Abstract

Automatic identification of mentioned entities in social media posts facilitates quick digestion of trending topics and popular opinions. Nonetheless, this remains a challenging task due to limited context and diverse name variations. In this paper, we study the problem of entity linking for Chinese news comments given mentions’ spans. We hypothesize that comments often refer to entities in the corresponding news article, as well as topics involving the entities. We therefore propose a novel model, XREF, that leverages attention mechanisms to (1) pinpoint relevant context within comments, and (2) detect supporting entities from the news article. To improve training, we make two contributions: (a) we propose a supervised attention loss in addition to the standard cross entropy, and (b) we develop a weakly supervised training scheme to utilize the large-scale unlabeled corpus. Two new datasets in entertainment and product domains are collected and annotated for experiments. Our proposed method outperforms previous methods on both datasets.

1. Introduction

Social media, including online discussion forums and commenting systems, provide convenient platforms for the public to voice opinions and discuss trending events [O’Connor et al., 2010, Lau et al., 2012]. Entity linking (EL), which aims to identify the knowledge base entry (or the lack thereof) for a given mention’s span, has become an indispensable tool for consuming the enormous amount of social media posts. Concretely, automatically recognizing entities can quickly inform who and what are popular [O’Connor et al., 2010, Zhao et al., 2014, Dredze et al., 2016], promote semantic understanding of social media content, and facilitate downstream tasks, such as relation extraction, opinion mining, questions answering, and personalized recommendation [Messenger and Whittle, 2011, Galli et al., 2015]. Although EL has been extensively investigated in newswire [Kazama and Torisawa, 2007, Ratinov et al., 2011], web pages [Demartini et al., 2012], and broadcast news [Benton and
Dredze, 2015], its study in the social media domain was only started more recently [Guo et al., 2013a, Yang and Chang, 2015, Moon et al., 2018], mostly focusing on English.

In this paper, we study the task of entity linking for user comments in online Chinese news portals. To the best of our knowledge, we are the first to investigate EL problem for the genre of news comments at a large scale. Besides issues present in the conventional EL work [Ji et al., 2010], social media text poses additional challenges: the lack of context and increased name variations due to its informal style. State-of-the-art EL methods [Francis-Landau et al., 2016, Gupta et al., 2017] heavily rely on modeling the text surrounding the mentions, as the abundant context from longer documents greatly helps identify entity related content. However, context is often scant for user comments. For instance, as shown in Figure 1, the entity mention \( m_2 \) may indicate “Bingbing Fan” or “Bingbing Li”, both being prominent actresses. By looking up the entities covered in the article, which contains unambiguous mention of the former entity, an EL system will be more confident to link \( m_2 \) to it. Moreover, the informal and evolving vocabulary on social media lead to enormous name variations based on aliases, morphing, and misspelling. For instance, in one of our newly annotated datasets, the maximum number of distinct mentions of an entity is 121.

In this work, we propose XREF, a novel entity linking model for Chinese news comments by exploiting context information of entity mentions as well as identifying relevant entities in reference articles. XREF, with its overview displayed in Figure 2, has three key properties. First, we enrich the mention representation with two sources of information through attentions. **Comment attention** pinpoints topics involving the target entity from comment context. For instance, words “star” and “actress” in comment C1 in Figure 1 provide useful information about entity types. **Article entity attention** detects target entities from the articles if they are discussed. Furthermore, we investigate a new objective function to drive the learning of article entity attention. Finally, we also exploit data augmentation with distant supervision [Mintz et al., 2009] to leverage large amounts of unlabeled comments and articles for model training.

Since there was no publicly-available annotated dataset, as part of this study, we collect and label two new datasets of Chinese news comments from the domain of entertainment and product, which are crawled from a popular Chinese news portal toutiao.com.1 Experimental results show that our best performing model obtains significantly better accuracy.

---

Bingbing Fan did not win the best actress in the Golden Horse Awards, she was attacked by Xiaogang Feng afterwards.

