Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Entity perception of Two-Step-Matching framework for public opinions

Ren-De Li, Hao-Tian Ma, Zi-Yi Wang, Qiang Guo, Jian-Guo Liu

A Library and Research Center of Computer Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, PR China
B School of Accountancy and Shanghai Key Laboratory of Financial Information Technology, Shanghai University of Finance and Economics, Shanghai 200433, PR China
C School of Humanities, Shanghai University of Finance and Economics, Shanghai 200433, PR China
D Institute of Sina WRD Big Data, Shanghai 201204, PR China

A R T I C L E   I N F O

Keywords:
Entity perception
BILSTM-CRF model
Jaro–Winkler distance algorithm
User comments
Public opinions

A B S T R A C T

Entity perception of ambiguous user comments is a critical problem of target identification for huge amount of public opinions. In this paper, a Two-Step-Matching method is proposed to identify the precise target entity from multiple entities mentioned. Firstly, potential entities are extracted by BILSTM-CRF model and characteristic words by TF-IDF model from public comments. Secondly, the first matching is implemented between potential entities and an official business directory by Jaro–Winkler distance algorithm. Then, in order to find the precise one, an industry-characteristic dictionary is developed into the second matching process. The precise entity is identified according to the count of characteristic words matching to industry-characteristic dictionary. In addition, associated rate (global indicator) and accuracy rate (sample indicator) are defined for evaluation of matching accuracy. The results for three data sets of public opinions about major public health events show that the highest associated rate and accuracy rate arrive at 0.93 and 0.95, averagely enhanced by 32% and 30% above the case of using the first matching process alone. This framework provides the method to find the true target entity of really wanted expression from public opinions.

1. Introduction

Public opinions usually contain critical information, pointing to the certain entity for social perception. For instance, the new Coronavirus 2019 (COVID-19) was initially linked to the Huanan Seafood Wholesale Market from public opinions [1]. Especially in emergency safety management, there is a delay from the occurrence of event to the perception of regulators/authorities. However, early signal can be found in social media in early time, such as the complaints/comments about the event. These signals may be ambiguous and hard to associate with real entities. Therefore, entity identification is one of the core techniques as natural language processing applications for information retrieval in the safety domain, which is initially proposed by Lim et al. [2]. They analyzed the use of extended keys, the union of keys in the relationships to be matched, and their corresponding identity rules to determine the equivalence between tuples in relationships that may not share any common keys. Hernández and Stolfo [3,4] developed a system to solve the merge/clear problem, which was to process the data multiple times with different keys in order to sort in each successive traversal. Arasu and co-workers [5,6] proposed a procedural framework for record matching, which took user-defined string con- versions as input. Fan et al. [7,8] suggested matching graphics patterns based on the concept of bounded simulation. Furthermore, matching methods take more account of semantics information. Cohen [9] extended the similarity measurement method based on TF-IDF similarity measurement in order to identify entities. Augsten et al. [10] and Schallehn et al. [11] defined the p-q-gram distance to exact matches of data items that represented the same real-world object. Rizzo and Troncy [12] proposed a framework, NERD, which unified entity extractors through natural language processing. Song et al. [13] produced PKDE4J (an extensible and flexible text mining system) to enable entity extraction and relationship extraction. Gradually, matching

* This work is partially supported by the National Natural Science Foundation of China (Grant Nos. 71901144, 71771152, 61773248), the Major Program of National Fund of Philosophy and Social Science of China (18ZDA088, 202DA060), Shanghai Planning Office of Philosophy and Social Science Foundation (Grant No. 2019EXW001), Foundation of University of Finance and Economics (Grant No. 2017110709), and S-Tech internet communication project (Grant Nos. 2018BHD005 and 2018TECH003).

* Corresponding author: School of Accountancy and Shanghai Key Laboratory of Financial Information Technology, Shanghai University of Finance and Economics, Shanghai 200433, PR China.
E-mail addresses: qiang.guo@usts.edu.cn (Q. Guo), liug004@ustc.edu.cn, liu.jianguo@sufe.edu.cn (J.-G. Liu).

