,e_q\in E_q} P\left(e_1,e_2,\cdots,e_q\right)\times Se\left(e_1,e_2,\cdots,e_q\right)\cdot\prod_{i=1}^{|M|}P(m_i|e_i)\\ =&\arg\max_{e_1\in E_1,\cdots,e_q\in E_q} Tree_1\left(e_1,e_2,\cdots,e_q\right)\times Tree_1\left(e_1,e_2,\cdots,e_q\right)\cdot\prod_{i=1}^{|M|}Lv(m_i,e_i)\end{aligned} \tag{11}$$

$$\begin{aligned}&\arg\max_{e_1\in E_1,\cdots,e_q\in E_q} P\left(e_1,e_2,\cdots,e_q\right)\times Se\left(e_1,e_2,\cdots,e_q\right)\cdot\prod_{i=1}^{|M|}P(m_i|e_i)\\ =&\arg\max_{e_1\in E_1,\cdots,e_q\in E_q} Tree_1\left(e_1,e_2,\cdots,e_q\right)\times Tree_2\left(e_1,e_2,\cdots,e_q\right)\cdot\prod_{i=1}^{|M|}Lv(m_i,e_i)\end{aligned} \tag{12}$$

$$\begin{aligned}&\arg\max_{e_1\in E_1,\cdots,e_q\in E_q} P\left(e_1,e_2,\cdots,e_q\right)\times Se\left(e_1,e_2,\cdots,e_q\right)\cdot\prod_{i=1}^{|M|}P(m_i|e_i)\\ =&\arg\max_{e_1\in E_1,\cdots,e_q\in E_q} Tree_2\left(e_1,e_2,\cdots,e_q\right)\times Tree_2\left(e_1,e_2,\cdots,e_q\right)\cdot\prod_{i=1}^{|M|}Lv(m_i,e_i)\end{aligned} \tag{13}$$

where $Tree_1\left(e_1,e_2,\cdots,e_q\right)$ denotes the overall relatedness of candidate entity group as measured by the strong relatedness sequence of 2-hop path, and $Tree_2\left(e_1,e_2,\cdots,e_q\right)$ represents the overall semantic similarity of candidate entity group as measured by the strong relatedness sequence of the semantic similarity.

3.3. Analysis of algorithm time complexity

In the implementation of SRSCL, the information of three types of 2-hop path is stored in a dictionary. According to the time complexity of dictionary query, the time complexity of three types of 2-hop path is $O(1)$. Similarly, the time complexity of one type of 2-hop path is $O(1)$. Therefore, compared with the CFEL model, taking three types of 2-hop path into consideration does not increase the time complexity. Similar to the CFEL, the time complexity of SRSCL is mainly reflected in calculating the relatedness of candidate entity group.

Given that the number of entity mentions in a sliding window is N, and the number of candidate entities for the ith entity mention is M_i, then a total of $M_1*M_2*\cdots*M_N$ entity mention groups is produced. According to the Prime algorithm in the maximum weight spanning tree, if the number of nodes in the graph is S, then the time complexity of the maximum weight spanning tree is $O\left(S^2\right)$. Therefore, if the number of entity mentions is N, the time complexity of calculating the overall relatedness or overall semantic similarity of an entity mention group is $O\left(N^2\right)$. Further, the time complexity of obtaining the optimal candidate entity group is $O\left(M_1*M_2*\cdots*M_N*N^2\right)$. According to the candidate entity pruning strategy, if the candidate entity is taken as top-k, the time complexity of obtaining the optimal candidate entity group is less than $O\left(k^N*N^2\right)$.

4. Experiment

4.1. Experimental data

To validate the effectiveness of the SRSCL proposed in this paper, part of the data in YAGO-Core is adopted as the HIN in entity linking, and AIDA CoNLL-YAGO, ACE2004 and AQUANT (Hoffart et al., 2011) are adopted as the datasets in entity linking. The detailed descriptions of the adopted HIN and datasets are as follows.

The YAGO 3.1 dataset is downloaded and the YAGO-Core is regarded as the HIN in the experiment in this paper. Considering the large scale YAGO-Core, a part of YAGO-Core (J. Li et al., 2021) is adopted as the HIN for the entity linking experiment. Specifically, 10 K entities and 770 K triples in the YAGO-Core are treated as the HIN in the experiment. Moreover, the public datasets AIDA CoNLL-YAGO, ACE2004 and AQUANT are employed to conduct the entity linking experiment on the above mentioned HIN. Specifically, AIDA CoNLL-YAGO contains 1393 articles related to the topics of country, government, sports events, etc. ACE2004, developed by the Linguistic Data Consortium (LDC), consists of various types of text data in English, Chinese and Arabic. Lastly. And AQUANT is composed of text data from three English news networks, including the Xinhua News Agency (the People's Republic of China), the New York Times News Agency and the Associated Press World Stream News Agency.

The processed datasets are summarized in Table 1, which includes the number of entity mentions, the number of entity mentions that have the corresponding entities in the HIN, and the number of entity mentions that have candidate entities via the candidate entity generation method. For instance, the AIDA CoNLL-YAGO dataset contains 34,929 entity mentions, among which 20,228 have the corresponding entities in the HIN and 19,716 have candidate entities via the candidate entity generation method. Consequently, only 20,228 entity mentions are employed in the collective entity linking experiment.

In the experiment, the above public datasets are processed in this paper. Specifically, all entity mentions whose corresponding entities are not present in the constructed HIN are deleted. The processed datasets are summarized in Table 1, which includes the number of entity mentions, the number of entity mentions that have the corresponding entities in the HIN, and the number of entity mentions that have candidate entities via the candidate entity generation method. For instance, the AIDA CoNLL-YAGO dataset contains 34,929 entity mentions, among which 20,228 have the corresponding entities in the HIN and 19,716 have candidate entities via the candidate entity generation method.

4.2. Experimental data preprocessing and sample generation

This section processes the HIN and entity linking datasets employed in the experiment in order to carry out entity linking experiments. Specifically, it includes: (1) category annotations of entities in the HIN. (2) entity mention group generation as sample input for entity linking.

(1) Category annotations of entities in the HIN

In the experiment, the candidate entity set needs to be pruned through entity types. Therefore, the entities in the HIN are labeled with category labels. Based on the work of some predecessors and the entity classification method in the Reference(Kalender et al., 2017), entities are manually divided into 9 categories, including: "Animal", "Concept", "Creative work", "Event", "Location", "Object", "Organization", "Person" and "Structure". In fact, entities in the YAGO dataset already possess corresponding entity subcategory labels, such as "officeholder", "ethnic group", "person", "mountain" and so on. Furthermore, the subcategories are divided into their respective categories. For example, "officeholder" is belong to "Person" and the type of the entity belonging to "officeholder" is labeled as "Person". "ethnic group" is belong to "Organization" and the type of the entity belonging to "ethnic group" is labeled as "Organization". Specifically, 1798 subcategories of entities are divided into 9 categories, as shown in Fig. 7. Eventually, "Animal", "Concept", "Creative work", "Event", "Location", "Object", "Organization", "Person" and "Structure" contain 29, 195, 151, 149, 292, 239, 236, 436 and 54 subcategories respectively.

