PAPER

Entity Identification on Microblogs by CRF Model with Adaptive Dependency*

Jun-Li LU[†a)], ***Nonmember***, **Makoto P. KATO**[†], ***Member***, **Takehiro YAMAMOTO**[†], ***and*** **Katsumi TANAKA**[†], ***Nonmembers***

SUMMARY We address the problem of entity identification on a microblog with special attention to indirect reference cases in which entities are not referred to by their names. Most studies on identifying entities referred to them by their full/partial name or abbreviation, while there are many indirectly mentioned entities in microblogs, which are difficult to identify in short text such as microblogs. We therefore tackled indirect reference cases by developing features that are particularly important for certain types of indirect references and modeling dependency among referred entities by a Conditional Random Field (CRF) model. In addition, we model non-sequential order dependency while keeping the inference tractable by dynamically building dependency among entities. The experimental results suggest that our features were effective for indirect references, and our CRF model with adaptive dependency was robust even when there were multiple mentions in a microblog and achieved the same high performance as that with the fully connected CRF model.

key words: *entity identification, indirect reference, Conditional Random Field*

1. Introduction

People write microblogs to share their opinions about information in the real world and refer to entities, such as people and places, in various ways. Entity identification, a problem of identifying entities referred to in a document, has been extensively addressed in the literature [1], [2]. Most related studies addressed the problem of identifying entities referred to by their full/partial name or abbreviation. In microblogs, however, there are many entities mentioned in an *indirect* way, which are often difficult to identify, especially in short text such as microblogs. Given a microblog “Jacoby Ellsbury is leaving for the rival”, *e.g.*, it must be difficult to understand what “the rival” is if one does not know the fact that “Jacoby Ellsbury, a baseball player who belonged to the Boston Red Sox, was being traded to the New York Yankees, a rival of the Boston Red Sox”.

In this paper, we address the entity identification problem with special attention to *indirect reference* cases in which entities are not referred to by their names. Compared with *direct reference* cases, where entities are referred to by their names, indirect references usually have more candidate entities and are accordingly difficult to identify. We therefore consider more evidence for entity identification that is especially important for indirect references. *E.g.*, we use syntactic patterns that are useful in identifying indirectly referred entities. If patterns, such as “Y known as X” and “X regarded as Y”, appear many times, they can be strong cues that X is used to refer to Y and be only applicable to indirect reference cases. Another example of effective features is a set of words in a microblog that is *translated into* a topic-specific language. *E.g.*, if we know that “baseball” is the topic of the microblog “Jacoby is the best player in the world!”, we may translate “player” to “outfielder” and recognize “Jacoby” as “Jacoby Ellsbury”.

We also use dependency among entities in a microblog for our entity identification problem. Since multiple entities can be mentioned in a microblog and be strong clues to identify other entities, predicting referred entities together is essential for accurate entity identification. We therefore use a Conditional Random Field (CRF) model [3], [4] to model the correlation among referred entities in a microblog. A limitation of a simple application of linear CRF models is that they cannot explicitly represent cases in which distant entities are dependent or neighboring entities are independent. Thus, we also model the non-sequential order dependency while keeping the inference tractable by dynamically building dependency among entities.

Experiments were carried out with 461 tweets that were randomly collected from Twitter[**]. The experimental results suggested that 1) our features were effective for indirect references, 2) our CRF model with adaptive dependency was robust even when there were multiple mentions in a microblog, and 3) it achieved the same high performance as the fully connected CRF model. We summarize the contributions of this work:

- We address the problem of identifying indirectly mentioned entities and argue that there is a considerable amount of indirect references in real microblog data.
- We propose features built upon characteristics of indirect references and a CRF model that dynamically builds dependency for identifying referred entities.
- We demonstrate that our features worked for both direct and indirect references and the CRF model with adaptive dependency is advantageous over several vari-

Manuscript received January 12, 2016.
Manuscript revised May 23, 2016.
Manuscript publicized June 20, 2016.

[†]The authors are with Dept. of Social Informatics, Graduate School of Informatics, Kyoto University, Kyoto-shi, 606–8501 Japan.

[*]accepted in Proc. of the 2015 IEEE/WIC/ACM International Conference on Web Intelligence. The main improvement of this journal is in Sect. 4, Sect. 5.2, Sect. 7.3, and Sect. 7.4,

a) E-mail: jllu@dl.kuis.kyoto-u.ac.jp

DOI: 10.1587/transinf.2016EDP7015

[**]https://twitter.com/.

ants of CRF models.

The rest of this paper is organized as follows. We discuss the related work on entity identification in Sect. 2. We explain how an entity is referred to by a direct/indirect reference and the entity identification problem in the presence of direct/indirect references in Sect. 3. We propose a method of generating candidate entities for a mention in Sect. 4, propose features for indirect references in Sect. 5, formulate the entity identification problem by a CRF model, and propose a method of dynamically acquiring dependency among entities in a CRF model in Sect. 6. We conduct a survey on real-world microblog data, discuss the experimental results on candidate entity generation and entity identification, and give a case study in Sect. 7. We conclude the paper in Sect. 8.

2. Related Work

Entity identification has been an area of active research and attracted many researchers. There are two kinds of problems relating to entity identification. The first is to identify which part of words (denoted as a mention) in a document refers to an entity [5]–[9]. The second problem is to identify which entity is referred to by that mention [1], [2], [10]–[12]. Some studies have attempted to address both problems together to identify the referred entity by the detected mention [7]. For identifying entities, one proposed mining additional context in addition to the knowledge base [13] and one proposed time-and-space-efficient algorithms [14]. Some tackled the entity identification problem on well-written documents [10], [13] and on short-and-noisy microblogs such as tweets [5]–[9]. To generate referred entities for a given mention, most studies used the name of entities (denoted as direct reference in our study). However, this approach is not applicable if entities are referred to by implicit expressions (denoted as indirect reference). *E.g.*, the text “the president” is used to refer to the entity “Barack Obama” and the text is not entity’s name. We shed light into indirect reference cases and proposed a method to address the entity identification problem for entities of direct/indirect references.

To evaluate a referred entity, the features used in previous studies include the probability that mentions entity’s name [7], [15], the similarity between the context of entities and a document [10], [13], [16], the correlation between multiple entities [17]–[19], and user interest, which was estimated from user’s microblogs [12]. As we discussed in our experiments, these features were not effective for indirect references. Since a mention in an indirect reference often consists of normal nouns, which can be abstract, and making acquiring the information of referred entities difficult, we suggest to translate a microblog text including mentions by the related topics and to use the translated content to accurately evaluate referred entities. Furthermore, contrary to previous studies, we use syntactic and linguistic features, such as “entity-known-as pattern”, which is to evaluate an entity by checking whether the entity is *known as* by another object or not. Finally, we predict referred entities in a microblog together using a Conditional Random Field (CRF) model [3], [9] while some studies predicted the entity on a single mention basis [2], [7], [12]. We also observed that mentions in a microblog may come from different topics, which implies that the dependency among corresponding referred entities are different. To improve prediction accuracy, we propose a greedy algorithm to dynamically acquire dependency among referred entities in a CRF model.