https://doi.org/10.1016/j.jnlsr.2020.06.005
Received 30 March 2020; Received in revised form 4 June 2020; Accepted 5 June 2020
2666-4496/© 2020 China Science Publishing & Media Ltd. Publishing Services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license. (http://creativecommons.org/licenses/by-nc-nd/4.0/)
methods of the content-based record are integrated into the rule-based methods. Recently, the disambiguation of entity identification attracts more attentions. Pappu et al. [14] suggested a lightweight, and multilingual named entity recognition (NER) and linking (NEL) system to detect entities in new languages with limited labelled resources. Nguyen et al. [15] offered J-REED, a joint approach for entity disambiguation and relation extraction that was based on probabilistic graphical models. Nguyen et al. [15] illustrated TwitterNEED, a hybrid approach for named entity extraction and named entity disambiguation for tweets. Canale et al. [16] illustrated a novel ensemble method for named entity recognition and disambiguation based on neural network. These researches distinguish taxonomies and disambiguation with different knowledge bases. However, it is a big challenge for matching entity precisely when public opinions contain multiple targets and which entity is referred ambiguously. For example, there is a piece of online user comment: “This Honeywell N95 mask bought from Taobao is very easy to use. It is very comfortable to wear after half day wearing. The air valve is not blocked in any way. Please see the attached photos which was taken by my Huawei mobile phone. Must be given high praise!”. In this comment, three entity names are mentioned regarding to mask. Accordingly, ambiguity has appeared in machine recognition since the content cannot locate to the specific target. As shown in Fig. 1, it can effectively extract targets and match entities using existing technologies. The first matching process can realize the matching between user comments and real entities, but how to distinguish the precise one is the main concern of entity perception.

In this work, a Two-Step-Matching (TSM) method is introduced into the framework. At the first stage, potential entity will be extracted from user comments by using BiLSTM-CRF model and characteristic words in comments will be recorded. The first matching process will pair the potential entity with an official business directory by Jaro-Winkler distance algorithm. Meanwhile, an industry-characteristic dictionary is constructed by combination of characteristic words in comments and description words in business directory. The next stage is to identify the precise entity if there are multiple targets in the first matching process. An industry-characteristic dictionary is introduced into the count of characteristic words matching with the user comments. As illustrated in Fig. 1, if a comment contains 6 characteristic words about entity A, while the other two entities match less characteristic words. It is generally agreed that the comment mainly points to the entity A, which is the precisely perceiving target. At the end, associated rate (Global indicator) and accuracy rate (sample indicator) are defined to evaluate the effect of entity identification. The results show that TSM framework performs significantly than using the first matching process alone.

2. The Two-Step-Matching perception method

2.1. Prepare the text ingredients

To obtain the potential entity from users’ comments, it is necessary to prepare the text string factors at the beginning stage. The research [17] indicated that the character-based bidirectional LSTM-CRF (BiLSTM-CRF) model included both effectiveness and efficiency, which is suitable for sentence-level [18] or document-level text [19].

The model divides Chinese phrase into a sequence which includes n characters as \( x = \{ x_1, x_2, \ldots, x_n \} \), \( x_i \) indicates the \( i \)th character in the phrase or sentence \( x \). Each character is represented as a q-dimensional one hot vector. Each character \( x_i \) in sentence \( x \) has a tag from the BIO Encoding set \{B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG, O\}. These tags are the common entity types to represent as follows: person (PER), location (LOC), organization (ORG) and none of any type (O). B and I are short for the “Beginning” and “Inside Characters”, respectively.

The 1st layer of the BiLSTM-CRF model, Look-up map, maps each character representation from a one hot vector into a lower dimensional dense word vector that includes character embedding and word embedding. Initialize the character embedding randomly and employ the pre-trained word embedding.

The inputs of the 2nd layer, BiLSTM layer, are those embeddings like \( (x_1, x_2, x_3, x_4) \). And the outputs of BiLSTM layer are the scores of each label. In the BiLSTM layer, it processes the data in both directions with two separate hidden layers. The BiLSTM computes the forward hidden sequence \( (f_{1}, f_{2}, f_{3}, f_{4}) \), and the backward hidden sequence \( (b_{1}, b_{2}, b_{3}, b_{4}) \). The representation of a character by BiLSTM layer is acquired by concatenating its forward and backward context representation, \( x_i = [f_i, b_i] \), which contains a representation of a character in context effectively and it is helpful to tag applications in next layer. Then, extracted sentence features are recorded as a matrix \( P = \{ p_1, p_2, p_3, p_4 \} \). Each \( p_i \) is regarded as the score value of classifying the \( x_i \) to the label.

These scores of each label predicted by the BiLSTM layer are the inputs of the 3rd layer, the CRF layer. And the outputs of this CRF layer are predicted labels from label sequence with the highest prediction score for characters \( x_i \) in sentence \( x \) initially. Considering \( P \) is the matrix of the scores outputted by the BiLSTM layer and its size is \( n \times k \), where \( n \) is the number of characters of sentence \( x \), and \( k \) is the number of distinct tags. \( p_y \) corresponds to the score of \( j \)th label of the \( i \)th character. Then if regarding a sequence of predicted label \( y = (y_1, y_2, \ldots, y_n) \), where \( n \) is equal to the length of sentence \( x \), and the scores of the label are:

\[
\text{score}(x, y) = \sum_{i=1}^{n} p_{y_i} + \sum_{i=1}^{n-1} A_{y_i, y_{i+1}}
\]
where \( A \) is a matrix of transition scores and \( A_{ij} \) represents the transition score from the \( i \)th label to the \( j \)th label. With the aim of having a more robust transition score matrix, two more labels, START and END, are put into sentence \( x \), so the size of matrix \( A \) is \((k + 2) \times (k + 2)\). The score is made up of two parts: one is the LSTM’s output \( p_i \), and the other one is determined by the transition matrix \( A \).