Table 1
The processed datasets.

Dataset	The number of entity mentions	The number of entity mentions that have the corresponding entities in the HIN	The number of entity mentions that have candidate entities
AIDA CoNLL-YAGO	34,929	20,228	19,716
AQUANT	727	186	180
ACE 2004	257	125	124

(2) Sample generation

In general, the length of sliding window needs to be set to obtain the entity mention group in collective entity linking, whereby all entity mentions in a sliding window form an entity mention group. Consequently, collective entity linking is achieved by computing the relatedness of the entity mention group. An example of this extraction process is illustrated Fig. 11, where the sliding window is set to 10. The corresponding entity of the entity mention is identified in YAGO by using its "Wikipedia URL" and the type of the corresponding entity is used to represent the entity mention type. Finally, the entity mention group of the first sliding window is represented in Fig. 11.

4.3. Model evaluation

In the evaluation of entity linking models, precision, recall and F1 score are typically used as metrics. However, when it comes to collective entity linking, the use of these metrics alone is not sufficient for effectively assessing the ability of the model to detect relatedness of candidate entity group. To address this, this section proposes a $M-k$ method for evaluating the model's effectiveness in capturing the overall relatedness of candidate entity group, in addition to precision, recall and F1 score.

In the experiment, precision, recall and F1 score are adopted to evaluate the performance of entity linking, which is shown in Eq. (14), Eq. (15) and Eq. (16)

$$P = \frac{correct\_link}{processed\_mentions} \tag{14}$$

$$R = \frac{correct\_link}{total\_mentions} \tag{15}$$

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{16}$$

where *correct_link* denotes the number of correctly linked entity mentions, and *processed_mentions* denotes the number of entity mentions that have candidate entities. The *total_mentions* denotes the number of entity mentions used in the experiment. For example, *processed_mentions* is 19,716 and *total_mentions* is 20,228 in AIDA CoNLL-YAGO.

Collective entity linking focuses more on the performance of entity linking methods in capturing the overall relatedness of candidate entity group. However, there is a lack of evaluation methods to estimate the performance of collective entity linking methods in this regard. To address this issue, this paper proposes an evaluation method, called $M-k$ (*Mention-k*) method. It is noteworthy that $M-k$ is based on the number of entity mentions contained in the sliding window. Specifically, $M-k$ first picks out entity mention groups whose number of entities with candidate entities is greater than k, then evaluates the effect of the collective entity linking method on these entity mention groups. Since the number of entity mentions with candidate entities in the sliding window is greater than 1, the relatedness between candidate entities of entity mentiones can be calculated. Finally, precision, recall rate and F1 score are adopted to represent the entity linking accuracy in $M-k$, as shown in Eq. (17), Eq. (18) and Eq. (19).

$$P = \frac{correct\_link\_k}{processed\_mentions\_k} \tag{17}$$

$$R = \frac{correct\_link\_k}{total\_mentions\_k} \tag{18}$$

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{19}$$

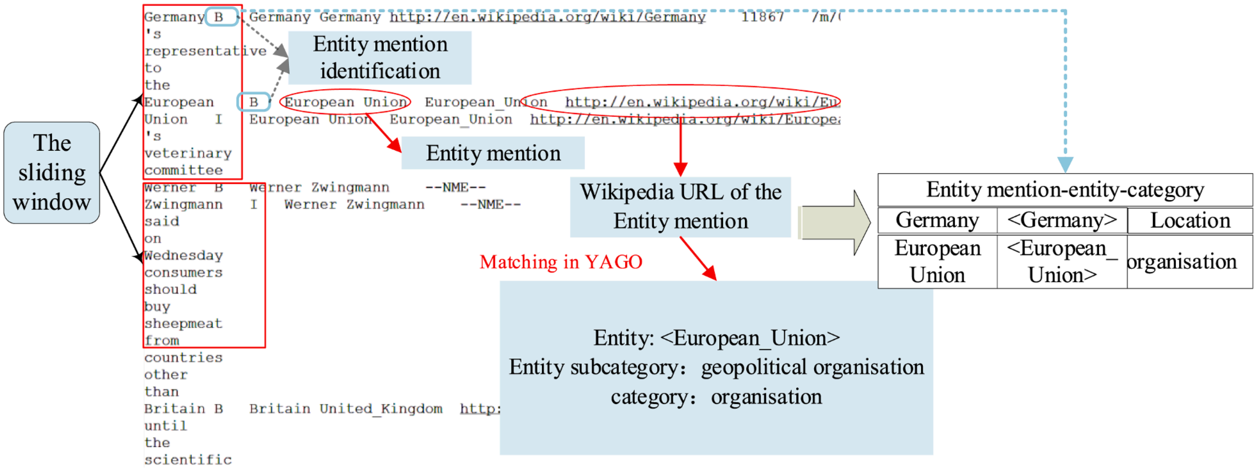

The sliding window
Entity mention identification
Entity mention
Wikipedia URL of the Entity mention
Matching in YAGO
Entity: <European_Union>
Entity subcategory: geopolitical organisation
category: organisation
Entity mention-entity-category
Germany
<Germany>
Location
European Union
<European_Union>
organisation
Germany B Germany Germany http://en.wikipedia.org/wiki/Germany 11867 /m/0
's
representative
to
the
European B European Union European_Union http://en.wikipedia.org/wiki/Eu
Union I European Union European_Union http://en.wikipedia.org/wiki/Europe
's
veterinary
committee
Werner B Werner Zwingmann --NME--
Zwingmann I Werner Zwingmann --NME--
said
on
Wednesday
consumers
should
buy
sheepmeat
from
countries
other
than
Britain B Britain United_Kingdom http:
until
the
scientific

Fig. 11. The entity mention group construction process for collective entity linking.

where *correct_link_k* denotes the number of entity mentions correctly linked in $M-k$ and *processed_mentions_k* denotes the number of entity mentions that have candidate entities in $M-k$. The *total_mentions_k* denotes the total number of entity mentions in $M-k$.

4.4. Experiment results and analysis

To validate the efficacy of the proposed SRSCL, comparative experiments of entity linking are conducted on three public datasets using the SRSCL and five baselines. In the experiments, precision, recall and F1 score are adopted to evaluate the entity linking effectiveness, and $M-k$ is adopted to evaluate the overall relatedness of the candidate entity group captured by the models. Therefore, this section is divided into two parts. The first part is to verify the efficacy of the SRSCL against the baselines on AIDA-CoNLL, YAGO, AQUANT and ACE2004 datasets. The second is to analyze the influence of different hyperparameters (e.g., β, γ, the size of candidate entity set and sliding window length.) on the entity linking performance of the SRSCL.

4.4.1. Comparision experiment

The comparison experiment is divided into two parts: the comparison of the proposed SRSCL with the baselines with respect to entity linking effect, and the comparison of the proposed SRSCL with the baselines regarding their capability of capturing the overall relatedness of candidate entity group.