3. Problem Definition: Entity Identification on Microblogs

We introduce several terms to define the problem addressed in this paper as shown in Fig. 1. *Microblog*: a microblog is a sequence of words posted by a user on a social media platform, *e.g.*, a tweet posted by a user on Twitter is a kind of microblog. *Mention*: a mention is a sequence of words in a microblog and is used by a user to refer to some entity, *e.g.*, a person, object, or place, in the real world. Note that we assume a one-to-one relationship between an entity and a Wikipedia article[†]; thus, an entity must correspond to a single Wikipedia article. *E.g.*, there are five mentions, each of which has underlined words, in the microblog in Fig. 1.

How an entity is mentioned. Previous studies [1], [5] used the name of entities to generate referred entities as follows. **Direct reference**: if mention m is used to refer to entity e and m is the full/partial name or the abbreviation of the name of e, the reference to e by m is a direct reference. Note that the name of e is defined by the title of the article of e in Wikipedia. *E.g.*, in Fig. 1, m is “Jacoby” and candidate entities e_1 and e_2 are “Jacoby Ellsbury” and “Jacoby transfer”, respectively, because m is the partial name of e_1 and that of e_2. However, the scope of referred entities by direct reference is limited because people can refer to entities without using their names. Motivated by the research of analogy and metaphor [20]–[22] in linguistics, we target referred entities of the following novel type.

Indirect reference: if mention m is used to refer to entity e and m is neither the full/partial name nor the abbreviation of the name of e, the reference to e by m is an indirect reference. *E.g.*, e =“Jacoby Ellsbury” can be indirectly referred to by each m of the following cases: 1) m is the attribute that describes e, *e.g.*, m =“the player”. 2) m is the relationship between e and another entity e_2, *e.g.*, m =“outfielder”, where e is an outfielder of e_2 =“New York Yankees”. 3) By metaphor, m contains the name of another entity e_2, *e.g.*, m =“our Ichiro Suzuki”, where e_2 =“Ichiro Suzuki” and the referred e are both outfielders. Note that we avoid unrealistic references between a mention and an entity by limiting a mention to be a noun-phrase. We therefore define the problem as follows:

Definition 1: (*Entity Identification on microblogs in the presence of Direct/Indirect References, EI-DIR*): given a microblog b from a user u, a set of mentions M_b extracted from

[†] https://en.wikipedia.org/wiki/.

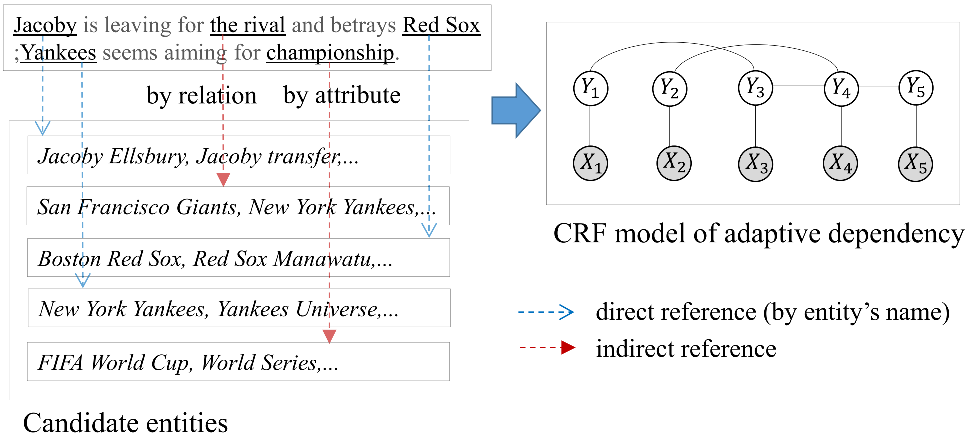

Jacoby is leaving for the rival and betrays Red Sox ;Yankees seems aiming for championship.
by relation
by attribute
Jacoby Ellsbury, Jacoby transfer,...
San Francisco Giants, New York Yankees,...
Boston Red Sox, Red Sox Manawatu,...
New York Yankees, Yankees Universe,...
FIFA World Cup, World Series,...
Candidate entities
Y1
Y2
Y3
Y4
Y5
X1
X2
X3
X4
X5
CRF model of adaptive dependency
direct reference (by entity's name)
indirect reference

Fig. 1 Entity identification on a microblog in the presence of direct/indirect references, where there are five mentions "Jacoby", "the rival", "Red Sox", "Yankees", and "championship" denoted as X_1, X_2, X_3, X_4, and X_5 and the corresponding referred entities denoted as Y_1, Y_2, Y_3, Y_4, and Y_5, respectively.

b, and a set of entities E, the *EI-DIR* problem is to identify the entity $e \in E$, which is referred to by each mention $m \in M_b$, through a direct or indirect reference.

4. Candidate Entity Generation

Given a mention of a microblog, we acquire candidate entities that are possibly referred to by the mention, as follows. Because an entity e is mentioned from a microblog b on a social media platform p, we argue that there are writer's other microblogs, other users' microblogs on platform p, or web pages (*e.g.*, news articles) containing information related to e and these data are useful to search for candidate entities. Therefore, we use search queries Q that are extracted from different document collections mentioned above. The search queries are

$$Q = \{(w_1, \varphi_1), (w_2, \varphi_2), \ldots, (w_i, \varphi_i)\},$$

where φ_i is the weight of word w_i. The process for extracting Q is shown as follows.

At first, we extract the set of top words Δ with the highest tf-idf weights, from different document collections denoted by S_1, S_2, and S_3. S_1 is the related writer's microblogs searched by mention m, S_2 is the related users' microblogs searched on platform p using microblog b's hashtags (*e.g.*, "#yankees" in Twitter), and S_3 is the related web pages searched on Web using m. Because proper-noun words could reflect about the topic of entity e, we keep only proper-noun words in S_1, S_2, and S_3 and add user's description on platform p from user-hashtags into S_1 and S_2 (*e.g.*, "@RedSox" is a user-hashtag in Twitter). Second, because Δ is extracted from the documents of different domains (*i.e.*, S_1, S_2, and S_3), we make the search query Q from the words of Δ with words of balanced weighting on different domains as follows.

1. At first, for the words of Δ, the weight $\varphi_{i,\mathcal{S}}$ of each word w_i from each domain $\mathcal{S}$ (*e.g.*, words of S_1 are from one domain) is normalized as $\varphi'_{i,\mathcal{S}}$ (we normalized by the max weight).
2. We determine the importance of each domain $\mathcal{S}$ as the weight $\varphi_{\mathcal{S}}$.
3. We sum up the weight of each word w_i from all domains as $\varphi_i = \sum_{\mathcal{S}\in\{S_1,S_2,S_3\}} \varphi_{\mathcal{S}} * \varphi'_{i,\mathcal{S}}$.