A softmax yields a probability for predicted label sequence \( y \):

\[
p(y|x) = \frac{\exp(score(x, y))}{\sum_{\bar{y} \in Y} \exp(score(x, \bar{y}))},
\]

where \( Y \) is all possible label sequences for the sentence \( x \).

Then, the log-probability of the correct label sequence during training process is given as:

\[
\log p(y|x) = score(x, y) - \log(\sum_{\bar{y} \in Y} \exp(score(x, \bar{y}))).
\]

The last output predicted label sequence with maximum score is given by:

\[
y^* = \arg \max_{y \in Y} score(x, y).
\]

Train the BiLSTM-CRF model, then test the entities from user comment texts, recording them as entity names \( e \).

The other text ingredient from user comments are characteristic words. After Chinese text segmentation, i.e. jieba, TF-IDF algorithm [20] is used to extract the characteristic words.

Each words list of comment is seen as the text document, the term frequency is expressed as:

\[
tf_{i,j} = \frac{g_{i,j}}{\sum_k g_{k,j}}.
\]

where \( g_{i,j} \) represents the number of occurrences of the term \( g_i \) in document \( w_j \) and \( \sum_k g_{k,j} \) represents the total number of all the words in the document \( w_j \). The inverse document frequency is:

\[
idf_i = \log \frac{|W|}{|j: g_i \in w_j|},
\]

where \(|W|\) is total number of documents and \(|j: g_i \in w_j|\) is the number of documents where the term \( g_i \) appears. Then \( tf \cdot idf \) is calculated as:

\[
tf \cdot idf_{i,j} = tf_{i,j} \cdot idf_i.
\]

A high weight in \( tf \cdot idf \) is reached by a high term frequency and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms. Top 20 highest weight words are chosen as the feature words of each documents.

Consequently, characteristic keywords \( a \) can be obtained. In the same method, the description words \( b \) can also be obtained from the description of business scope of entity in the business directory by TF-IDF algorithm.

### 2.2. Matching the target entity

Usually, Chinese entities contain four basic elements for their full names, which are administrative divisions, short name, industry and organization format. This work just focuses on Short Name \( C \) and Industry \( I \) to build the business directory. In Table 1, the third and fourth columns are examples, and the business scope is demonstrated in the last column.

String matching techniques can find string \( A \) within string \( B \) regardless of the order of characters, then it can index and retrieve information from database [21]. In this work, string matching happens between entity names from user comments and short name from business directory. This first matching process can identify the entities pairs, then the second matching process occurs between characteristic keywords and industry-characteristic dictionary to pick out the target entity.

Jaro–Winkler distance [22] is an algorithm for calculating the similarity between two strings. It is first proposed to determine whether the two names on the health record are the same. It is suitable for calculation between the short characters such as strings. Given two strings, the Jaro similarity is calculated as:

\[
d_i = \frac{1}{3} \left( \frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m - t}{m} \right).
\]

where \( d_i \) is the Jaro similarity score. \( |s_1| \) and \( |s_2| \) are the length of two strings. \( m \) is the number of characters matched (the order is guaranteed to be the same), and \( t \) is the number of transpositions that is the distance between two characters \( |s_1| \) and \( |s_2| \), which is no more than \( \frac{Max(|s_1|, |s_2|)}{2} - 1 \).

The Jaro–Winkler algorithm gives the starting part a higher score for the same string, which is defined as a prefix range \( p \). If the prefix part has the same length as the partial string, the Jaro–Winkler distance is:

\[
d_{w} = d_i + L \cdot p(1 - d_i),
\]

where \( L \) is the length of the prefix part matched, and \( L = 2 \) is set in this work. \( p \) is a range factor constant to adjust the prefix matching. The Jaro–Winkler distance is from 0 to 1, which 0 represents no similarity between two strings and 1 represents an exact match.

In the first match, Jaro–Winkler distance algorithm is applied between entity names \( e \) and short name \( C \) in business directory.

However, after the first match based on Jaro–Winkler distance, multiple matching pairs appear. Therefore, a subsequent second matching process is presented with an industry-characteristic dictionary, which is constructed as Fig. 2.

As shown in Fig. 2(a), the first string match happens between keywords \( a = \{a_1, a_2, a_3, \ldots \} \) and the names \( e = \{e_1, e_2, e_3, \ldots \} \) from user comments, and the short names from the business directory \( C = \{C_1, C_2, C_3, \ldots \} \). These names are corresponding to the Industry \( I = \{I_1, I_2, I_3, \ldots \} \).

