Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2026–2035
Marseille, 11–16 May 2020

CLEEK: A Chinese Long-text Corpus for Entity Linking

Weixin Zeng[†], Xiang Zhao[*†‡], Jiuyang Tang[†‡], Zhen Tan[†], Xuqian Huang[†]
[†]Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, China
[‡]Collaborative Innovation Center of Geospatial Technology, China

Abstract

Entity linking, as one of the fundamental tasks in natural language processing, is crucial to knowledge fusion, knowledge base construction and update. Nevertheless, in contrast to the research on entity linking for English text, which undergoes continuous development, the Chinese counterpart is still in its infancy. One prominent issue lies in publicly available annotated datasets and evaluation benchmarks, which are lacking and deficient. In specific, existing Chinese corpora for entity linking were mainly constructed from noisy short texts, such as microblogs and news headings, where long texts were largely overlooked, which yet constitute a wider spectrum of real-life scenarios. To address the issue, in this work, we build CLEEK, a Chinese corpus of multi-domain long text for entity linking, in order to encourage advancement of entity linking in languages besides English. The corpus consists of 100 documents from diverse domains, and is publicly accessible. Moreover, we devise a measure to evaluate the difficulty of documents with respect to entity linking, which is then used to characterize the corpus. Additionally, the results of two baselines and seven state-of-the-art solutions on CLEEK are reported and compared. The empirical results validate the usefulness of CLEEK and the effectiveness of proposed difficulty measure.

Keywords: Entity linking, Named entity disambiguation, Data annotation

1. Introduction

With the exponential proliferation of unstructured data on the Internet, automatic extraction of valuable information from raw data becomes increasingly crucial. As one of the fundamental steps of bridging raw text and regularized knowledge, entity linking plays an indispensable role in various knowledge-centric tasks, such as knowledge fusion, knowledge base construction, etc.

Entity linking (EL), or entity disambiguation, aims at mapping ambiguous mentions in text to the true entities in a target knowledge base (KB). *Entities* are unique identifiers of objects, while *mentions*, as the surface forms of entities, usually possess various appearances. An instance of EL is depicted in Figure 1: there are three mentions underlined in the text, namely, *The Bulls*, *Hinrich* and *Chandler*. For each mention, EL first retrieves candidate entities from a target KB. As for mention *The Bulls*, the candidate entities include the basketball team *Chicago Bulls*, the animal *Bull*, and possibly many others. Subsequently, EL selects the true entity out of the candidates. For example, the entity *Chicago Bulls* is the true entity for the mention *The Bulls*.

EL is intrinsically non-trivial, since the diverse forms of mentions render it challenging to generate possible candidate entities, let alone selecting true entities out of a group of similar candidates. Over the recent years, endeavours have been devoted to designing accurate and efficient EL systems. In particular, English EL has undergone continuous development, with the aid of up-to-date KBs and evaluation benchmarks. Thus far, numerous English EL corpora have been constructed from different types of sources, including news (Hoffart et al., 2011; Cucerzan, 2007; Rosales-Méndez et al., 2018), tweets (Rowe et al., 2014) and RSS feeds (Röder et al., 2014). The wide range of data sources and textual forms enable comprehensive evaluation of robustness of English EL methods.

By examining the length of source text, we can broadly put the aforementioned datasets into two categories, made from short text, e.g., (Rowe et al., 2014), and from long text, represented by (Hoffart et al., 2011; Cucerzan, 2007). Comparatively, there are more datasets available of the latter kind. This is intuitive, as long text usually provides information with higher quality, in comparison with social media, and it embodies a wider range of real-life textual data. In addition, from the perspective of EL solutions, it is observed that (1) EL on short text tends to require excessive hand-crafted features specific to a certain kind of application, which makes it not necessarily applicable to others; and (2) short-text oriented corpus finds itself inappropriate for evaluating the cluster of EL methods based on collective schemes, since short text is unable to supply enough contextual mentions. As a consequence, long-text oriented corpora are considered to be at least of equal, if not greater, significance to verifying the effectiveness and robustness of EL methods.

In contrast to the advancement in English, however, Chinese EL systems suffer from lagged development, partially due to the lack of appropriate Chinese KBs and evaluation benchmarks. In particular, almost all existing publicly available Chinese EL datasets are based on short text, such as microblogs (NLPCC 2013 [1], NLPCC 2014 [2] , NLPCC 2015 [3]) and news headings (Chen et al., 2018)[4]. Among others, mentions in the NLPCC serial corpora were annotated to noisy and incomplete KBs, which substantially lim-

[*]Corresponding author: Xiang Zhao (xiangzhao@nudt.edu.cn)

[1] http://tcci.ccf.org.cn/conference/2013/pages/page04_eva.html
[2] http://tcci.ccf.org.cn/conference/2014/pages/page04_eva.html
[3] http://tcci.ccf.org.cn/conference/2015/pages/page05_evanotice.html
[4] https://github.com/clhisawolfman/dataset-cnel