5. Features

We use features to evaluate a direct or indirect reference between a mention and a referred entity. We especially focus on indirect reference cases and design features for them by comparing previous studies. Given a set of mentions M (denoted as M for convenience) from a microblog, for each candidate entity e of each mention $m \in M$ and for each pair of candidates e and e_2 of a given pair of mentions $m \in M$ and $m_2 \in M$, we use the following features. At first, we consider two basic features. **Keyword**: we consider whether the keywords K_e, which describe entity e and are extracted from the main (the first-k) paragraphs and the titles of categories in the article of e in Wikipedia, can match mention m. *E.g.*, "outfielder" is one keyword for "Jacoby Ellsbury". Therefore,

$$sim(K_e, \{m\}),\ \ sim(K_e, \{m, m_2, n_{e_2}\}).$$

Note that K_e is the *top-k words* that are picked according to the weight tf$_i\cdot$ idf$_i$ of each given word w_i, where tf$_i$ is the frequency of w_i in all the specified documents[†] (or words-windows) and idf$_i$ is the inverted document frequency of w_i; $sim(A, B) = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$ is the similarity between the bag-of-words model of A and that of B, where the i-th element of **a** (or **b**) belonging to A (or B) is the weight tf$_i\cdot$ idf$_i$ of word w_i. We also examine the neighboring mention m_2 and the name n_{e_2} of the neighboring entity e_2 in the same microblog by the similarity $sim(K_e, \{m, m_2, n_{e_2}\})$.

Occurrence frequency: we consider the co-occurrence

[†] A document is an article in Wikipedia.

among entities e and e_2 and mention m in documents as

$$\log r(n_e),\ \log r(n_e, m),\ \log r(n_e, n_{e_2}),$$

where $r(n_e, m)$ is the number of times the name n_e of e and m co-occur in documents.

5.1 Topic-Specific Translation on Microblogs

The difficulty in an indirect reference is that we cannot acquire referred entities by the mention. The reason is that a mention in this case often consists of normal nouns, which can be *abstract* and difficult to match entity's information, *e.g.*, the *topic* of entities. For instance, in the microblog "X is the best player", the mention "player" can refer to entities of a variety of topics, *e.g.*, a referred entity is an "outfielder" in the topic of baseball or "goalkeeper" in the topic of football. **Topic-specific translation**: we therefore translate a microblog including mentions based on the related topics and use the translated content to accurately evaluate referred entities. As in the above example, we may translate the mention "player" based on the topic of the microblog, into either "outfielder" or "goalkeeper" in an ideal situation. The process of translating the content W_b of a microblog b is as follows.

1. We extract the top proper-nouns P_b, which are related to microblog b of writer u. The words of P_b are extracted from W_b, news linked by b, and descriptions of users tagged in b, and the recent microblogs written by u.
2. We then extract the words I_b from Wikipedia documents, which often co-occur with the words of P_b.
3. Using I_b, we translate word $w \in W_b$ by word $w' \in I_b$ if w and w' are semantically similar, *i.e.*, $T_b = \{w' | \lambda(w, w') \geq \theta, w \in W_b, w' \in I_b, 0 \leq \theta \leq 1\} \cup W_b$. Note that stop-words and proper nouns in microblog b are not needed to be translated and we keep the weight of each translated word from I_b.

The semantic similarity $\lambda(w, w')$ between words w and w' is measured by WordNet Similarity [23], which takes into account the lexicalized concept, synset, taxonomy, and dictionary definitions for all words and is provided by this tool "WS4J"[†].

With the translated microblog T_b, we evaluate an entity e by computing the similarity between T_b and the context C_e of e, which is a set of words from each words-window surrounding the name n_e of e from documents, *e.g.*, a words-window is "n_e hits a home-run". Therefore,

$$sim(C_e, T_b),\ sim(C_{ee_2}, T_b). \tag{1a-b}$$

We also consider the context of e and the neighboring entity e_2 as C_{ee_2}. To consider the diversity of the context of entities, we separate the context C_e based on parts of speech into a set of nouns $C_{e,n}$ and a set of verbs $C_{e,v}$. *E.g.*, the noun "outfielder" belongs to $C_{e,n}$ and the verb "hit" belongs to $C_{e,v}$. Therefore,

$$sim(C_{e,n}, T_b),\ sim(C_{e,v}, T_b),\ sim(C_{ee_2,n}, T_b),$$
$$sim(C_{ee_2,v}, T_b).$$

5.2 Evidence from Linguistics

Furthermore, we consider linguistic features to acquire more clues based on the following two reasons. First, armed with topic-specific translation, we can identify entities of a certain topic; however, we are still ambiguous about that how people thinks of entities. Especially, how a microblog's writer describes about an entity. Second, for the entity identification problem, few studies went into detail about the following linguistic features we propose.

Entity-known-as pattern: according to a pattern p, where people can utilize p to link an entity e with some object, we check whether e can be represented by mention m. *i.e.*, we check if the words in p appear in sentences containing the name n_e of e and m. *E.g.*, "$n_e \ldots p \ldots m$" is a sentence such that n_e appears and is followed by p and m, where n_e ="Jacoby Ellsbury", p ="known as", and m ="outfielder". Therefore,

$$\log \sum_{p \in P} l(\text{``}n_e \ldots p \ldots m\text{''}),\ \log \sum_{p \in P} l(\text{``}n_e \ldots p \ldots m \ldots n_{e_2}\text{''}), \tag{2a-b}$$

where $l(\text{``}n_e \ldots p \ldots m\text{''})$ denotes the number of times sentence "$n_e \ldots p \ldots m$" appears in documents. We further consider whether a neighboring entity e_2 is involved in by sentence "$n_e \ldots p \ldots m \ldots n_{e_2}$".

Entity's action: people could describe the action, which is performed by an entity e. Therefore, we measure whether e performed an action by checking if e is reasonable to co-occur with the verbs v_s (or v_o), where v_s (or v_o) is the verb of a sentence in microblog b such that the corresponding mention m of e is the subject (or the object), by

$$p(e|v_s),\ p(e|v_o). \tag{3a-b}$$

We compute $p(e|v_s) = \frac{l(\text{``}e \ldots v_s\text{''})}{\sum_{e'} l(\text{``}e' \ldots v_s\text{''})}$ and $p(e|v_o) = \frac{l(\text{``}v_o \ldots e\text{''})}{\sum_{e'} l(\text{``}v_o \ldots e'\text{''})}$ from documents, where the subject in sentence "$e \ldots v_s$" is e and the object in sentence "$v_o \ldots e$" is e.