Industry-characteristic dictionary \( I \) includes the reorganized business directory, the first matching entity names \( e \) and short names \( C \). Then, industry dictionary needs to add the corresponding characteristic keywords \( a \) from user comment texts, description words \( b \), shown in Fig. 2(b).

In the second matching process, industry-characteristic dictionary is needed. By using Jaro–Winkler distance, each keyword \( a \) (in test set) is compared to the characteristic keywords \( a \) and description words \( b \) (in train set) from the industry-characteristic dictionary. Then the unique entity-industry pair can be found between the characteristic keywords and industry-characteristic dictionary (in Fig. 3), which is the target entity identified.

The second matching process is to solve the ambiguous problem. In Fig. 3, three potential entities correspond to the three names in business directory of comment 1 after the first matching. Which entity is the user most concerned? Offering one of the answers in this work is to find that the characteristic keywords matching, such as characteristic keyword \( a_1 \) in both user comments and industry-characteristic dictionary. Then \( C_1 \) is the identified as target entity from comment 1. For comment 2, a name \( c_2 \) corresponds to three entities in business directory multiply. According to second matching characteristic keywords both in comments and industry-characteristic dictionary, like industry \( I_2 \) appears twice, and industry \( I_3 \) once. The name \( C_3 \) corresponding to industry \( I_3 \) is considered as target entity.

Besides, two functions are applied in TSM process. One is removing comments irrelevant to the entity. As shown in Fig. 3, the characteristic word \( a_1 \) related to name \( c_1 \) corresponding to industry \( I_1 \), but an irrelevant keyword \( b_3 \) is removed during the process. The other one is to exclude redundant comments. As shown in Fig. 3, one name \( c_2 \) corresponds to two industries, but the second matching result ignores the less relevant industry \( I_2 \) which only includes one characteristic keyword \( a_1 \).
Table 1
Examples in business directory.

| Full name                  | Administrative divisions | Short name | Industry | Organizational form       | Description (partial)                                      | Added-value telecommunication service; Wholesale and Retail; Computer Hardware etc. |
|----------------------------|--------------------------|------------|----------|---------------------------|------------------------------------------------------------|----------------------------------------------------------------------------------|
| Honeywell Investment Co., Ltd. | Shanghai                | Honeywell | Investment Co., Ltd. | Agent, R&D, sale etc. Zhejiang Taobao Network Co., Ltd. |                                                                                           |                                                                                  |
| Shenzhen Huawei Technology Co., Ltd. | Shenzhen                | Huawei     | Technology Co., Ltd. | Development, production, sales and technical services of wireless communication equipment, software, etc. |                                                                                           |                                                                                  |

Fig. 2. The construction of industry-characteristic dictionary. The first match happens between the potential names c from user comment texts and short name C from business directory, shown in (a). Then, reconstruct pairs according to industry, shown in (b). For instance, two entities $c_1(C_1)$ and $c_2(C_2)$ belong to industry $I_1$, where the two entities are mentioned twice (through keywords $a_1$). In this illustration, filter the keywords $a$ that appear more than 1 time. In empirical experiment, filter the keywords $a$ that appear more than 10 times in each industry. Then, extract the descriptions of entities ($C_1, C_2$) as description words ($b_1, b_2$) for the second matching process. So industry-characteristic dictionary contains both characteristic keywords $a$ and industry description words $b$.

Fig. 3. The illustration of two examples for TSM process to identify one target entity. The first match happens between potential names $c$ from user comment texts and short names $C$ to find the entities mentioned by users. The second matching process involves user comment keywords and the industry-characteristic dictionary. For comment 1, the target entity is $C_1$, since only keyword $a_1$ from user comments matches characteristic keyword $a_1$ in the industry-characteristic dictionary. For comment 2, although name $c_3$ could catch three entities uncertainly, keyword $a_2$ and $b_2$ can match the industry-characteristic dictionary in the second matching process. They belong to the industry $I_2$ and correspond to name $C_3$ according to the first match result. So the target entity for comment 2 is $C_3$. 
2.3. Evaluating the Two-Step-Matching framework

In order to evaluate the results of the TSM entities, two indicators are defined as:

1) association rate, $R_{ass}$: the number of records matched by the TSM framework/the number of comment texts in the test set. This indicator is a global indicator, which does not involve artificial verification;

$$R_{ass} = \frac{\sum_{c=1}^{N} |\mathcal{C}^{\text{match}}_c|}{|\mathcal{T}|}$$  

2) accuracy rate, $R_{acc}$: the number of correct matching records / the number of artificial matching records.

$$R_{acc} = \frac{\sum_{c=1}^{N} |\mathcal{C}^{\text{match}}_c \cap \mathcal{C}^{\text{artificial}}_c|}{|\mathcal{C}^{\text{artificial}}_c|},$$  

where $|c|$ is the total number of potential names and $N$ is the number of comments in the test set, $|\mathcal{C}^{match}_c|$ is the matching set of the second matching method, and $|\mathcal{C}^{artificial}_c|$ refers to the matching set of artificial labelling method. In the test sets, when one comment includes multi-entities, for the first match as benchmark, consider the first mentioned entity as correct matching record $\mathcal{C}^{match}_c$. Since the second match method could identify the multi-entities, the $\mathcal{C}^{match}_c$ is only one.