(1) The comparison experiment in the entity linking effect

In this paper, AIDA CoNLL-YAGO, AQUAINT and ACE2004 are utilized to evaluate the effectiveness of the proposed SRSCL model. Additionally, SRSCL is compared with five entity linking methods/models, namely the classical entity popularity model (POP) (Shen, Han, et al., 2014), the latest collective entity linking method (CFEL) (J. Li et al., 2021), EMDD, TTHP, and CFELS. The details of each model/method for heterogeneous information network information extraction and overall relatedness capturing of candidate entity group are presented in Table 2. The POP model calculates the popularity of each entity by extracting degree information from the HIN and ranking the candidate entities accordingly. The CFEL model computes the relatedness between entities through the directed 2-hop path, representing the co-occurrence probability between entities, and eventually completes the entity linking through a probability model. To further improve the relatedness measurement, the TTHP method utilizes three types of 2-hop path instead of the directed 2-hop path used in CFEL. Additionally, the EMDD model uses semantic vectors instead of the directed 2-hop path to measure the relatedness of entities. Finally, the CFELS achieves collective entity linking by introducing proposed strong relatedness sequence to CFEL.

Table 2
Characteristic analysis of the proposed SRSCL and baseline models.

Model/ Method	Information extraction methods for HINs	Overall relatedness capturing
POP	Adjacency matrix	No
CFEL	Directed 2-hop path	Cumulative entity pair relatedness
EMDD	Knowledge representation learning	No
TTHP	Three types of 2-hop path	No
CFELS	Directed 2-hop path	Strong relatedness sequence
SRSCL	Directed 2-hop path and knowledge representation learning	Strong relatedness sequence

To verify the effectiveness of the SRSCL proposed in this paper, the most classical POP and the latest collective entity linking method CFEL are used as comparison methods. Furthermore, to verify the effectiveness of the proposed strong relatedness sequence, CFELS is adopted as a comparison method. Additionally, to verify the efficacy of the three 2-hop path defined in this paper, TTHP is adopted as a comparison method. Lastly, to validate that the knowledge representation learning introduced in this paper can effectively extract the semantic information of entities and relationships in the HIN, EMDD is utilized as a comparison method. In addition, the above three datasets are taken as validation datasets and the HIN built in this paper is treated as the entity linking database. The precision, recall and F1 score of entity linking are taken as indicators to evaluate the effect of entity linking. The entity linking results of the baselines and the SRSCL on the aforementioned datasets are reported in Table 3.

Table 3 illustrates that compared with the five baselines, the SRSCL achieves the best results on AIDA CoNLL-YAGO, ACE2004 and AQUAINT (i.e., highest precision, highest recall and highest F1 score). Specifically, compared with POP, CFEL, EMDD, TTHP and CFELS, the SRSCL improves the precision of AIDA CoNLL-YAGO by 0.3%-7.24%, with an average increase of 2.77%. Compared with POP, CFEL, EMDD, TTHP and CFELS, the SRSCL improves the recall of AIDA CoNLL-YAGO by 0.29%-6.9%, with an average increase of 2.64%. And compared with POP, CFEL, EMDD, TTHP and CFELS, SRSCL improves the F1 score of AIDA CoNLL-YAGO by 0.3%-7.07%, with an average increase of 2.71%. Significantly, CFEL, EMDD, TTHP, CFELS and SRSCL all employ one strategy or two strategies (i.e. 2-hop path or knowledge representation learning) to measure the semantic similarity or relatedness between entity pair. In contrast, POP model captures the features around entities through the adjacency matrix of entities, which is not able to utilize the information of relations between entities. Unsurprisingly, POP achieves

Table 3
The entity linking results of the baselines and the SRSCL.

Model/Method		POP	CFEL	EMDD	TTHP	CFELS	**SRSCL**
AIDA CoNLL-YAGO	P(%)	74.99	80.21	79.96	81.93	80.21	**82.23**
	R(%)	71.39	76.36	76.12	78.00	76.36	78.29
	F1(%)	73.15	78.24	77.99	79.92	78.24	**80.22**
ACE2004	P(%)	83.87	89.52	89.52	90.32	89.52	**90.32**
	R(%)	83.87	89.52	89.52	90.32	89.52	90.32
	F1(%)	83.87	89.52	89.52	90.32	89.52	**90.32**
AQUAINT	P(%)	75.71	86.44	88.70	86.75	88.14	**89.83**
	R(%)	74.44	85.00	87.22	85.06	86.67	88.33
	F1(%)	75.07	85.71	87.96	85.90	87.40	**89.08**

the lowest entity link precision, recall and F1 score. The experimental results verify that the strategies of 2-hop path and knowledge representation learning can effectively extract the features of entities and relations. CFEL, EMDD and TTHP only use one of the two strategies to capture the features of entities, while the SRSCL adopts both strategies to capture the features of entities, making feature extraction more comprehensive. Therefore, the SRSCL achieves the best entity linking effect, which is consistent with the experimental results. Additionally, CFELS utilizes one type of 2-hop path to capture features around entities. The SRSCL, however, takes three types of 2-hop path and knowledge representation learning to capture features around entities. The experimental results verify that the three types of 2-hop path and knowledge representation learning model proposed in this paper can effectively extract the features of entities.

On AQUAINT, POP model achieves the lowest entity link precision, recall and F1 score, while CFEL achieves the second-worst results. The results of the proposed SRSCL, EMDD, TTHP and CFELS are significantly better than those of CFEL. Among these models, the SRSCL achieves the best results. The SRSCL, EMDD, TTHP, and CFELS employ either three types of 2-hop path to extract the features of entities, or a knowledge representation learning model to extract the features of entities. Experimentally, compared to the directed 2-hop path, these methods are able to more effectively and comprehensively extract the information of the HIN. Consequently, compared with the POP and CFEL, the entity linking effect of the SRSCL is significantly improved. For example, the F1 score of the SRSCL is 14.01% and 3.37% higher than that of the POP and CFEL, respectively. The experimental results verify the validity of the proposed SRSCL.

In addition, the ACE2004 benchmark is also used to evaluate the performance of the proposed entity linking method. Specifically, the POP achieves the lowest entity link precision, recall and F1 score. The CFEL, EMDD and CFELS achieves the second-best entity link precision, recall and F1 score, while the SRSCL and TTHP achieve the best scores. The results demonstrate that, due to the sparse data in ACE2004, multiple models yield similar entity linking effects. For instance, the precision, recall and F1 scores of the SRSCL and TTHP are identical. Notably, the proposed SRSCL obtained the best entity linking results of all the methods compared in this paper. This indicates the efficacy of the SRSCL in utilizing the information in HIN, and validates its effectiveness in entity linking.

The experimental results presented above show that the SRSCL proposed in this paper has attained the best entity linking effect on the three datasets (i.e., on ACE2004 and AQUAINT with sparse data, as well as on AIDA CoNLL-YAGO with dense data). Therefore, the SRSCL proposed in this paper is demonstrated to be highly effective for entity linking tasks.

The CFEL, EMDD, TTHP, CFELS, and SRSCL are members of the same family of entity linking methods. To verify the validity of the three types of 2-hop path, knowledge representation learning and strong relatedness sequence proposed in this paper, a comparison of the EMDD, TTHP, CFELS and SRSCL with CFEL is conducted. The precision and F1 score of the entity linking methods on the three datasets are illustrated in Fig. 12 and Fig. 13 respectively.