6. CRF Model

Because there are multiple mentions in a microblog, we want to predict the referred entities of these mentions together by considering their correlation. We therefore formulate the *EI-DIR* problem by using a Conditional Random Field (CRF) model with the advantage of relaxing the independence assumption, which is made by using a Hidden Markov Model (HMM), and avoids the label bias problem [3], [4]. Furthermore, for making the inference that predicts referred entities on a CRF model tractable, we use

[†] https://code.google.com/p/ws4j/

a CRF model with non-cycle dependency, which does not contain cycle connections of dependency among referred entities. That is, the time complexity of the inference on the non-cycle CRF model is $O(nc^2)$, which is linear to the number n of mentions in a microblog and we prove at Appendix C, where c is the number of candidate entities of a mention; however, this complexity on a cycle-connected CRF model can be exponential to n, *i.e.*, $O(c^k)$, where k is the length of the longest cycle and $3 \leq k \leq n$.

Given a set of mentions M from a microblog b, we convert it to a CRF model (X, Y, L) as follows. Each mention $m_i \in M$, $1 \leq i \leq |M|$ is denoted as X_i as an observed value, which is connected to a label Y_i. Y_i is a random variable and the value of Y_i is a candidate of the referred entity of m_i. Note that microblog b is denoted as X_0. Formally $X = X_0 \times X_1 \times \ldots \times X_n$, $Y = Y_1 \times \ldots \times Y_n$, and L is a set of undirected connections of dependency among labels in Y. One example of formulating the *EI-DIR* problem into a non-cycle CRF model is shown in Fig. 1. Given a set of mentions from a microblog denoted as $\mathbf{x} = (x_0, x_1, \ldots, x_n) \in X$, the probability of a candidate of referred entities $\mathbf{y} = (y_1, \ldots, y_n) \in Y$ is

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{z(\mathbf{x})} \exp\Big(\sum_{i=1}^{n} [\sum_{f \in F_\alpha} w_f f(y_i) + \sum_{f \in F_\beta} w_f f(x_i, y_i)] + \sum_{(i,j) \in L} \sum_{f \in F_\gamma} w_f f(y_i, y_j) \Big), \tag{4}$$

where F_α, F_β, and F_γ are a set of features of a single entity, a set of features between a mention and an entity, and a set of features between two entities, respectively, w_f is the weight of feature f, $(i, j) \in L$ denotes that label Y_i is connected to label Y_j, and $z(x)$ is the normalizing constant. The best candidate of referred entities ranked using the CRF model is

$$\mathbf{y}' = \arg\max_{\mathbf{y} \in Y} p(\mathbf{y}|\mathbf{x}). \tag{5}$$

6.1 Adaptive Dependency

Given a microblog, we observed that some mentions in a microblog may come from different topics, which implies that dependency among the corresponding referred entities are different. To enhance the inference on a CRF model, we acquire dependency among referred entities, *i.e.*, to determine and give the weight to any connection between labels in Y in the CRF model. *E.g.*, given a set of mentions from a microblog denoted as $\mathbf{x} = (x_0, x_1, \ldots, x_4) \in X$, suppose mentions x_1 and x_2 are from the topic of "baseball" and those denoted as x_3 and x_4 are from the topic of "education", the connections between Y_1 and Y_2 or between Y_3 and Y_4 are more dependent than those between Y_1 and Y_3 or between Y_1 and Y_4.

We propose a greedy algorithm to dynamically acquire dependency among referred entities on a given microblog as follows. Given a CRF model of $(X, Y, L = \emptyset)$ from a microblog, we sequentially add one important connection $l = (i, j) \in U$ of the dependency between labels Y_i and Y_j into L from unselected connections in U, where l does not induce a cycle with existing connections in L, until there are no more important connections in U. The connection l is selected by

$$\arg\max_{l \in U} \delta(l)$$

and we filter out l with the extreme-low-importance if $\delta(l) \leq \rho\tau$, where $\rho\tau$ is the threshold of importance value, τ is the mean of importance values of all connections, and $\delta(l)$ is the importance of $l = (i, j)$ between labels Y_i and Y_j that we measure by summing the top-k values of

$$\sum_{f \in F_\beta} w_f f(x_i, y_i) \sum_{f \in F_\beta} w_f f(x_j, y_j) \sum_{f \in F_\gamma} w_f f(y_i, y_j),$$

where $\forall y_i \in Y_i$, $\forall y_j \in Y_j$, and each weight w_f is acquired by training. Note that the importance $\delta(l)$ can be measured by different ways. With importance values of connections, we modify Eq. (4), the probability $p(\mathbf{y}|\mathbf{x})$ of a candidate of referred entities, as

$$\frac{1}{z(\mathbf{x})} \exp\Big(\sum_{i=1}^{n} [\sum_{f \in F_\alpha} w_f f(y_i) + \sum_{f \in F_\beta} w_f f(x_i, y_i)] + \sum_{l=(i,j) \in L} \sum_{f \in F_\gamma} \delta'(l) w_f f(y_i, y_j) \Big), \tag{6}$$

where $\delta'(l) = \frac{\delta(l)}{\sum_{l' \in L} \delta(l')}$.

7. Experimental Results

We investigated direct and indirect reference cases in real-world microblog data. We described the baseline method consisting of the existing features and model. We discussed the experimental results of candidate entity generation and entity identification and gave a case study. Table 1 shows the summary of experimental data.

7.1 Microblog Annotation

We investigated 461 tweets (which were microblogs) from Twitter, which were randomly collected using five Twitter-tags, as shown in Table 2, during October 22–27, 2014. Then, three people annotated these tweets according to the following three rules. Note that one of them was an author of this paper.

- A mention is a noun-phrase because words of other parts of speech are difficult to refer to entities. *E.g.*, the verb "run" is obviously not used to refer to entities.
- Entities that are specific should be targeted. *E.g.*, entity e ="Jacoby Ellsbury" is specific and entity e_2 ="Outfielder" is general. The reason is that if an entity (*e.g.*, e_2) is not specific, there exists other entities replaceable for e_2.
- Annotations should not be nested. *E.g.*, if the text "United State of America" in a microblog is annotated

Table 1 Data summary.

Data	#
Tweets	461
Mentions	1202
Direct references	850
Indirect references	570
Writers	354
News tagged in tweets	375
Twitter-users tagged in tweets	210
Articles (entities) in Wikipedia (version 2014-12-08)	4,732,221

Table 2 Annotation results on 461 randomly collected tweets.

Twitter-tag			#	
	tweets	mentions	direct ref.	indirect ref.
#Yankees	86	228	153	108
#Obama	92	227	167	87
#Ebola	97	241	151	156
#Nobel	94	287	228	124
#Islam	92	219	151	95
Mean per tweet		2.61	1.84	1.24

Table 3 Parameters used in experiments.

Parameter	Value
# of top candidate entities	1000
# of keywords K_e	0.1k
# of top-k words P_b	0.1k
# of top-k words I_b	1k
# of top-k words Δ	50
# of patterns	10
# of past microblogs B_u per writer	0.1k
Size of words-window for context of entities	0.1k
Threshold of semantic similarity θ	0.7, range [0,1]
ρ for threshold of connection's importance	0.001

as a mention to refer to a certain entity e, then "America" in this text cannot be annotated to refer to the same e.