3. Matching framework and results

3.1. DATA description

This study is based on public opinion monitoring of major public health events, related to Shopping (Consumption of E-commerce Platforms), Tourism (Reservation of Hotels and Air Tickets) and Rent (Rent of Houses and Cars). Therefore, the team selected 3 categories for sample pool including: Shopping(454 comments), Tourism(285 comments), Rent (391 comments), from 1 Jan to 1st Feb 2020. User comments are sentence-level texts. To construct business directory and industry-characteristic dictionary, entity registration details from China State Administration for Market Regulation are acquired. This will be used in the TSM process to identify the target entity below.

3.2. The construction of matching framework

The BiLSTM-CRF model is employed to obtain the entity names from the user comment texts at first stage. Then it matches with short name in business directory, but there could be some multiple pairs. After that matching process is the complement for targeting the most accurate entity. The training set and the test set is divided by 6:4. The training set is to obtain the industry-characteristic dictionary. The details of TSM process are demonstrated in Fig. 4.

The input of this method is user comments, and the output is the target matching entity list. For each comment text, comment set {a, c} is to match the business directory (C, I) by Jaro-Winkler distance algorithm (Fig. 4 step 1). If only one name corresponds to the business directory, $c_i = C_j$, then complete the match and the record goes into the final matching list. If there is a name that corresponds to multiple entity names $c_i \in \{C_j, I_j\}$ (Fig. 4 step 2), then it is necessary to check whether each entity name appears in the comment. If there is an entity name in the business directory $C_j + I_k \in \{a\}$ (Fig. 4 step 3), the match is complete as well.

However, if there are multiple names in the business directory \{c\} n (C, I) (Fig. 4 step 4), it needs to count the number of the entity names that appears in the comment (Fig. 4 step 5). If the number of occurrences of a name is higher than other names $n_1 > n_2 \ldots$ (Fig. 4 step 6), then it needs to check whether the entity name is in potential name set $C_j + I_k \in \{c\}$ (Fig. 4 step 7). If the result comes out as true, the match is completed; otherwise, it is unmatched. If there are multiple entity names with the same count at the same time $n_1 = n_2 = \ldots = n_m$, then it needs to check whether each entity name that appears in the characteristic words set again. If only one entity name hits and other entity names do not $C_j + I_k \notin \{c\}$. $C_j + I_k \notin \{c\} \ldots C_j + I_k \notin \{c\}$ (Fig. 4 step 8), the match is completed.

In order to improve the efficiency of second matching process, the industry characteristic words set is excluded from the entity set \{a\} - \{I\} (Fig. 4 step 9). After that, if only one name remains in the name set which corresponds to multiple entity names $c_i \in \{C_j, I_j\}$ (Fig. 4 step 10), then returns to the step (3) and check as first match. If there is a single mapping relationship between the name set and the business directory set \{c\} → (C, I) (Fig. 4 step 11), that is, multiple names correspond to the entities one by one, the second matching process is started.

In the second matching process, the industry-characteristic dictionary is needed. Firstly, the characteristic words between user comments and industry-characteristic dictionary need to be matched again (Fig. 4 step 12). If the number of second matching of a entity name is higher than other names, $n_1 > n_2 > \ldots$ (Fig. 4 step 13), this second matching process completes, then puts the most frequent occurring entity names into final matching list. The step (13) already includes a single mapping relationship, and repeated checking does not need to do like the step (7) in Fig. 4. If there are multiple entity names with the same number at the same time $n_1 = n_2 = \ldots = n_m$, it needs to look up whether each entity name emerges in the comment name set. If only one entity name hits and other entity names do not $C_j + I_k \notin \{c\}$. $C_j + I_k \notin \{c\} \ldots C_j + I_k \notin \{c\}$ (Fig. 4 step 14), the second match is completed. Eventually, the second matching results form a matching list.

3.3. Experiment results

Take Table 2 as an example. After the first match, three entities are identified, but still it cannot confirm the specific target one. Furthermore, the second matching process identifies “Honeywell” by characteristic keywords, i.e. “N95 mask”, “wearing” and “air valve” are industry characteristics. Although the words “photos” and “mobile phone” also represent the characteristic of “Huawei”, the number of characteristics of “Honeywell” occurs more. Therefore, the target entity can be identified.