TTHP outperforms CFEL in precision and F1 score on the three datasets, especially on AIDA CoNLL-YAGO (as shown in Fig. 12 and Fig. 13). This is attributed to the fact that TTHP considers three types of 2-hop path to measure the relatedness between entity pair, whereas CFEL only uses one type of 2-hop path (i.e., directed 2-hop path). The experimental results verify that the three types of 2-hop path can more comprehensively extract the features of entities, which significantly improves the effect of entity linking. The precision and F1 score of EMDD on AIDA CoNLL-YAGO are basically the same as those of CFEL. This may be due to the limitation of the current knowledge representation learning, and the ability of the knowledge representation learning model to extract the deep features of entities needs to be improved. Interestingly, EMDD significantly outperforms CFEL in precision and F1 score on AQUAINT. This suggests that the knowledge representation learning can improve the entity linking effect of sparse data compared with the directed 2-hop path. It is also observed that CFELS achieves comparable performance with CFEL on the three datasets, with a significant advantage on AQUAINT, where the precision and F1 score of CFELS are significantly better than CFEL. This verifies that the strong relatedness sequence can effectively mine the relatedness among mutiple entities, thereby significantly improving the effect of entity linking.

The SRSCL integrates the three types of 2-hop path, knowledge representation learning and strong relatedness sequence. Theoretically, the SRSCL is expected to achieve the best entity linking effect. This expectation is further confirmed by the empirical results shown in Fig. 12 and Fig. 13. In these figures, it is evident that the SRSCL outperforms the latest CFEL in terms of precision and F1 score on the three datasets. The results thus demonstrate the validity of the proposed SRSCL for entity linking.

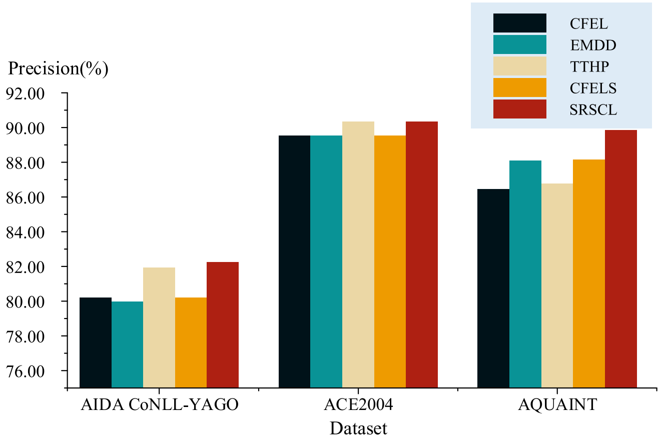


Fig. 12. The comparison results of entity linking precision.

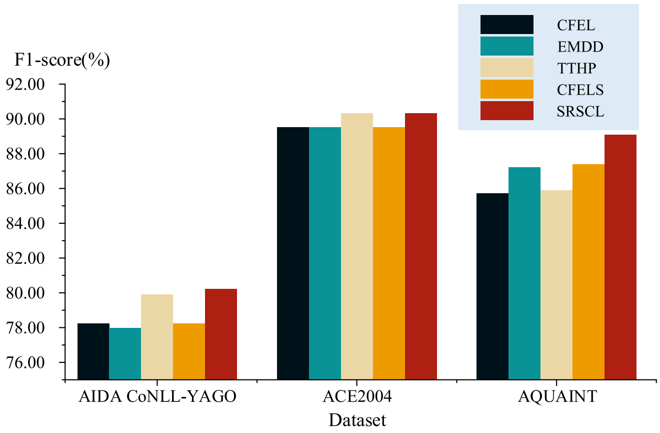


Fig. 13. The comparison results of entity linking F1 score.

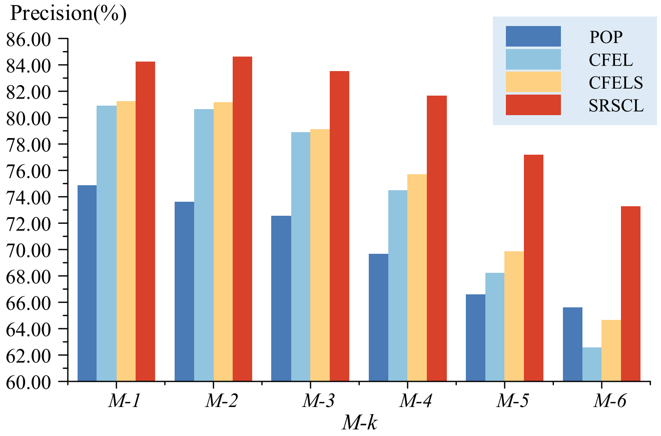


Fig. 14. Entity linking precisions of the POP, CFEL, CFELS and SRSCL on $M-k$.

On the one hand, the main advantage of the SRSCL is to mine the overall relatedness of candidate entity group. On the other hand, the overall relatedness of candidate entity group of entity mentioned can be calculated when the number of entity mentions in the sliding window is greater than 1. To evaluate the performance of the SRSCL in capturing the overall relatedness of candidate entity group, the $M-k$ method is employed, which is illustrated in the following description.

(2) The comparison experiment in capturing the overall relatedness of candidate entity group

To measure the performance of the entity linking methods in capturing the overall relatedness of candidate entity group, the $M-k$ method is adopted to estimate the experimental results on AIDA CoNLL-YAGO. Specifically, $M-1$, $M-2$, $M-3$, $M-4$, $M-5$ and $M-6$ are adopted to estimate the performance of the entity linking methods, where $M-k$ denotes that the number of entity mentions that have candidate entities in a sliding window, being greater than k. Finally, the entity linking results on $M-k$ are evaluated through precision, recall and F1 score. The experimental results are presented in Table 4.

Table 4 illustrates that the recall and F1 score of the four entity linking methods are positively correlated with the precision, i.e., the higher the precision is, the higher the recall is, and the higher the F1 score is. This suggests that the precision can be used as a representative indicator to evaluate the effect of entity linking in the experiments. To intuitively compare the entity linking performance of POP, CFEL, CFELS and SRSCL on $M-k$, the entity linking precisions of the four methods are presented in Fig. 14.

The SRSCL significantly outperforms the POP, CFEL and CFELS on $M-1$-$M-6$, as evidenced by Table 4 and Fig. 14. CFELS achieves the second-highest entity linking precision on $M-1$-$M-6$. CFEL and POP achieve the third and fourth best results on $M-1$-$M-6$, respectively. Compared with POP, CFEL demonstrates lower precision on $M-6$, indicating that it is difficult for CFEL to capture the overall relatedness of the candidate entity group when the number of entity mentions is large. In contrast, POP, which uses adjacency matrix, is able to effectively capture the overall relatedness of the candidate entity group. It is worth noting that the SRSCL and CFELS are outperforms POP and CFEL significantly on $M-1$-$M-6$. The experimental results demonstrate that the SRSCL and CFELS proposed in this paper are capable of capturing the overall relatedness of candidate entity group regardless of the number of entity mentions. The results confirm the effectiveness of the strong relatedness sequence, three types of 2-hop path and knowledge representation learning proposed or introduced in this paper for capturing the overall relatedness of candidate entity group.