Yesterday, Dongyu Zhou and Sichun Ma won the best actress in the Golden Horse Awards for their performance in the movie "Soul Mate".

Later, the camera moved towards Bingbing Fan... She is just a star...
Table 1: Statistics of crawled datasets from entertainment and product domains.

<table>
<thead>
<tr>
<th></th>
<th>Entertainment</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td># News Articles</td>
<td>10,845</td>
<td>8,275</td>
</tr>
<tr>
<td>Avg # Sents per Article</td>
<td>18.5</td>
<td>13.5</td>
</tr>
<tr>
<td>Avg # Chars per Sentence</td>
<td>40.4</td>
<td>50.7</td>
</tr>
<tr>
<td># Comments</td>
<td>967,763</td>
<td>410,790</td>
</tr>
<tr>
<td>Avg # Chars per Comment</td>
<td>21.3</td>
<td>23.2</td>
</tr>
<tr>
<td># Annotated Comments</td>
<td>30,630</td>
<td>5,189</td>
</tr>
<tr>
<td># Annotated Mentions</td>
<td>46,942</td>
<td>7,497</td>
</tr>
<tr>
<td># Annotated Unique Entities</td>
<td>1,846</td>
<td>470</td>
</tr>
</tbody>
</table>

similar content [Huang et al., 2014]. However, users in news commenting systems might be anonymous, and few additional posts would be available for newly published articles. We therefore study a more practical setup, without using any of the aforementioned information as input.

3. Data Collection and Annotation

We collect user comments along with corresponding news articles from toutiao.com, a popular Chinese online news portal. A sample article snippet with comments is displayed in Figure 1. Two popular domains are selected for annotation: entertainment (Ent) and product (Prod). Articles and comments in Ent focus on movies, TV shows, and celebrities, whereas most topics in Prod are automobiles and electronic products. The statistics of the crawled dataset after filtering are in Table 1. As illustrated, there are only an average of 20 characters in a comment, highlighting the lack of context.

Annotation Procedure. We randomly sample 995 articles from Ent and 783 articles from Prod, and annotate the corresponding user comments. Articles and comments that are not in the samples are used for model pre-training via data augmentation (§ 4.6).

Annotators are presented with both comments and corresponding articles during the annotation process. They first identify mention spans, where named, nominal, and pronominal mentions of entities are labeled. Each mention is then linked to an entity in a knowledge base, or labeled as NIL if no entry is found. Though not in our knowledge base, the word “小編” (editor) is included as an entity due to its popularity. We also allow one mention to be linked to multiple entities, e.g. plural pronoun “他們” (they/them). Comments without any mention are discarded. 13 professional annotators, who are native Chinese speakers with extensive NLP annotation experience, are hired, each annotating a different subset. An additional human annotator conducts the final check.

Statistics. Final statistics for the datasets are displayed in Table 1. On average, there are 4.4 distinct mentions per entity, with a maximum number of 121 for domain Ent. For Prod, the average mention number is 2.9 with a maximum number of 48. Sample mentions are shown in Table 2.

We categorize the samples into the following types, based on the entities mentioned by: (1) canonical names as defined in knowledge base; (2) nicknames as the popular aliases included in the knowledge base for each entity; (3) pronominal mentions indicating one
Table 2: Entities with the most unique mentions from entertainment and product domains.

<table>
<thead>
<tr>
<th>Entity (uniq. mentions)</th>
<th>Sample Mentions</th>
</tr>
</thead>
<tbody>
<tr>
<td>范冰冰 (121)</td>
<td>“戏子(actress)”, “姐姐 (sister Bing)”, “范 (Fan)”, “国际女神 (international goddess)”</td>
</tr>
<tr>
<td>那英 (111)</td>
<td>“戏子 (actress)”, “满族后裔 (descendant of Manchu people)”, “自个 (herself)”, “演员 (actress)”</td>
</tr>
<tr>
<td>别克英朗 (48)</td>
<td>“手动精英 (stick shift elite)”, “这款车 (this car)”, “我的车子 (my car)”, “2016款英朗 (2016 Excelle)”</td>
</tr>
<tr>
<td>马自达3昂克赛拉 (47)</td>
<td>“两厢 (hatchback)”, “自动舒适型 (automatic and comfortable)”, “昂克赛拉 (Axela)”, “昂克赛拉1.5自动舒适型 (Axela 1.5T automatic)”</td>
</tr>
</tbody>
</table>

Table 3: Mention type distribution.