In order to check the results matched the real entities precisely, we randomly select 150 matching records from three data sets, and manually label entities according to the business directory. Then the artificial matching results are compared with the list of the second matching method.

The results of three kinds of text corpus are illustrated in Table 3. The first matching process uses the BiLSTM-CRF before Jaro-Winkler match. It is obvious from this table that the matching method of BiLSTM-CRF included is better than that of CRF model and Jaro-Winkler alone. BiLSTM-CRF matching can avoid errors caused by segmentation cutting errors compared to Jaro-Winkler method. For example, “I bought a Jingdongfang screen”. If Jaro-Winkler’s matching method is used only, the word that is eventually cut out is Jingdong, which is a wrong result. Conversely, the word that BiLSTM-CRF finally cuts out according to the longest matching field is Jingdongfang. So the $R_{ass}$ and $R_{acc}$ obtained by the BiLSTM-CRF matching method are better than the matching method of Jaro-Winkler. In three different categories, associated rate $R_{ass}$ has increased from 0.39, 0.43 and 0.35 to 0.55, 0.60 and 0.49, respectively. Meanwhile, accuracy rate increased from 0.44, 0.45 and 0.41 to 0.66, 0.69 and 0.61. Due to BiLSTM-CRF model combining the advantages of BiLSTM model and CRF model, compared to the traditional CRF model, the average value of $R_{ass}$ and $R_{acc}$ of BiLSTM-CRF model has increased from 0.46 and 0.57 to 0.55 and 0.65, respectively, the results of BiLSTM-CRF model are better than that of CRF model. The BiLSTM model can automatically extract the features of the observation sequence according to the target, but the disadvantage is that it cannot learn the relationship between the state sequences. Whereas the advantage of CRF is that it can model the implied state and learn the characteristics of the state sequence, but its disadvantage is that it needs to manually extract the...
sequence features. So the general approach is to add a layer of CRF after the BiLSTM to get the best of both. Then, the results of first matching with BiLSTM-CRF are as a benchmark for comparison.

Both $R_{att}$ and $R_{acc}$ of second matching framework proposed in this work perform much better than only one matching results. In Shopping and Rent categories, $R_{att}$ has increased to 0.93 (increased 38%) and 0.95 (increased 29%), respectively. The increments indicate that the comments in the two categories include multiple entities mapping relations, i.e., users' intents are on one entity mainly, but other irrelevant entities are also mentioned. The second matching process, moreover, improves the matching results simply because of the industry characteristic terms. Meanwhile, in rent category, $R_{att}$ has increased from 0.49 to 0.77 (increased 28%), probably since involving the wide coverage of the industries. $R_{acc}$ of the second matching results are all over 0.9 in three categories (averagely increased 30%), indicating the TSM framework is much closer to the human recognition.

![Fig. 4. The framework of Two-Step-Matching processes.](image)

| Table 2 | The example of matching results. |
|---------|----------------------------------|
| **Matching rule** | **A User’s Comment** | **Entity 1** | **Entity 2** | **Entity 3** |
| First match | This **Honeywell** N95 mask bought from **Taobao** is very easy to use. It is very comfortable to wear after half day wearing. The air valve is not blocked in any way. Please see the attached photos which was taken by **Huawei** mobile phone. Must be given high praise! | **Honeywell** | **Taobao** | **Huawei** |
| Second match | This **Honeywell** N95 mask bought from **Taobao** is very easy to use. It is very comfortable to wear after half day wearing. The air valve is not blocked in any way. Please see the attached photos which was taken by **Huawei** mobile phone. Must be given high praise! | **Honeywell** | **Taobao** | **Huawei** |

| Table 3 | Results of empirical experiments. |
|---------|---------------------------------|
| **Category** | **First match** | **Second match** | **Improvement vs BiLSTM-CRF** |
| | **Jaro–Winkler alone** | **with CRF (compared)** | **with BiLSTM-CRF** | **Jaro–Winkler alone** | **with CRF (compared)** | **with BiLSTM-CRF** | **Jaro–Winkler alone** | **with CRF (compared)** | **with BiLSTM-CRF** |
| Shopping | $R_{att}$ | $R_{acc}$ | $R_{att}$ | $R_{acc}$ | $R_{att}$ | $R_{acc}$ | $R_{att}$ | $R_{acc}$ | $R_{att}$ | $R_{acc}$ | $R_{att}$ | $R_{acc}$ |
| 0.39 | 0.44 | 0.50 | 0.61 | 0.55 | 0.66 | 0.93 | 0.95 | 38% | 29% |
| Tourism | 0.43 | 0.45 | 0.53 | 0.62 | 0.60 | 0.69 | 0.90 | 0.97 | 30% | 28% |
| Rent | 0.35 | 0.41 | 0.44 | 0.54 | 0.49 | 0.61 | 0.77 | 0.95 | 28% | 34% |
| **Avg.** | 0.39 | 0.43 | 0.46 | 0.57 | 0.55 | 0.65 | 0.87 | 0.96 | 32% | 30% |
4. Conclusion and discussions