Because the boundary of a mention text in a short-and-noisy tweet was not clear and sometimes the referred entities for the same mention from different annotators were similar in their Wikipedia articles, *e.g.*, articles of "Ebola virus" and "Ebola virus disease" are similar, we relaxed annotation results in the same tweets annotated by different annotators as follows. In a tweet, given one annotation in which entity e_1 is referred to by mention m_1 from annotator 1 and another annotation in which entity e_2 is referred to by mention m_2 from annotator 2, if m_1 and m_2 overlap at least one word in the tweet and e_1 and e_2 are similar in their articles, we made a new annotation such that its mention, which is the union of m_1 and m_2, refers to both e_1 and e_2. As shown in Table 2, the annotation results showed that there were multiple mentions in a microblog (2.61 per tweet) and the number of indirect references was not small (1.24 per tweet). This result implied that acquiring dependency among multiple referred entities and applying features for indirect references were required.

7.2 Baseline Method

We used the existing features and model for comparison. *Baseline-feature*: previous studies [7], [12], [17] focused on the following features. **Mention entity's name**: the probability that mention m refers to the name n_e of entity e by counting the number of anchor-link $a(m, n_e)$ where $a(m, n_e)$ links to the article of e by the text m in an article in Wikipedia, *i.e.*, $p(n_e|m) = \frac{\#a(m,n_e)}{\sum_{e'\in\Gamma}\#a(m,n_{e'})}$ where $\#a(m, n_e)$ is the number of $a(m, n_e)$ in Wikipedia and Γ is a set of entities linked by the text m. **Context similarity**: the similarity between the context C_e of e and the content W_b of microblog b, *i.e.*, $sim(C_e, W_b)$. We also measured the context of e and neighboring entity e_2 in b using $sim(C_{ee_2}, W_b)$. **Entities' correlation**: the correlation between two referred entities e and e_2 in b by measuring the similarity between the set V_e of e and that V_{e_2} of e_2, which is $\frac{|V_e \cap V_{e_2}|}{|V_e \cup V_{e_2}|}$, where V_e is a set of entities, for each of which there exists an anchor-link linking to e in its article or linked by in e's article. **User interest**: to evaluate user interest by computing the possibility that the name n_e of e appears on past microblogs B_u of writer u of b, *i.e.*, $sim(\{n_e\}, B_u)$. **Baseline-model**: previous studies [2], [12] ranked candidate entities on a single mention basis as follows. For each mention $m_i \in M$ in a microblog, to select the best candidate entity $e' = \arg\max_{e_i \in D_{m_i}} p(e_i|e_1, \ldots, e_{i-1}, e_{i+1}, \ldots, e_n, x)$, where D_{m_i} is a set of candidate entities of m_i and $e_1, \ldots, e_{i-1}, e_{i+1}, \ldots, e_n$ are currently predicted entities of other mentions. For a fair comparison with our results, the predicted entity of each mention was determined by repeating the above ranking until these predicted entities were not changed or reached 10,000 rounds.

Experiment setting: we list the parameters used in experiments in Table 3. The performance was measured using the mean reciprocal rank (*MRR*), which is

$$MRR = \frac{1}{|T|}\sum_{i=1}^{|T|} 1/rank_i,$$

where T is a set of tests and $rank_i$ is the ranked position of the ground-truth entity compared with other candidates at a single mention of a microblog by score at test i. Note that if the set of extracted candidate entities for a mention did not contain the ground-truth one, the corresponding $1/rank_i$ was set as zero. We ran 4-fold cross-validation for the given data. We trained a CRF model by using Structured SVM (SSVM) with the objective function of Hinge loss [24], [25], *cf.* Appendix A. Also, because our CRF model did not contain cyclic-dependency, we used the Belief Propagation algorithm [26] to speed up the training process. Because the number of combinations of candidate entities is exponential to the number of mentions of a microblog, we used Gibbs sampling to select quality candidates (*i.e.*, candidates of high score) to efficiently obtain the ranked position of ground-truth entities, *cf.* Appendix B. Note that we list all

Table 4 A list of features.

Feature	Description	Input	Output
Keyword	The similarity between the top words K_e in entity e's article and mention m.	K_e, m	$sim(K_e, \{m\})$
Occurrence frequency	The frequency that the name n_e of e and m co-occur in documents.	n_e, m	$log\ r(n_e, m)$
Topic-specific translation	The similarity between the context C_e of e and the translated content T_b of microblog b.	C_e, T_b	$sim(C_e, T_b)$
Entity-known-as pattern	The frequency that n_e, pattern p, m co-occur in order in a sentence.	n_e, p, m	$log\ l(``n_e \ldots p \ldots m")$
Entity's action	The probability that e, as a subject, co-occurs with verb v_s in a sentence.	e, v_s	$p(e \vert v_s)$
Mention entity's name	The probability that m refers to the name n_e in Wikipedia.	n_e, m	$p(n_e \vert m)$
Context similarity	The similarity between the context C_e and the content W_b of b.	C_e, W_b	$sim(C_e, W_b)$
Entities' correlation	The similarity between the set of entities V_e linking with e and that set V_{e_2} of entity e_2.	V_e, V_{e_2}	$\frac{\vert V_e \cap V_{e_2} \vert}{\vert V_e \cup V_{e_2} \vert}$
User interest	The similarity between the name n_e and the past microblogs B_u of writer u.	n_e, B_u	$sim(\{n_e\}, B_u)$

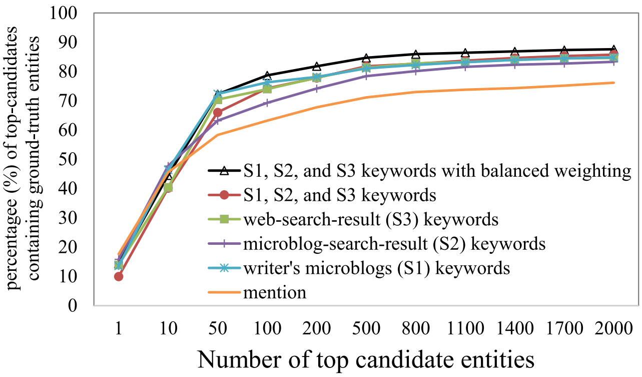


Fig. 2 Quality of candidate entity generation.

features including ours and existing ones in Table 4 and denote our proposed *CRF* model for entity identification as *CRF*, which is Eq. (5).