Table 4
The performance comparison results in capturing the overall relatedness of candidate entity group.

$M-k$		$M-1$	$M-2$	$M-3$	$M-4$	$M-5$	$M-6$
POP	P(%)	74.85	73.60	72.53	69.66	66.58	65.58
	R(%)	73.80	72.84	72.06	69.21	66.22	65.28
	F1(%)	74.32	73.22	72.29	69.44	66.40	65.43
CFEL	P(%)	80.90	80.64	78.90	74.48	68.21	62.56
	R(%)	79.76	79.81	78.39	74.00	67.84	62.27
	F1(%)	80.32	80.23	78.64	74.24	68.02	62.41
CFELS	P(%)	81.25	81.16	79.12	75.70	69.84	64.65
	R(%)	80.11	80.32	78.60	75.21	69.46	64.35
	F1(%)	80.68	80.74	78.86	75.46	69.65	64.50
SRSCL	P(%)	**84.25**	**84.62**	**83.53**	**81.67**	**77.17**	**73.26**
	R(%)	83.06	83.75	82.99	81.14	76.76	72.92
	F1(%)	**83.65**	**84.19**	**83.26**	**81.40**	**76.96**	**73.09**

Compared with CFEL, the precisions of CFELS on $M-1$-$M-6$ are improved by 0.35%, 0.52%, 0.22%, 1.22%, 1.63% and 2.09%, respectively. The experimental results show that the entity linking precision by CFELS increases with the number of entity mentions in the sliding window, indicating that the strong relatedness sequence can effectively capture the overall relatedness of candidate entity group, especially when there are many entity mentions in the sliding window. In comparison, the precisions of the SRSCL on $M-1$-$M-6$ are improved by 3.00%, 3.46%, 4.41%, 5.97%, 7.33% and 8.61%, illustrating that the three types of 2-hop path, knowledge representation learning can effectively improve the performance of entity linking method in capturing the overall relatedness of candidate entity group. Notably, the precisions of the SRSCL on $M-1$-$M-6$ are improved by 3.35%, 3.98%, 4.63%, 7.19%, 8.96% and 10.70% compared with CFEL. The experimental results show that the precision of the SRSCL is significantly improved with the increase of the number of entity mentions in the sliding window, as illustrated in Fig. 14.

To further verify the performance of the SRSCL in capturing the overall relatedness of candidate entity group, comparison experiments are conducted on other two datasets, AQUAINT and ACE2004. Due to sparse data in these datasets, the number of entity mentions in the sliding window is small. Consequently, the comparison experiments are only made on $M-1$ and $M-2$, the results of which are shown in Table 5.

The results of the experiments conducted on the AQUAINT and ACE2004 datasets show that the strong relatedness sequence and the SRSCL can effectively capture the overall relatedness of candidate entity group, thus performing better than the CFEL, CFELS, and POP in terms of precision, recall and F1 score. This is evidenced in the results presented in Table 5. Specifically, on $M-1$ of AQUAINT, the SRSCL significantly outperforms CFELS, CFEL and POP in precision, recall and F1 score, and the CFELS outperforms CFEL in precision, recall and F1 score. The results indicate that the strong relatedness sequence and SRSCL can

Table 5
The comparison results in capturing the overall relatedness of candidate entity group.

Dataset		AQUAINT		ACE2004	
		$M-1$	$M-2$	$M-1$	$M-2$
POP	P	74.47%	44.44%	82.93%	**100.00%**
	R	74.47%	44.44%	82.93%	100.00%
	F1	74.47%	44.44%	82.93%	**100.00%**
CFEL	P	78.72%	66.67%	87.80%	86.67%
	R	78.72%	66.67%	87.80%	86.67%
	F1	78.72%	66.67%	87.80%	86.67%
CFELS	P	85.11%	66.67%	90.24%	86.67%
	R	85.11%	66.67%	90.24%	86.67%
	F1	85.11%	66.67%	90.24%	86.67%
SRSCL	P	**91.49%**	**66.67%**	**90.24%**	86.67%
	R	91.49%	66.67%	90.24%	86.67%
	F1	**91.49%**	**66.67%**	**90.24%**	86.67%

effectively capture the overall relatedness of candidate entity group. On $M-2$, POP achieves the worst results, while the performance of other methods is comparable. This could be attributed to the sparse data of AQUAINT, which might lead to similar entity linking behaviour of multiple entity linking methods. Similarly, on $M-1$ of ACE2004, the SRSCL and CFELS significantly outperform CFEL and POP. However, on $M-2$, POP achieves the best results, while the other methods achieve the second best results, which may be due to the sparsity of data and the fewer data samples available in ACE2004. The experimental results on AQUAINT and ACE2004 demonstrate that the strong relatedness sequence and SRSCL can effectively capture the overall relatedness of candidate entity group, leading to greater effectiveness of the SRSCL in entity linking. These experimental conclusions on AQUAINT and ACE2004 are consistent with those obtained from AIDA CoNLL-YAGO.

To conclude, the proposed SRSCL is more effective than the classical POP model and the latest CFEL method in capturing the overall relatedness of candidate entity group. Its effectiveness is especially notable when the number of entity mentions is large in one sliding window, significantly improving the success rate of collective entity linking. Therefore, the SRSCL can effectively realize the task of collective entity linking.

4.4.2. Discussion of the influence of different hyperparameters on the performance of the SRSCL

This section focuses on discussing the effects of the hyperparameters of SRSCL on collective entity linking performance. The SRSCL method introduces the semantic similarity of entities. Therefore, semantic vectors of entities in HIN are first obtained by the RotatE model. Subsequently with the semantic vectors fixed, the effects of the hyperparameters of the SRSCL model and the three SRSCL computation methods proposed in Section 3.2.5 assessed in terms of collective entity linking performance..

In the proposed SRSCL, the RotatE model is first adpoted to learn the embedding vector for each entity to obtain the semantic representation. The YAGO-Core fact triples are utilized to train the RotatE, where the size of the training set is 770 k, the size of the validation set is 8000 and the size of the test set is 8000. In experiment, when batch_size = 4800 (i.e., the hyperparameter that controls the size of the training batch), embedding_dimension = 512 (i.e., the entity and relation embedding dimension), fixed_margin = 6.0 (i.e., the fixed distance threshold used in the Rotate model that determines whether entities have a relation or not), train_times = 3000 (i.e., the hyperparameter that controls the training epoch) and self-adversarial_sampling_temperature = 2e-4 (i.e., the probability hyperparameter for extracting negative triplets), the RotatE obtains a good performance with the hits@10 (filter) of 0.8869. Hence, the entity embedding vector learned from the RotatE, with the above hyperparameters, can effectively capture the semantic information of the entity. Then, the influence of different hyperparameters on the performance of the SRSCL is discussed using the AQUAINT, as described below.