In the safety science, such as the public opinion, emergencies and so on, the entity identification plays crucial role for the early warning and diagnosis and intervention which could reducing the response time greatly. This paper demonstrates the Two-Step-Matching method designed to identify the precise target from ambiguous user comments of public opinions. Firstly, potential entity from user comments are extracted by BiLSTM-CRF model and user characteristic keywords are extracted by TF-IDF algorithm. Secondly, the potential entity is matched with the official business directory by Jaro–Winkler distance algorithm to form the match pair. Then, the industry-characteristic dictionary is used to find the precise target, which includes characteristic keywords and description words. In the second matching process, the target company can be found with large count of matching among the characteristic keywords mentioned in comments and industry-characteristic dictionary. Meanwhile, two evaluation indicators are defined in this work to verify the proposed method: 1) association rate, \( R_{ass} \): the number of records matched via the TSM matching method/the number of comment records; 2) accuracy rate, \( R_{acc} \): the number of accurate matched records/the number of artificial matched records. In the empirical experiments, three categories of user comments about major public health events, i.e., Shopping, Tourism and Rent are verified. Comparing with the case of using the first matching process alone, the TSM method shows that the associated rate \( R_{ass} \) has increased 38%, 30%, 28%, respectively. Meanwhile, accuracy rate has increased 29%, 28%, 34%, respectively. Average speaking, the association rate and accuracy rate for the shopping, tourism and rent data sets could be enhanced by 32% and 30%.

For practical application, this proposed method is convenient to be employed in perception of social media [23]. For example, in public opinion application, entity identification techniques can be employed to draw attention to the trends of the public opinion, to mine the opinions about competing entities in typical opinion mining applications, such as users’ attitudes about the competing products or brands, and to discover the early warning signals about the product’ drawbacks in the details of comment texts in product quality management [24, 25]. Therefore, it is significant for a system to automatically identify the true target entities that users really talk about from relevant comment texts.

Nevertheless, several other issues should be investigated in future. For instance, the TSM method proposed in this paper can also be employed in the knowledge graph research to identify the named entity and to construction the relationship among entities, so comparing the results with other knowledge graph construction techniques is one of the future works. It can also calculate the similarities by mapping the characters into the high-dimensional with various graph embedding techniques, such as the node embedding with node information and network information [26]. What’s more, the main operational limitation is that our method is suitable for the ambiguity of multiple targets appeared simultaneously. If the user’s expressions with multiple targets appear different times, such as entity A appears three times, entity B appears twice times, and entity C appears once, but the user is actually concerned about entity C. In this case, our method cannot recognize the result because we recognize entity A as the final target. Therefore, semantic information of context can be considered to explore more information from users, such as synonyms, upper and lower words. Professional word dictionary may be helpful to speed up the identification and training of the industry-characteristic dictionary could be iterated according to text corpus. Moreover, it is necessary to keep balance among complicated text corpus structure, time consuming of training process, and accuracy of the results.

CRediT authorship contribution statement

Ren-De Li: Methodology, Writing - original draft, Writing - review & editing. Hao-Tian Ma: Data curation, Software. Zi-Yi Wang: Writing - original draft, Writing - review & editing. Qiang Guo: Supervision, Funding acquisition. Jian-Guo Liu: Conceptualization.

Acknowledgments

We thank Qiang Yue for the preliminary experience of industry application.