<table>
<thead>
<tr>
<th></th>
<th>Canon.</th>
<th>Nick.</th>
<th>Pron.</th>
<th>Others</th>
<th>Plural</th>
<th>NIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entertainment</td>
<td>29.8%</td>
<td>4.0%</td>
<td>12.9%</td>
<td>21.9%</td>
<td>2.9%</td>
<td>28.4%</td>
</tr>
<tr>
<td>Product</td>
<td>33.9%</td>
<td>0.6%</td>
<td>2.7%</td>
<td>41.1%</td>
<td>0.2%</td>
<td>21.6%</td>
</tr>
</tbody>
</table>

Knowledge Base. Baidu Baike\(^2\), a large-scale Chinese online encyclopedia, is used to construct the knowledge base (KB). A snapshot of Baike containing 68,067 unique entities was collected on May 10th, 2017. Four attributes are leveraged for feature engineering: (1) **gender**, (2) **nicknames** as a list of common aliases for an entity, (3) **entity type**, and (4) **entity relation**.

4. The Proposed Approach

Our model takes as input a phrasal mention \(m\) in a comment \(C\), which is posted under an article. Given a knowledge base, we aim to predict the KB entity that \(m\) refers to, or to label it as NIL if no such entity exists. Concretely, a list of candidate entities will be first selected as \(E_m = \{e\}\) based on string matching and knowledge graph expansion (see § 4.1). Then a linking probability will be computed over each candidate given \(m\).

4.1 Candidate Construction

Our candidate construction algorithm consists of two steps. For each mention, we consider all entities that appear in the same comment and corresponding article by matching their canonical names. This forms the initial candidate list. In the second step, a new entity

\(^2\) https://baike.baidu.com
is selected if it has a relation with any entity in the initial list according to our KB. The initial list, the expanded entities, and NIL comprise the final candidate set.

Following this procedure, 96% gold-standard entities are retrieved in the candidate sets for ENT, and 62% are covered for PROD. To improve the coverage for the PROD domain, we collect unambiguous aliases (no other entity with the same alias) that are not pronominal mentions from training data for each entity, and use these as additional entity nicknames for candidate construction. The coverage is increased to 93%.

4.2 Entity Representation

Prior work for entity representation learning usually relies on entity-word co-occurrence statistics derived from the entities’ English Wikipedia pages [Francis-Landau et al., 2016, Gupta et al., 2017, Ganea and Hofmann, 2017, Eshel et al., 2017]. Unfortunately, Wikipedia has low coverage of entities in our newly collected Chinese datasets. We thus consider two sources of information, both acquired from news headlines. First, a graph-based node2vec [Grover and Leskovec, 2016] embedding \( u^\text{nod} \) is induced from an entity-entity co-occurrence matrix extracted from 65 million news titles after applying canonical name matching [Zwicklbauer et al., 2016, Yamada et al., 2016]. \( u^\text{nod} \) is expected to capture entity relations. Second, a Singular Vector Decomposition (SVD)-based representation \( u^\text{word} \) is obtained from an entity-word co-occurrence matrix constructed from the same set of news titles. We concatenate them as \( u \) and apply a one-layer feedforward neural network over it to form the entity representation \( v_e = \tanh(W_e u + b_e) \), where \( W_e \in \mathbb{R}^{300 \times 600} \) and \( b_e \in \mathbb{R}^{300 \times 1} \) are trainable parameters.