7.3 Performance of Candidate Entity Generation

We evaluated the candidate entity generation by examining whether ground-truth entities are included in generated candidates. Figure 2 shows that with the search query Q of balanced weighting words from different domains, we could boost the performance (2.76% gain when using top-1000 candidates) compared with Q of equally weighing words. In addition, the experimental result showed that extracting Q from different domains (*i.e.*, writer's or users' microblogs or Web) was effective to search for correct entities (3.27%, 4.87%, 3.20%, and 12.65% gain compared with S_1, S_2, S_3, and mention m, respectively, when using top-1000 candidates) compared with a single domain or without searching (*i.e.*, using only mention m). For the experiments of entity identification in Sect. 7.4, we used 1000 candidates for a direct/indirect reference, by which 86.4% of direct/indirect reference cases contained ground-truth entities in the top-1000 candidates.

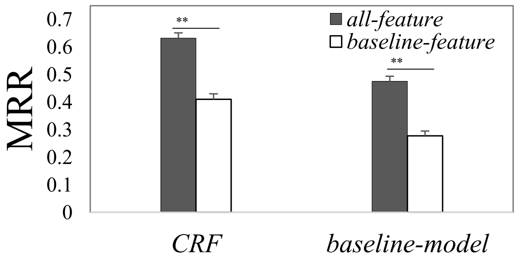


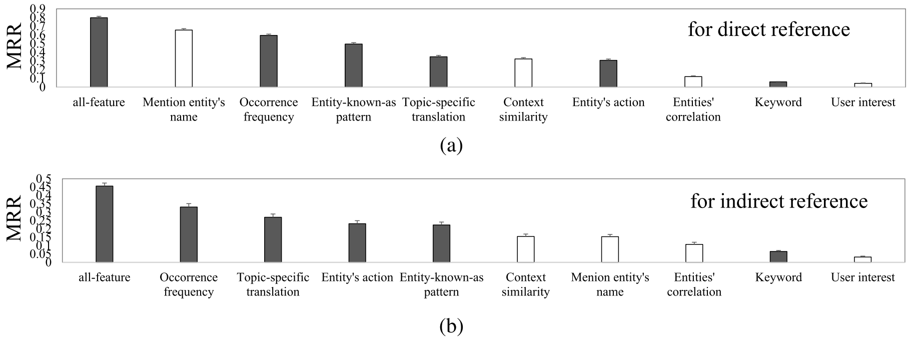
Fig. 3 Overall performance: comparison on features and models (+SEM).

7.4 Performance of Entity Identification

Overall performance. As shown in Fig. 3, our *CRF* model using all features achieved the performance of MRR 2.28 times[†] compared with the baseline model using the baseline features. With *CRF* model, using all features boosted performance by 1.54 times compared with the baseline features. On using all features, applying *CRF* model boosted performance by 1.33 times compared with applying the baseline model. The results suggest that not only our CRF model with adaptive dependency but also integrating existing features with our proposed features for direct and indirect references were important for the *EI-DIR* problem. We investigated the effect of ours/the baseline features and *CRF*/the baseline model in detail.

Feature comparison. We showed the effect of features on direct/indirect references in Figs. 4 (a)–(b). Our proposed features performed well on both direct/indirect references, *cf.* three and four of the top-five single features in direct and indirect references, respectively, were ours. For indirect references, multiple features were more effective than

[†] We showed that the experimental results were trustable by applying the t-Test for assessing whether the means between two groups of values were significantly different. We also applied the Bonferroni correction for multiple comparisons. Note that $*$ and $**$ denote $p < .05$ and $p < .01$, respectively.



(a)

(b)

Fig. 4 The effect of features on (a) direct and (b) indirect references (+SEM).

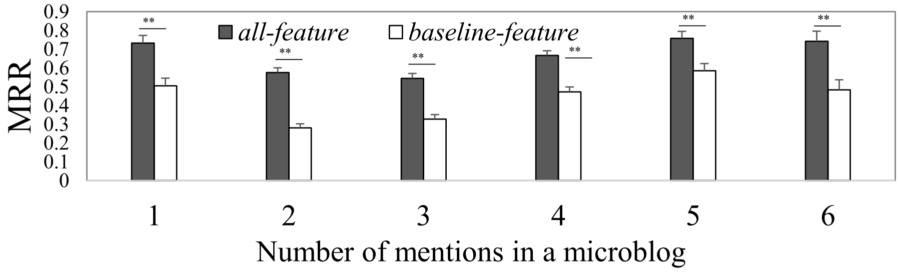


Fig. 5 Feature comparison (+SEM).

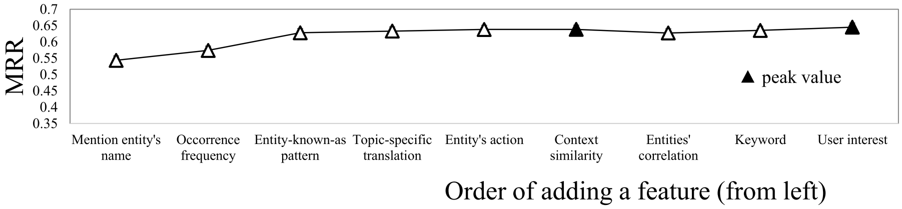


Fig. 6 Performance of best feature combination (+SEM).

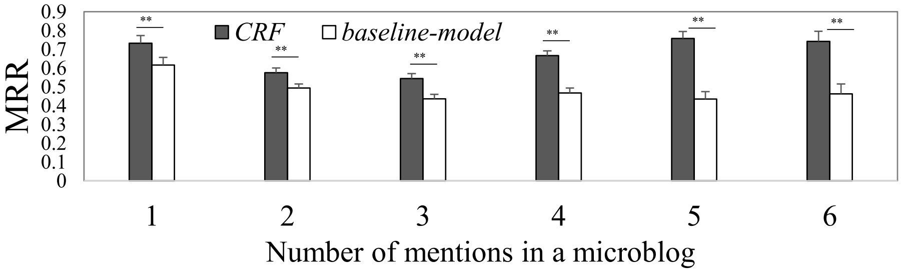


Fig. 7 Model comparison (+SEM).

using a single feature as shown in Fig. 4 (b). The reason might be that the referred entities induced from indirect references were of different types and from different topics; thus we required more features to evaluate them. For direct references, we observed that the existing feature of mentioning entity's name was important. Furthermore, we compared the effect of all features with that of the baseline features as shown in Fig. 5. All features including ours were robust when there were multiple mentions in a microblog. The reason might be that some of our features could obtain quality measurement for multiple referred entities and mentions together, *e.g.*, topic-specific translation and entity-known-as pattern. To identify which features are important, we step-by-step added one important feature from currently unselected features into our CRF model as shown in Fig. 6. We found that for enhancing performance, multiple features, which were effective for direct/indirect references, were required collectively, *cf.* requiring the top-six features to reach the first peak MRR including mention entity's name and our entity-known-as pattern and topic-specific translation. **Model comparison**. As shown in Fig. 7, we compared our CRF model with the baseline model. The result showed

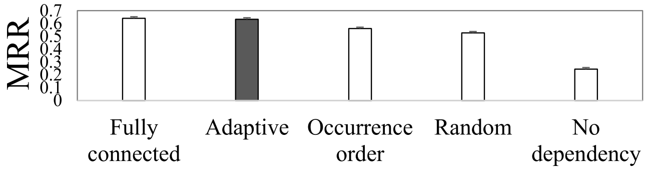


Fig. 8 Performance of CRF models with varying dependency (+SEM).