References

[1] J.T. Wu, K. Leung, G.M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study, The Lancet 395 (10225) (2020) 689–697.
[2] E.-P. Lin, J. Srivastava, S. Wehkamp, J. Richardson, Entity identification in database integration, Inf. Sci. 89 (1–2) (1996) 1–38.
[3] M.A. Hernández, S.J. Stolfo, The merge/purge problem for large databases, ACM Sigmod Record 24 (2) (1995) 127–138.
[4] M.A. Hernández, S.J. Stolfo, Real-world data is dirty: data cleansing and the merge/purge problem, Data Min. Knowl. Discov. 2 (1) (1998) 9–37.
[5] A. Arasu, S. Chaudhuri, R. Kaushik, Transformation-based framework for record matching, in: Proceedings of the 24th IEEE International Conference on Data Engineering, IEEE, 2008, pp. 40–49.
[6] A. Arasu, R. Kaushik, A grammar-based entity representation framework for data cleaning, in: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, 2009, pp. 233–244.
[7] W. Fan, X. Wang, Y. Wu, Incremental graph pattern matching, ACM Trans. Datab. Syst. (TODS) 38 (3) (2013) 1–47.
[8] W. Fan, J. Li, X. Wang, Y. Wu, Query preserving graph compression, in: Proceedings of the 2012 ACM SIGMOD International Conference on Management of data, 2012, pp. 167–168.
[9] W.W. Cohen, Data integration using similarity joins and a word-based information representation language, ACM Trans. Inf. Syst. (TIS) 18 (3) (2000) 288–321.
[10] N. Augsten, M. Bohnen, C. Dyreson, J. Gamper, Approximate joins for data-centric xml, in: Proceedings of the 24th IEEE International Conference on Data Engineering, IEEE, 2008, pp. 814–823.
[11] E. Schallenth, K.-U. Sattler, G. Saake, Efficient similarity-based operations for data integration, Data Knowl. Eng. 48 (3) (2004) 361–387.
[12] G. Rizzo, R. Troncy, Nerd: a framework for unifying named entity recognition and disambiguation extraction tools, in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2012, pp. 73–76.
[13] M. Song, W.C. Kim, D. Lee, G.E. Heo, K.Y. Kang, Pked4j: entity and relation extraction for public knowledge discovery, J. Biomed. Inform. 57 (2015) 320–332.
[14] A. Pappu, R. Blanco, Y. Mehdad, A. Stent, K. Thadani, Lightweight multilingual entity extraction and linking, in: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, 2017, pp. 365–374.
[15] D.B. Nguyen, M. Theoibald, G. Weikum, J-reed: joint relation extraction and entity disambiguation, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 2227–2230.
[16] L. Canale, P. Liseva, R. Troncy, A novel ensemble method for named entity recognition and disambiguation based on neural network, in: Proceedings of the International Semantic Web Conference, Springer, 2018, pp. 91–107.
[17] C. Dong, J. Zhang, C. Zong, M. Hattori, H. Di, Chord-based LSTM-CRF with radical-level features for chinese named entity recognition, in: Natural Language Understanding and Intelligent Applications, Springer, 2016, pp. 239–250.
[18] T. Chen, R. Xu, Y. Fei, X. Wang, Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN, Exp. Syst. Appl. 72 (2017) 221–230.
[19] L. Ling, Z. Yang, Y. Pei, Z. Yin, W. Lei, H. Lin, W. Jien, An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition, Bioinformatics (8) (2017) 8.
[20] K.S. Jones, A statistical interpretation of term specificity and its application in retrieval, J. Doc. 28 (1) (1972) 11–21.
[21] S.I. Halak, A. Kamsin, P. Shivakumara, G.A. Gilkar, W.Z. Khan, M. Imran, Exact string matching algorithms: survey, issues, and future research directions, IEEE Access 7 (2019) 69614–69627.
[22] W.E. Winkler, String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage, in: Proceedings of the Section On Survey Research Methods, American Statistical Association, 1990, pp. 354–359.
[23] I.B. Derczynski, R. Yang, C.S. Jensen, Towards context-aware search and analysis on social media data, in: Proceedings of the 16th International Conference on Extending Database Technology, 2013, pp. 137–142.
[24] L. Zhang, B. Liu, Aspect and entity extraction for opinion mining, in: Data Mining and Knowledge Discovery for Big Data, Springer, 2014, pp. 1–40.
[25] B. Zhong, X. Xing, H. Luo, Q. Zhou, H. Li, T. Rose, W. Fang, Deep learning-based extraction of construction procedural constraints from construction regulations, Adv. Eng. Inf. 43 (2020) 101063.
[26] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, in: Proceedings of the Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017, pp. 1024–1034.
Technology. His research interests are information science, human behavior and big data mining.

Hao-Tian Ma received the B.S. Engineering degree in Automation from Harbin University of Science and Technology, China, 2013 and M.S. Science degree in Finance from the University of Dundee, UK, 2019. His research interests are option volatility and private equity financing.

Zi-Yi Wang received the B.S., M.S., and Ph.D. degrees in Management Science and Engineering in Shanghai University of Finance and Economics, China, in 2009, 2012, and 2017, respectively. She is currently a Research Assistant in Economic Journalism Department, School of Humanities, Shanghai University of Finance and Economics. Her research interests include social network analysis, computational communication, and natural language processing.

Qiang Guo is Professor of University of Shanghai for Science and Technology, Ph.D. supervisor. She was funded several National Natural Science Foundations and published 119 papers including 66 SCI indexed papers. Her H index is 18. His current research interests include: big data analysis, knowledge management and network analysis.

Jian-Guo Liu is Chair Professor of Shanghai University of Finance and Economics, Ph.D. supervisor, Dean of Sina WRD Big Data Research Institute, Distinguished Professor of Shanghai Oriental Scholar, Shanghai Shuguang Scholar, and Deputy Dean of Shanghai University of Finance and Economics. His current research interests include: Media big data modeling and analysis, knowledge management, and financial management.