4.3 Mention Representation

We train character embeddings from the 327 million user comments with word2vec [Mikolov et al., 2013]. A bidirectional Long Short-Term Memory (biLSTM) network is then applied over comment character embeddings \( x_c^i \), with hidden state \( h_i = [h_i^-; h_i^+] \) for each time step \( i \). We append a one-bit mask \( q_i \) to the character embeddings to indicate the mention span. If a character is within the mention span, \( q_i = 1 \); otherwise, it is 0. \( h_i \) is calculated recurrently as \( h_i = g(h_{i-1}, [x_c^i; q_i]) \), where \( g \) is the 200-dimensional biLSTM network. The last hidden state \( h_T \) is taken as the base form of mention representation \( v_m^\text{base} \).

**Comment Attention.** Preliminary studies show that \( v_m^\text{base} \) focuses on the local context, and does not capture long-distance information well. Hence we propose to learn an importance distribution over all comment characters through a bilinear attention [Luong et al., 2015] with query \( \tilde{m} \), the average character embeddings of the mention:

\[
\tilde{m} = \frac{1}{m_e - m_s} \sum_{i=m_s}^{m_e} x_c^i
\]

\[
\alpha_{i}^{\text{cnt}} = \frac{\exp(h_i^T W_c \tilde{m})}{\sum_{i'=1}^{T} \exp(h_i'^T W_c \tilde{m})}
\]

\[
v_{m}^{\text{cnt}} = \sum_{i=1}^{T} \alpha_{i}^{\text{cnt}} h_i
\]
Article Entity Attention. Intuitively, users tend to comment on entities covered in the news. We thus design an article entity attention to identify target entities if they appear in the article, or indicate non-existence otherwise. Concretely, articles are segmented into words by Jieba, an open source Chinese word segmentation tool. Each word is matched with canonical entity names in the knowledge base, and the article is represented as a set of unambiguous entities, $E_a$. Each entity is represented as $u = [u_{nod}; u_{wrd}]$. We also add one absent padding entity (denoted as $ABS$), a 300-dimension zero vector, into the set to indicate that the entity is not in the article. The article entity representation $v_{art}$ is calculated as:

$$
\beta_{j}^{art} = \text{softmax}(u_j^T W_a \tilde{m})
$$

$$
v_{art} = \sum_{j=1}^{|E_a|} \beta_{j}^{art} u_j
$$

where $u_j$ is the entity representation for $j$-th entity in $E_a$. $W_a \in \mathbb{R}^{600 \times 300}$ is the bilinear matrix parameter.

4.4 Learning Objective

XREF learns to align the mention representation $v_m$ and the candidate entity representation $v_e$ after transforming them into a common semantic space. Specifically, the base form $v_m^{base}$, comment attended $v_m^{cmt}$, and article attended $v_m^{art}$ are concatenated as the input to a feedforward neural network to form $v_m = \tanh(W_m[v_m^{base}; v_m^{cmt}; v_m^{art}] + b_m)$. Given a mention $m$ represented as $v_m$, the probability for $m$ being linked to an entity $e$ (represented as $v_e$) is computed by applying the softmax function over the dot product between their representations, over all candidates in $E_m$: $P(e|m) = \text{softmax}_{e \in E_m}(v_e \cdot v_m)$ The entity with the highest positive likelihood is selected as prediction. Previous work [Yang et al., 2016] has found that surface features can further improve representation learning-based EL models. We thus append features (§ 4.5) to the dot product via $P(e|m) = \text{softmax}_{e \in E_m} \left( w \cdot [v_e \cdot v_m; \Phi(m)] \right)$, where $\Phi(m)$ is the feature vector and $w$ are learnable weights.

During training time, we use the same candidate construction algorithm in § 4.1 to collect negative samples, where all candidates except the gold-standard are treated as negative. The cross-entropy loss on training set is defined as:

$$
\mathcal{L}_{EL}(\theta) = - \sum_n \sum_k y_{n,k}^* \log(P(e_k|m_n))
$$

where $P(e_k|m_n)$ is the predicted probability for the $k$-th entity candidate for $n$-th mention in training set. $y_{n,k}^*$ represents the gold-standard, it has a value of 1.0 for positive samples, and 0.0 for negative ones.