Table 5 Case study. Note that ground-truth entities are underlined.

Mention	**Candidate entities**	**Difficulty**
"our Suzuki"	"Jacoby Ellsbury", "Ichiro Suzuki", "Suzuki Kei"	Metaphor-like reference
"the prospect"	"Luis Severino", "Karl-Anthony Towns"	Abstract term
"pitcher of Yankees"	"Masahiro Tanaka", "New York Yankees", "Derek Jeter"	Ambiguous between entity's name and attribute

that our CRF model was effective when there were multiple mentions, which suggest that our CRF model with adaptive dependency could acquire proper dependency among multiple entities for correctly inferring multiple entities together, compared with the baseline model.

Performance of CRF model with adaptive dependency. We further compared the CRF model of our proposed adaptive dependency with other dependency including connecting each pair of referred entities (fully connected), according to the occurrence order of mentions in the microblog (occurrence order), randomly selecting dependency, and no dependency as shown in Fig. 8. The results suggest that our CRF model with adaptive dependency was slightly worse than that with the fully connected dependency, *cf.* the MRR difference was 0.0083; however, a CRF model with the fully connected dependency induced the exponential time-complexity for predicting entities. The dependency of occurrence order provided quality results, which implied that people write the related content together in a microblog and this characteristic is useful to model dependency when we have little information about the dependency.

7.5 Discussion

We investigated difficult cases and listed possible difficulty in Table 5. One case was a metaphor-like reference, where a mention was obviously the name of a certain entity but was used to refer to another entity. The second case was that a mention was abstract words and thus it could generate candidate entities from a variety of topics. The last case was an ambiguous condition between entity's name and attribute. *E.g.*, in the mention "pitcher of Yankees", "pitcher" is a normal noun and it could target candidate entities of indirect references; however, candidates targeted by "Yankees", which is a proper noun, may come from direct references. Furthermore, we discuss that how to detect a noun-phrase n in a microblog that indirectly refers to a certain entity e as follows. The features of topic may be useful for detection. For example, n indirectly refers to e if n is an attribute/relation of e or the sentence containing n shows a status or action about e. Furthermore, if we have a data set with labels indicating whether a noun-phrase in a microblog is an indirect reference or not, we can learn a detection rule by machine learning techniques (*e.g.*, SVM).

8. Conclusion and Future Work

We addressed the problem of entity identification on a microblog and focused on the cases of indirect references, in which entities are not referred to by their names. Because a mention of an indirect reference may consist of abstract nouns and it makes estimating referred entities difficult, we translated a microblog by related topics and used the translated results to measure referred entities. We also used syntactic or linguistic features. To identify referred entities in a microblog together with adaptive correlation, we applied the CRF model with dynamically building dependency among them. We surveyed 461 random tweets and found that the number of indirect references was not small. The experimental results suggest that our features were effective for indirect references and the CRF model with adaptive dependency was robust when there were multiple mentions in a microblog and it achieved the same high performance as the CRF model with the fully connected dependency. In future work, the systematic analysis for indirect reference cases (*e.g.*, clustering of indirect reference cases) is required, which can be beneficial to design the features for indirect references. The on-line application may need to automatically detect which words in a microblog are a mention of a direct or indirect reference and to identify the referred entities with the detected mentions together.

Acknowledgments

This work was supported in part by Grants-in-Aid for Scientific Research (Nos. 25240050, 15H01718, 26700009, and 16K16156) from MEXT of Japan.

References

[1] W. Shen, J. Wang, and J. Han, "Entity linking with a knowledge base: Issues, techniques, and solutions," IEEE Trans. Knowl. Data Eng., vol.27, no.2, pp.443–460, 2015.

[2] W. Shen, J. Han, and J. Wang, "A probabilistic model for linking named entities in web text with heterogeneous information networks," in SIGMOD, pp.1199–1210, 2014.

[3] C. Sutton and A. McCallum, "An introduction to conditional random fields," Foundations and Trends in Machine Learning, vol.4, no.4, pp.267–373, 2012.

[4] A. McCallum, "Efficiently inducing features of conditional random fields," CoRR, vol.abs/1212.2504, 2012.

[5] L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, J. Petrak, and K. Bontcheva, "Analysis of named entity recognition and linking for tweets," Inf. Process. Manage., vol.51, no.2, pp.32–49, 2015.

[6] C. Li, A. Sun, J. Weng, and Q. He, "Tweet segmentation and its application to named entity recognition," IEEE Trans. Knowl. Data Eng., vol.27, no.2, pp.558–570, 2015.

[7] S. Guo, M. Chang, and E. Kiciman, "To link or not to link? A study on end-to-end tweet entity linking," NAACL HLT, pp.1020–1030, 2013.

[8] X. Liu, F. Wei, S. Zhang, and M. Zhou, "Named entity recognition for tweets," ACM TIST, vol.4, no.1, 2013.

[9] X. Liu and M. Zhou, "Two-stage NER for tweets with clustering," Inf. Process. Manage., vol.49, no.1, pp.264–273, 2013.

[10] J. Hoffart, Y. Altun, and G. Weikum, "Discovering emerging entities with ambiguous names," in WWW, pp.385–396, 2014.

[11] Z. Guo and D. Barbosa, "Robust entity linking via random walks," in CIKM, pp.499–508, 2014.

[12] W. Shen, J. Wang, P. Luo, and M. Wang, "Linking named entities in tweets with knowledge base via user interest modeling," in KDD, pp.68–76, 2013.

[13] Y. Li, C. Wang, F. Han, J. Han, D. Roth, and X. Yan, "Mining evidences for named entity disambiguation," in KDD, pp.1070–1078, 2013.

[14] R. Blanco, G. Ottaviano, and E. Meij, "Fast and space-efficient entity linking for queries," in WSDM, pp.179–188, 2015.

[15] C. Wang, K. Chakrabarti, T. Cheng, and S. Chaudhuri, "Targeted disambiguation of ad-hoc, homogeneous sets of named entities," in WWW, pp.719–728, 2012.

[16] E. Meij, W. Weerkamp, and M. de Rijke, "Adding semantics to microblog posts," in WSDM, pp.563–572, 2012.

[17] B. Skaggs and L. Getoor, "Topic modeling for wikipedia link disambiguation," ACM Trans. Inf. Syst., vol.32, no.3, 2014.

[18] X. Han and L. Sun, "An entity-topic model for entity linking," in EMNLP-CoNLL, pp.105–115, 2012.