According to Eq. (11), Eq. (12) and Eq. (13), three different representations of SRSCL are presented, namely SRSCL-11, SRSCL-12, and SRSCL-13. The objective of this section is to obtain the optimal SRSCL. The hyperparameters that have a significant impact on the performance of SRSCL include β, γ, the size of candidate entity set (i.e., top-k), the length of the sliding window (denoted as l). By default, the SRSCL-12 is selected as the optimal SRSCL, and the hyperparameters is set to $\beta = 0.5$, $\gamma = 0.1$, top-k = top-10 and $l = 15$.

(1) The influence of β and γ on the performance of the SRSCL

First, the size of candidate entity set and the length of the sliding window are set to the default values. Based on experience and knowledge, the optimization range of β and γ are established as follows: $\beta \in [0.3, 0.4, 0.5, 0.6, 0.7]$ and $\gamma \in [0.1, 0.2, 0.3, 0.4, 0.5]$. Subsequently, the set hyperparameters are used to optimize the SRSCL-11, SRSCL-12 and SRSCL-13 respectively to select the optimal calculation strategy.

The experimental results of the SRSCL-11, SRSCL-12 and SRSCL-13 with different hyperparameters are shown in Table 6. Since β and γ are not included in the SRSCL-13, the results of the SRSCL-13 remain unchanged under different hyperparameters, as shown in Table 5. When $\beta = 0.3$ and $\gamma = 0.4$, the SRSCL-11 achieves better precision, recall and F1 score. When $\beta = 0.3$ and $\gamma = 0.4$, the SRSCL-12 achieves better precision, recall and F1 score. Among multiple hyperparameters, the SRSCL-11 and SRSCL-12 obtain better results when $\beta = 0.3$. The results illustrate that the other two types of 2-hop path proposed in this paper have an importance of 0.7, which verifies the effectiveness of the three types of 2-hop path proposed in this paper.

Table 6 demonstrates that the SRSCL-12 outperforms the SRSCL-11 and SRSCL-13 in terms of precision, recall and F1 score. Specifically, the precision, recall and F1 score of the SRSCL-12 are 2.82%, 2.77%, and 2.81% higher than those of the SRSCL-11, respectively. The difference between the SRSCL-12 and SRSCL-11 is that the former calculates semantic similarity using semantic vectors, whereas the latter utilizes three types of 2-hop path to estimate semantic similarity. Comparing the SRSCL-12 with SRSCL-11, a conclusion can be drawn that knowledge representation learning is more effective in capturing semantic information of entities in comparison with three types of 2-hop path. In addition, the results of the experiments demonstrate that the SRSCL-12 outperforms the SRSCL-13 in terms of precision, recall, and F1 score. This is due to the difference in the approaches taken by the two algorithms: while the SRSCL-12 uses three types of 2-hop path to calculate the relatedness between the entity pair, the SRSCL-13 employs the knowledge representation learning to estimate the relatedness. The findings suggest that the relatedness between the entity pair can be effectively measured through the three types of 2-hop path, rather than through knowledge representation learning. In fact, semantic similarity may pay more attention to the meaning of entities in all contexts, while the relatedness may be more focus on the local characteristics between entity pair. Therefore, the three types of 2-hop path is more effective in measuring the relatedness between entity pair, and the knowledge representation learning is more effective in measuring the semantic similarity between entities. From above analysis, the SRSCL-12 not only captures the semantic information of entities through the knowledge representation learning effectively, but also captures the relatedness of entity pair through the three types of 2-hop path. As such, the SRSCL-12 can be considered the most effective entity linking method among the three SRSCL methods.

Table 6 illustrates that the SRSCL-12 outperforms SRSCL-11 and SRSCL-13 in the performance of entity linking. Therefore, the SRSCL-12 is selected as the optimal SRSCL, and β is set to 0.3 and γ is set to 0.4. Then the size of candidate entity set and the length of the sliding window

Table 6
The results of the SRSCL with different hyperparameters.

Number	β	γ	SRSCL-11			SRSCL-12			SRSCL-13		
			P(%)	*R*(%)	*F*1(%)	*P*(%)	*R*(%)	*F*1(%)	*P*(%)	*R*(%)	*F*1(%)
1	0.3	0.1	86.44	85.00	85.71	89.27	87.78	88.52	89.27	87.78	88.52
2	0.3	0.2	87.01	85.56	86.27	89.27	87.78	88.52	89.27	87.78	88.52
3	0.3	0.3	87.01	85.56	86.27	89.83	88.33	89.08	89.27	87.78	88.52
4	**0.3**	**0.4**	**87.01**	**85.56**	**86.27**	**89.83**	**88.33**	**89.08**	89.27	87.78	88.52
5	0.3	0.5	87.01	85.56	86.27	88.70	87.22	87.96	89.27	87.78	88.52
6	0.4	0.1	84.75	83.33	84.03	89.27	87.78	88.52	89.27	87.78	88.52
7	0.4	0.2	87.01	85.56	86.27	89.27	87.78	88.52	89.27	87.78	88.52
8	0.4	0.3	87.01	85.56	86.27	89.27	87.78	88.52	89.27	87.78	88.52
9	0.4	0.4	86.44	85.00	85.71	88.70	87.22	87.96	89.27	87.78	88.52
10	0.4	0.5	87.01	85.56	86.27	88.70	87.22	87.96	89.27	87.78	88.52
11	0.5	0.1	84.75	83.33	84.03	89.27	87.78	88.52	89.27	87.78	88.52
12	0.5	0.2	87.01	85.56	86.27	89.27	87.78	88.52	89.27	87.78	88.52
13	0.5	0.3	87.01	85.56	86.27	89.27	87.78	88.52	89.27	87.78	88.52
14	0.5	0.4	86.44	85.00	85.71	88.70	87.22	87.96	89.27	87.78	88.52
15	0.5	0.5	86.44	85.00	85.71	88.70	87.22	87.96	89.27	87.78	88.52
16	0.6	0.1	85.31	83.89	84.59	89.27	87.78	88.52	89.27	87.78	88.52
17	0.6	0.2	85.31	83.89	84.59	89.27	87.78	88.52	89.27	87.78	88.52
18	0.6	0.3	86.44	85.00	85.71	88.14	86.67	87.39	89.27	87.78	88.52
19	0.6	0.4	86.44	85.00	85.71	88.70	87.22	87.96	89.27	87.78	88.52
20	0.6	0.5	85.88	84.44	85.15	88.70	87.22	87.96	89.27	87.78	88.52
21	0.7	0.1	85.31	83.89	84.59	89.27	87.78	88.52	89.27	87.78	88.52
22	0.7	0.2	84.75	83.33	84.03	88.14	86.67	87.39	89.27	87.78	88.52
23	0.7	0.3	86.44	85.00	85.71	88.14	86.67	87.39	89.27	87.78	88.52
24	0.7	0.4	85.88	84.44	85.15	88.14	86.67	87.39	89.27	87.78	88.52
25	0.7	0.5	85.88	84.44	85.15	88.70	87.22	87.96	89.27	87.78	88.52

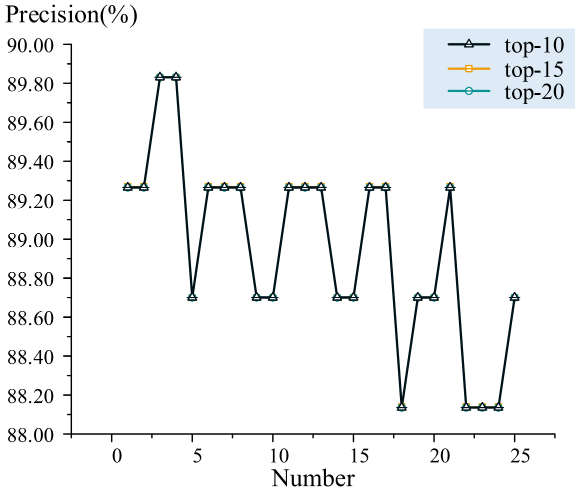


Fig. 15. The entity linking precisions of the SRSCL-12 on different sizes of candidate entity sets. (The precisions of top-10, top-15 and top-20 are same).