3. https://github.com/fxsjy/jieba
Supervised Attention Loss. Notice that the article entity attention naturally learns an alignment between the mention and entity representation $u$. To help learn high quality alignment, we design a new learning objective to provide direct supervision to the article entity attention. To the best of our knowledge, we are the first to design supervised attention mechanism to guide entity linking. Concretely, during training, if an entity in $\mathcal{E}_a$ matches the gold-standard, we assign a relevance value of 1.0 to it; otherwise, the score is 0.0. If none from $\mathcal{E}_a$ matches, the absent padding entity is labeled as relevant. We thus design the following objective for article attention learning:

$$L_{Att}(\theta) = -\sum_{n} \sum_{j} \beta^*_n,j \log(\hat{\beta}_n,j)$$ (7)

$\beta^*_n,j$ is the true relevance value for $j$-th article entity, and $\hat{\beta}_n,j$ is the attention calculated as in Eq. 4, both are extended with mention index $n$ (i.e. the $n$-th mention in the training set). The final learning objective becomes $\mathcal{L}(\theta) = \mathcal{L}_{EL}(\theta) + \lambda \cdot L_{Att}(\theta)$. $\lambda$ is set to 0.1 in all experiments below.

4.5 Features

We optionally append 20 features to the output layer, as detailed in Table 4, where the last 11 features are adopted from [Zheng et al., 2010].

4.6 Weakly Supervised Pre-training

We leverage the unlabeled samples for data augmentation. Concretely, mentions and entities are automatically labeled if an entity’s canonical name or nickname can be matched in a comment unambiguously (i.e., no other entity with the same name). In total, this procedure automatically labeled 502,858 comments for the ENT domain, which is split into 453,080 for training and 49,778 for validation. For the PROD domain, we create 175,951 comments, among which 158,336 are for training and 17,615 are for validation. Each dataset is used to pre-train XREF, which is then trained on the annotated data.

5. Experimental Setup

Each dataset is split into training, validation, and test sets based on articles, with statistics displayed in Table 5. Articles in test sets are published later than those in training and validation sets. For this study, we focus on the task of entity linking, therefore gold-standard mention spans are assumed to have been provided. A mention detection component will be developed in future work.

Hyperparameters. For all experiments, Adam optimizer [Kingma and Ba, 2015] is used with an initial learning rate of 0.0001. We adopt gradient clipping with a maximum norm of 5. Model batch size is set to 128.