[19] D. Milne and I.H. Witten, "An effective, low-cost measure of semantic relatedness obtained from wikipedia links," AAAI, pp.25–30, 2008.

[20] M.P. Kato, H. Ohshima, S. Oyama, and K. Tanaka, "Query by analogical example: relational search using web search engine indices," in CIKM, pp.27–36, 2009.

[21] E. Shutova, S. Teufel, and A. Korhonen, "Statistical metaphor processing," Computational Linguistics, vol.39, no.2, pp.301–353, 2013.

[22] E. Shutova, B.J. Devereux, and A. Korhonen, "Conceptual metaphor theory meets the data: a corpus-based human annotation study," Language Resources and Evaluation, vol.47, no.4, pp.1261–1284, 2013.

[23] T. Pedersen, S. Patwardhan, and J. Michelizzi, "Wordnet::Similarity: measuring the relatedness of concepts," in AAAI, pp.38–41, 2004.

[24] T. Finley and T. Joachims, "Training structural svms when exact inference is intractable," in ICML, pp.304–311, 2008.

[25] T. Joachims, T. Finley, and C.-N.J. Yu, "Cutting-plane training of structural svms," Machine Learning, vol.77, no.1, pp.27–59, 2009.

[26] E.B. Sudderth, A.T. Ihler, M. Isard, W.T. Freeman, and A.S. Willsky, "Nonparametric belief propagation," Commun. ACM, vol.53, no.10, pp.95–103, 2010.

[27] A. Vedaldi, "A MATLAB wrapper of SVM$^{\text{struct}}$," http://www.vlfeat.org/~vedaldi/code/svm-struct-matlab.html, 2011.

[28] S. Chatterjee and S.J. Russell, "A temporally abstracted viterbi algorithm," CoRR, vol.abs/1202.3707, 2012.

Appendix A: Training a CRF Model by Structured SVM

Given a CRF model (X, Y, L), we learned the optimal weight $\mathbf{w}' \in \mathbb{R}^m$ of features in Eq. (4) based on the objective function of Hinge loss [25], *i.e.*,

$$\mathbf{w}' = \arg\min_{\mathbf{w}\in\mathbb{R}^m} \Bigg(\epsilon\|\mathbf{w}\| + \qquad \text{(A· 1a)}$$

$$\sum_{i=1}^{\nu} \max_{\mathbf{y}\in Y} \left(l(\mathbf{y}) + \mathbf{w}^T \phi(\mathbf{x}^i, \mathbf{y}) \right) - \mathbf{w}^T \phi(\mathbf{x}^i, \mathbf{y}^i) \Bigg), \qquad \text{(A· 1b)}$$

where m is the number of features, ν is the number of training instances of microblogs, $\mathbf{y} = (y_1, \ldots, y_n) \in Y$ is a candidate of referred entities, $\mathbf{x}^i = (x_0^i, x_1^i, \ldots, x_n^i) \in X$ is the i-th microblog with mentions, $\mathbf{y}^i = (y_1^i, \ldots, y_n^i)$ is the ground-truth of referred entities for the i-th microblog, $l(\mathbf{y})$ is the difference between $\mathbf{y}$ and $\mathbf{y}^i$ which is $l(\mathbf{y}) = \sum_{j=1}^{n} |\{1|y_j = y_j^i\}|$, and $\phi(\mathbf{x}^i, \mathbf{y})$ is a vector of feature values of entities $\mathbf{y}$ given $\mathbf{x}^i$. Because there were multiple outputs (*i.e.*, entities) in a $\mathbf{y}$ and there existed correlation between outputs, we used Structured SVM (SSVM) to obtain the optimal weight $\mathbf{w}'$ by using the tool [27]. Because we used the CRF model with non-cycle dependency, we used the Belief Propagation algorithm to efficiently get the maximum value in Eq. (A· 1b).

Appendix B: Getting Quality Candidates by Gibbs Sampling

As the number of candidate entities for entity identification could be large, we efficiently picked the top-k candidates C with high score by Gibbs sampling. For the first round of Gibbs sampling, we picked candidate $\mathbf{y}^{(1)}$ s.t. $\mathbf{y}^{(1)} = \arg\max_{\mathbf{y}\in Y} \mathbf{w}^T \phi(\mathbf{x}, \mathbf{y})$, which had the maximum score based on feature values using the Belief Propagation algorithm. Given candidate $\mathbf{y}^{(i)}$ sampled at the current round i, we sampled candidate $\mathbf{y}^{(i+1)} = (y_1^{(i+1)}, \ldots, y_n^{(i+1)})$ at the next round $i + 1$ as follows:

$$\begin{aligned} &\textit{sample } y_1^{(i+1)} \sim p(y_1 | y_2^{(i)}, \ldots, y_n^{(i)}), \\ &\qquad\qquad \ldots \\ &\textit{sample } y_n^{(i+1)} \sim p(y_n | y_1^{(i+1)}, \ldots, y_{n-1}^{(i+1)}). \end{aligned}$$

Appendix C: Proof of Time Complexity of Inference on a Non-Cycle CRF Model

Given a CRF model, which does not contain cycle connections of dependency, comprises a set of trees S, where each pair of trees in S is not connected, the inference on each tree $s \in S$ by the Viterbi algorithm takes time $O(n_s c^2)$ [28] and then the overall time is $\sum_{s\in S} O(n_s c^2) = O(c^2) \sum_{s\in S} n_s = O(nc^2)$, where n_s is the number of mentions in s and $\sum_{s\in S} n_s = n$.

Jun-Li Lu received the M.S. degree in Department of Electrical Engineering from National Taiwan University, Taiwan, in 2010. From 2013, he became a Ph.D. student in Graduate School of Informatics, Kyoto University, Japan. His research interests include text mining and social-network mining.

Makoto P. Kato received the B.S., M.S., and Ph.D. degrees from Kyoto University, Japan, in 2008, 2009, and 2012, respectively. He is currently an assistant professor at Kyoto University. He serves as a program co-chair of NTCIR from 2015. He is also a program committee member of the SIGIR and WSDM conferences. His research interests include interactive information retrieval, user behavioral analysis in search, and search intent detection.

Takehiro Yamamoto received the M.S. and Ph.D. degrees in Informatics from Kyoto University, Japan, in 2007 and 2011, respectively. He has worked as a research fellow at Kyoto University from 2012 to 2013. Since 2014 he has been an assistant professor at Kyoto University. His research interests include information retrieval, human-computer interaction, and web mining.

Katsumi Tanaka received the B.S., M.S., and Ph.D. degrees in Information Science from Kyoto University, Japan, in 1974, 1976 and 1981, respectively. In 1986, he joined Department of Instrumentation Engineering, Faculty of Engineering at Kobe University, Japan, as an associate professor. In 1994, he became a full professor in Faculty of Engineering, Kobe University. Since 2001, he has been a professor of Graduate School of Informatics, Kyoto University. His research interests include database theory and systems, web information retrieval, and multimedia retrieval.