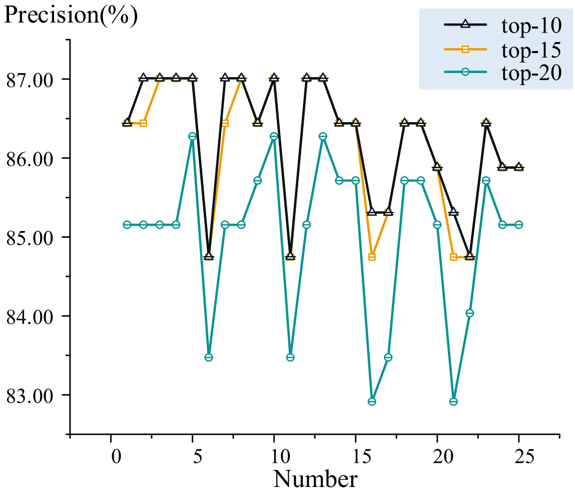


Fig. 16. The entity linking precisions of the SRSCL-11 on different sizes of candidate entity sets.

are optimized as follows.

(2) The influence of the size of candidate entity set on the performance of the SRSCL

To investigate the effect of the size of candidate entity set on the SRSCL, three sizes of candidate entity set are evaluated: top-10, top-15 and top-20. The results, depicted in Fig. 15, show that the precision of the SRSCL-12 remained unchanged for the three sizes of the candidate entity set. Additionally, note that the larger the size of candidate entity set is, the higher the complexity of the SRSCL-12 is. Consequently, the size of candidate entity set is set to top-10.

To further verify the effectiveness of the candidate entity generation method and the rationality of setting the size of candidate entity set as top-10, experiments are repeated on the SRSCL-11 and the results are shown in Fig. 16. As can be observed, the entity linking precision of the SRSCL-11 gradually decreases with the size of the candidate entity set ranging from top-10 to top-20. This indicates that, when the size of candidate entity set is increased, the noise of the candidate entities is increased in tandem. Moreover, the presence of candidate entities unrelated to the entity mention interferes with the correct entity linking, thus resulting in a decrease in the entity linking precision of the SRSCL-11. These results demonstrate that a small number of candidate entities can achieve a higher entity linking precision, thereby verifying the effectiveness of the proposed candidate entity generation method and justifying the choice to set the size of candidate entity set to top-10 in this paper.

(3) The influence of the length of the sliding window on the performance of the SRSCL

To investigate the influence of the length of the sliding window on the performance of the SRSCL, *l* is set to 10, 15, 20 and other hyperparameter settings are fixed. The precision, recall and F1 score of the SRSCL-12 under different sliding window lengths are illustrated in Fig. 17.

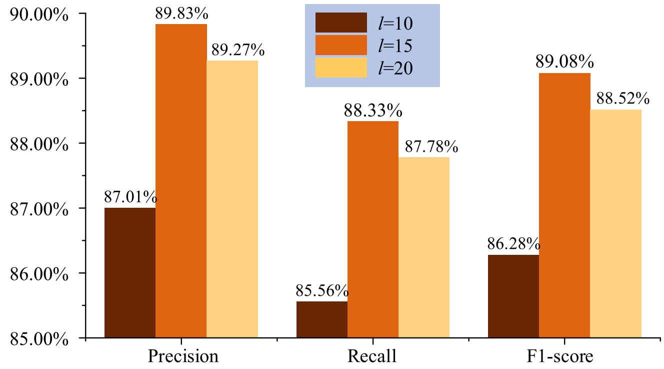


Fig. 17. Entity linking results of the SRSCL-12 under different sliding window lengths.

Generally, the number of entity mentions in the sliding window increases as the length of sliding window increases. Fig. 17 demonstrates that the precision, recall and F1 score of the SRSCL-12 rise steadily when the sliding window is set from 10 to 15. This indicates that the SRSCL-12 can effectively capture the overall relatedness of candidate entity group, leading to an improvement in the entity linking effect with an increase in the length of the sliding window. However, the precision, recall and F1 score of the SRSCL-12 decrease gradually when the sliding window is set from 15 to 20. This may be attributed to the fact that when describing a thing, the keywords that are closer to the central words tend to be more relevant to the thing, whereas those that are farther away tend to be less relevant. Therefore, when the number of entity mentions in the sliding window is continuously increased, it will introduce more unrelated entity mentions and, consequently, more noise to the entity linking process, leading to a degraded performance of entity linking. Based on the aforementioned analysis, the SRSCL can obtain an optimal entity linking effect when the length of the sliding window is set to 15.

In conclusion, the SRSCL-12 demonstrates the best entity linking performance when the hyperparameter β is set to 0.3, γ is set to 0.4, the size of candidate entities set is set to top-10 and the length of the sliding window is set to 15.

5. Conclusions

The collective entity linking for heterogeneous information networks has been a long-standing challenge in the application of such networks. An in-depth analysis of the latest collective entity linking method, CFEL, reveals three key unsolved issues. To address these issues, this paper proposes a strong-relatedness-sequence-based fine-grained collective entity linking method (SRSCL) for heterogeneous information networks. First, in view of the fact that the CFEL does not fully follow the "collective linking" idea in solving the objective function, SRSCL introduces a knowledge representation learning model and proposes an overall semantic similarity model for entity mention group and candidate entity group to closely adhere to the "collective linking" idea. Second, to address the difficulty of CFEL in highlighting the importance of strong logical associations in measuring the overall relatedness of candidate entity group, SRSCL proposes the concept of strong relatedness sequence, and a strong-relatedness-sequence-based overall relatedness measurement model for candidate entity group. Third, aiming at the problem that CFEL only uses one types of 2-hop path to measure entity relatedness, SRSCL defines three types of 2-hop path and considers the contribution of each path to entity relatedness. Accordingly, an entity pair relatedness measurement model based on three 2-hop path is investigated to accurately measure entity relatedness. In addition, this paper establishes a $M-k$ method to evaluate the performance of collective entity linking in capturing the overall relatedness of candidate entity group. Finally, the effectiveness of the SRSCL is validated by a series of experiments. Experimental results show that the proposed SRSCL in this paper improves the precision, recall and F1 score by 10.7% in comparison with the latest model when the number of entity mentions contained in the sliding window is greater than 6.