Baselines. We design five baselines: (1) MatchCanon matches the mention with canonical names in KB, and outputs an entity if a match is found, otherwise predicts NIL; (2) MatchCanonAndNick further matches nicknames if MatchCanon returns NIL, (3) FrequencyInArt predicts the most frequent entity in the article; (4) FirstInArt pre-
### Table 4: Features used in our model and comparisons.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CanonMatch</td>
<td>Whether the mention text exact-match the canonical KB name</td>
</tr>
<tr>
<td>NicknMatch</td>
<td>Whether the mention text exact-match the canonical KB name</td>
</tr>
<tr>
<td>CharJaccard</td>
<td>The char-level Jaccard score btw. the mention and entity’s canonical name</td>
</tr>
<tr>
<td>PinyJaccard</td>
<td>The Jaccard similarity between the Pinyin of the mention and entity’s canonical name</td>
</tr>
<tr>
<td>GendMatch</td>
<td>Whether the gender of pronominal mention matches that in KB</td>
</tr>
<tr>
<td>EntArtFreq</td>
<td>The frequency of candidate entity in article, considering both exact canonical name searching and nickname searching</td>
</tr>
<tr>
<td>CommentDist</td>
<td>The distance between the canonical name of the candidate entity and the mention in the comment, if the canonical name is not present set to 100</td>
</tr>
<tr>
<td>PriorProb</td>
<td>Probability $P(e</td>
</tr>
<tr>
<td>Special</td>
<td>Whether the mention is a domain-specific entity, such as “小编” (editor)</td>
</tr>
<tr>
<td>EditDist</td>
<td>The edit distance between mention and entity on character level</td>
</tr>
<tr>
<td>StartWithMent</td>
<td>Whether any of the entity’s canonical name or nickname starts with the mention string</td>
</tr>
<tr>
<td>EndWithMent</td>
<td>Whether any of the entity’s canonical name or nickname ends with the mention string</td>
</tr>
<tr>
<td>StartInMent</td>
<td>Whether any of the entity’s canonical name or nickname is a prefix of the mention string</td>
</tr>
<tr>
<td>EndInMent</td>
<td>Whether any of the entity’s canonical name or nickname is an affix of the mention string</td>
</tr>
<tr>
<td>EqualWordCnt</td>
<td>The maximum number of same words between mention and entity’s canonical name and nicknames</td>
</tr>
<tr>
<td>MissWordCnt</td>
<td>The minimum number of different words between mention and entity’s canonical name and nicknames</td>
</tr>
<tr>
<td>ContxtSim</td>
<td>TF-IDF similarity between entity’s Baike article and comment</td>
</tr>
<tr>
<td>ContxtSimRank</td>
<td>Inverted rank of ContxtSim across all candidates</td>
</tr>
<tr>
<td>AllInSrc</td>
<td>Whether all words in candidate entity’s canonical name exist in comment</td>
</tr>
<tr>
<td>MatchedNE</td>
<td>The number of matched named entities between entity’s Baike page and comment</td>
</tr>
</tbody>
</table>

### Table 5: Experimental setup statistics.

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th></th>
<th>Valid</th>
<th></th>
<th>Test</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>article</td>
<td>comment</td>
<td>article</td>
<td>comment</td>
<td>article</td>
<td>comment</td>
</tr>
<tr>
<td>Entertainment</td>
<td>734</td>
<td>23,046</td>
<td>98</td>
<td>3,153</td>
<td>149</td>
<td>4,415</td>
</tr>
<tr>
<td>Product</td>
<td>587</td>
<td>3,943</td>
<td>78</td>
<td>473</td>
<td>118</td>
<td>773</td>
</tr>
</tbody>
</table>

Predicts the first entity in the article; (5) PRIORPROB predicts the most likely entity based on $P(e|m)$, estimated from entity-mention co-occurrence in the training set.

**Comparisons.** We further compare against the following models: (1) Vector Space Model (VSM) computes TF-IDF cosine similarity between mention context and entity KB pages, with the most similar candidate as prediction. (2) Logistic Regression (LOGREG) trained with features described in the next paragraph. (3) ListNet is a learning-to-rank approach that outperforms all methods in the EL track of TAC-KBP2009 [Zheng et al., 2010]. (4) CEMEL expands mention representation with similar posts and then applies VSM [Guo et al., 2013b]. We retrieve all comments containing the mention string from the training set as similar posts. (5) MENT-NORM [Le and Titov, 2018] is the state-of-the-art EL model on

---

4. We include all content from entity’s Baidu Baike page.
Table 6: Entity linking results on singular mentions with and without NIL (non-existence in KB) considered. The best performing learning-based models are highlighted in bold per column. No MRR result is reported for baselines where only one entity is returned. Our models that are statistically significantly better than all the baselines and comparisons are marked with * ($p < 0.0001$, approximation randomization test [Noreen, 1989]).

AIDA-CoNLL [Hoffart et al., 2011a], which consists of English news articles. It leverages latent relations among mentions to find global optimal linking results. Important parameters of MENT-NORM, such as the number of latent relations, are tuned on our development set. The same entity and character embeddings as in our model are utilized.

6. Results and Analysis

Main Results. We report evaluation results based on accuracy and Mean Reciprocal Rank (MRR) [Voorhees et al., 1999], which considers the positions of gold-standard entities ranked by each system. Table 6 displays evaluation results for entity mentions excluding plural pronominal mentions. We experiment with two setups based on whether NIL is considered for training and prediction.

Overall, our model achieves significantly better results than all other comparisons on the ENT domain for both setups ($p < 0.0001$, approximation randomization test). For PROD domain, our model also obtains the best accuracy and MRR when NIL is not included. When NIL is considered, while the strong baseline based on prior probability $p(e|m)$ achieves marginally better accuracy, our model yields higher MRR. This is because the ENT domain has much more pronominal mentions (12.9%) than the PROD domain (2.7%). On ENT, our models perform especially well at resolving pronominal cases; on PROD, the prior baseline memorizes the names better, yet our model still obtains the best MRR when NIL is considered.

Results on Plural Pronominal Mentions. Though rarely studied in prior work [Ji et al., 2016], it is common to observe pronominal mentions linked to multiple entities in social media. Here we report results on plural pronominal mentions only in Table 7. We assume the true number of entities is given as $K$, which varies among samples; top $K$ candidates
output by each model are compared against the gold-standards. In addition to accuracy@K and MRR, Normalized Discounted Cumulative Gain (NDCG) [Järvelin and Kekäläinen, 2002] that considers multiple target predictions is reported. As can be seen, XREF with article entity attention significantly outperforms other comparisons. This is likely because plural nominal often refers to the entities in the article, suggesting the effectiveness of article attentions in these samples. Further experiments show that the full model with additional comment attention and features actually yields marginally lower scores.

We further show sample comment attention and articles entity attention output by our model in Figure 3. For the plural pronominal mention (“their”) in comment C1, the article entity attention correctly identifies both “Xiaoming Huang” and “Angelababy” from the news. Comment attention also pinpoints phrases related to the entities, e.g., “acting skills” and “charities”. For mention $m_2$, the article entity attention also correctly indicates entity’s non-existence in the article by giving a high weight to the absent padding entity.

**Error Analysis.** We break down the errors made by each model based on different mention types, as illustrated in Figure 4. Our model XREF produces much less errors in pronominal mentions and other name variations than the comparisons. However, the name matching-based baseline achieves better performance on canonical mentions, indicating a future direction for designing better representation learning over names.

**Effect of Data Augmentation and Ablation Study.** We examine the effect of data augmentation by evaluating models that are trained with manually labeled data only. As can be seen in Table 8, for both domains, there are significant accuracy drops. Moreover,
Figure 4: Error breakdown based on mention type. Our model makes less errors on pronominal mentions and name variations (Others) not captured by KB. “Base Model” represents XREF without attentions and features.

<table>
<thead>
<tr>
<th></th>
<th>Entertainment</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>XREF</td>
<td>57.93</td>
<td>65.33</td>
</tr>
<tr>
<td>w/o Comment Attn</td>
<td>51.94</td>
<td>62.22</td>
</tr>
<tr>
<td>w/o Comment + Article Attn</td>
<td>44.90</td>
<td>55.33</td>
</tr>
</tbody>
</table>

Table 8: Accuracy by our models without data augmentation (NIL not considered).

accuracy drops further when attentions or features are removed. This again demonstrates the effectiveness of comment attention and article entity attention proposed by this work.

7. Conclusion

We present a novel entity linking model, XREF, for Chinese online news comments. Attention mechanisms are proposed to identify salient information from comments and corresponding article to facilitate entity resolution. Model pre-training based on data augmentation is conducted to improve performance. Two large-scale datasets are annotated for experiments. Results show that our model significantly outperforms competitive comparisons, including previous state-of-the-art. For future work, additional languages, including low-resource ones, will be investigated.

References


Stephen Guo, Ming-Wei Chang, and Emre Kiciman. To link or not to link? a study on end-to-end tweet entity linking. In *Proceedings of the 2013 Conference of the North


Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual


Wei Shen, Jianyong Wang, and Jiawei Han. Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Transactions on Knowledge and Data Engineering, 27 (2):443–460, 2015.


Yi Yang and Ming-Wei Chang. S-mart: Novel tree-based structured learning algorithms applied to tweet entity linking. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 504–513, Beijing, China,