In this paper, the proposed SRSCL relies on categories of entity mention and entity, thus making it difficult to effectively implement entity linking without category information. Consequently, future research should be devoted to methods for predicting categories of entity mention and entity. Furthermore, the time complexity of the algorithm should also be addressed. To this end, we plan to focus on improving the efficiency of entity linking, with the ultimate goal of applying our proposed entity linking method to larger heterogeneous information networks.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

This work was funded by National Natural Science Foundation of China [No. U2133202], the Fellowship of China Postdoctoral Science Foundation [No. 2022M720955] and the Fellowship of Heilongjiang Province Postdoctoral Science Foundation [No. LBH-Z22187].

References

Akabe, K., Takeuchi, T., Aoki, T., & Nishimura, K. (2021). Information retrieval on oncology knowledge base using recursive paraphrase lattice. *Journal of Biomedical Informatics, 116*, Article 103705.

Biggs, N. (2002). *Discrete mathematics*. Oxford University Press.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems, 26*.

Chong, W.-H., Lim, E.-P., & Cohen, W. (2017). Collective entity linking in tweets over space and time. Advances in Information Retrieval: 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings 39.

Daiber, J., Jakob, M., Hokamp, C., & Mendes, P. N. (2013). Improving efficiency and accuracy in multilingual entity extraction. Proceedings of the 9th international conference on semantic systems.

Ganea, O.-E., Ganea, M., Lucchi, A., Eickhoff, C., & Hofmann, T. (2016). Probabilistic bag-of-hyperlinks model for entity linking. Proceedings of the 25th International Conference on World Wide Web.

Geng, Z., Chen, G., Han, Y., Lu, G., & Li, F. (2020). Semantic relation extraction using sequential and tree-structured LSTM with attention. *Information Sciences, 509*, 183–192.

Hoffart, J., Seufert, S., Nguyen, D. B., Theobald, M., & Weikum, G. (2012). KORE: keyphrase overlap relatedness for entity disambiguation. Proceedings of the 21st ACM international conference on Information and knowledge management.

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., . . . Weikum, G. (2011). Robust disambiguation of named entities in text. Proceedings of the 2011 conference on empirical methods in natural language processing.

Huang, Z., Zheng, Y., Cheng, R., Sun, Y., Mamoulis, N., & Li, X. (2016). Meta structure: Computing relevance in large heterogeneous information networks. Proceedings of the 22nd ACM SIGKDD International conference on knowledge discovery and data mining.

Kalender, M., Korkmaz, E. E., & Engineering, D. (2017). THINKER-entity linking system for Turkish language. *IEEE Transactions on Knowledge and Data Engineering, 30*(2), 367–380.

Li, C., & Tian, Y. (2020). Downstream model design of pre-trained language model for relation extraction task. *arXiv preprint arXiv:.03786*.

Li, J., Bu, C., Li, P., & Wu, X. (2021). A coarse-to-fine collective entity linking method for heterogeneous information networks. *Knowledge-Based Systems, 228*, Article 107286.

Li, Z., Zhao, Y., Li, Y., Rahman, S., Wang, F., Xin, X., & Zhang, J. (2021). Fault localization based on knowledge graph in software-defined optical networks. *Journal of Lightwave Technology, 39*(13), 4236–4246.

Liu, M., Gong, G., Qin, B., & Liu, T. (2019). A multi-view-based collective entity linking method. *ACM Transactions on Information Systems, 37*(2), 1–29.

Nasar, Z., Jaffry, S. W., & Malik, M. K. (2021). Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys, 54*(1), 1–39.

Oliveira, I. L., Fileto, R., Speck, R., Garcia, L. P., Moussallem, D., & Lehmann, J. (2021). Towards holistic entity linking: Survey and directions. *Information Systems, 95*, Article 101624.

Onoe, Y., & Durrett, G. (2020). Fine-grained entity typing for domain independent entity linking. Proceedings of the AAAI Conference on Artificial Intelligence.

Ravi, M. P. K., Singh, K., Mulang, I. O., Shekarpour, S., Hoffart, J., & Lehmann, J. (2021). CHOLAN: A modular approach for neural entity linking on Wikipedia and Wikidata. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume.

Sevgili, Ö., Shelmanov, A., Arkhipov, M., Panchenko, A., & Biemann, C. (2022). Neural entity linking: A survey of models based on deep learning. *Semantic Web, 13*(3), 527–570.

Shen, W., Han, J., & Wang, J. (2014). A probabilistic model for linking named entities in web text with heterogeneous information networks. Proceedings of the 2014 ACM SIGMOD international conference on Management of data.

Shen, W., Han, J., Wang, J., Yuan, X., Yang, Z., & Engineering, D. (2017). SHINE+: A general framework for domain-specific entity linking with heterogeneous information networks. *IEEE Transactions on Knowledge and Data Engineering, 30*(2), 353–366.

Shen, W., Wang, J., Han, J., & Engineering, D. (2014). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering, 27*(2), 443–460.

Song, B., Li, F., Liu, Y., & Zeng, X. (2021). Deep learning methods for biomedical named entity recognition: A survey and qualitative comparison. *Briefings in Bioinformatics, 22*(6), bbab282.

Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., & Wu, T. (2009). Rankclus: integrating clustering with ranking for heterogeneous information network analysis. Proceedings of the 12th international conference on extending database technology: advances in database technology.

Sun, Z., Deng, Z.-H., Nie, J.-Y., & Tang, J. (2019). RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space [preprint]. *Arxiv*. https://doi.org/arXiv:1902.10197.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., & Bouchard, G. (2016, Jun 20-22). Complex Embeddings for Simple Link Prediction.*Proceedings of Machine Learning Research* [International conference on machine learning, vol 48]. 33rd International Conference on Machine Learning, New York, NY.

Wang, H., Zheng, J. G., Ma, X., Fox, P., & Ji, H. (2015). Language and domain independent entity linking with quantified collective validation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.

Wang, Z., Zhang, J. W., Feng, J. L., Chen, Z., & Aaai. (2014, Jul 27-31). Knowledge graph embedding by translating on hyperplanes. *AAAI Conference on Artificial Intelligence* [Proceedings of the twenty-eighth aaai conference on artificial intelligence]. 28th AAAI Conference on Artificial Intelligence, Quebec City, Canada.

Wu, X., Tang, Y., Zhou, C., Zhu, G., Song, J., & Liu, G. (2022). An intelligent search engine based on knowledge graph for power equipment management. 2022 5th International Conference on Energy, Electrical and Power Engineering (CEEPE).

Xia, Y., Wang, X., Gu, L., Gao, Q., Jiao, J., & Wang, C. (2020). A collective entity linking algorithm with parallel computing on large-scale knowledge base. *The Journal of Supercomputing, 76*(2), 948–963.

Xie, L., Hu, Z., Cai, X., Zhang, W., Chen, J., & Systems, I. (2021). Explainable recommendation based on knowledge graph and multi-objective optimization. *Complex, 7*, 1241–1252.

Ye, Q., Hsieh, C.-Y., Yang, Z., Kang, Y., Chen, J., Cao, D., … Hou, T. (2021). A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. *Nature Communications, 12*(1), 6775.