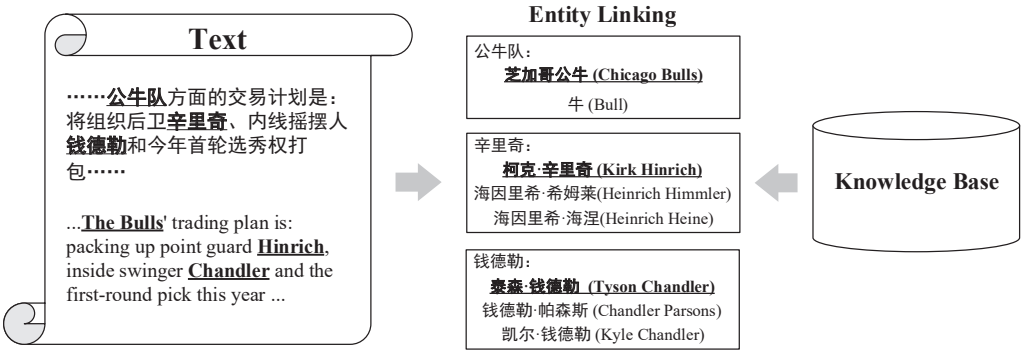

Figure 1: An Example of Entity Linking.

its EL performance; the corpora in (Chen et al., 2018) were annotated with CN-DBpedia (Xu et al., 2017), which has merely provided free APIs so far, hence making it not fully accessible.

Worse still, these Chinese corpora might fail to serve as qualified EL benchmarks. It has been noted that a simple baseline method using prior probability can achieve fairly promising results in many of the datasets (Guo and Barbosa, 2016), leaving an illusion of not much space for further improvement. The underlying reason is that mentions in these datasets are not even ambiguous, since most of them were derived from text with high clarity. As a consequence, there is a pressing need to construct a corpus with a certain level of difficulty, so as to better examine various EL methods.

In short, the drawback of existing Chinese EL corpora is two-fold: (1) All of EL corpora are derived from short text, and hence, fail to cover many real-life scenarios, being unsuitable for examining the *effectiveness* of EL methods; and (2) The difficulty of corpora is not well characterized, and some of the datasets tend to have a bias towards mentions with negligible ambiguity, rendering them inadequate for evaluating the *robustness* of EL methods.

In this work, we propose CLEEK, a Chinese long-text corpus for entity linking, which comprises 100 documents and 2,786 mentions, along with a measure for characterizing corpus difficulty. Specifically, we first elaborate the process of corpus construction and annotation, and then provide an in-depth analysis of corpus properties, in particular the difficulty of dataset. To validate the usefulness of CLEEK, we implement two baselines, **Prior** and **Ctx**, and seven state-of-the-art solutions, **Babelfy**, **Pan**, **PR**, **PPRSim**, **REL-RW**, **NeuPL** and **PairLink**, and report their linking performance on CLEEK. The experiment results also demonstrate that the proposed difficulty measure is capable of indicating the ambiguity of documents, as well as the whole dataset. Noteworthily, through comparisons with existing EL corpora in Table 1, it can be observed that the size of CLEEK is not small in the field of EL. Moreover, the quality of dataset, instead of quantity, is the more important aspect in this paper.

Contributions. The main contributions of this article can be summarized into three ingredients:

- We introduce a Chinese multi-domain long-text corpus for entity linking, namely, CLEEK. To the best of our knowledge, this is among the *first publicly-available* Chinese EL dataset derived from long text and annotated with two major Chinese KBs (Chinese Wikipedia and CN-DBpeida). [5]
- An evaluation measure for characterizing difficulty of EL corpora is put forward to quantify the ambiguity of documents in CLEEK, and empirical results reveal that CLEEK indeed contains documents with various levels of difficulty and validate the usefulness of our proposed difficulty measure.
- Two baselines and seven state-of-the-art solutions are implemented on CLEEK to verify the effectiveness of the presented corpus and serve as references for follow-up research.

2. Related Work

We discuss related work from two perspectives—EL corpora and EL methods.

EL Datasets. There are at least nine datasets in common use for English EL evaluation (Ling et al., 2015), the majority of which are derived from news or web pages. The large number of available corpora inevitably results in unjust comparisons among EL solutions. Moreover, the datasets are of different qualities and difficulties, which might also affect EL performances. We are not aware of any direct research in characterizing EL corpus in terms of difficulty.

With regards to Chinese EL corpora, CLP 2012 [6] is the first to introduce Chinese personal name disambiguation task, whereas the dataset is centered on person names, and for the time being it is not publicly available. Currently, the serial corpora provided by NLPCC are the mainstream evaluation benchmarks for Chinese EL. However, all of them stem from Chinese microblogs, which can be fairly short and noisy. Additionally, the datasets are also annotated to noisy and incomplete KBs, which might well restrain the effectiveness of EL systems.

The knowledge base population (KBP) track [7] includes Chinese EL as a component since 2015. Particularly, for the Chinese EL dataset in KBP2016 task (Ji et al., 2016),

[5] https://github.com/DexterZeng/CLEEK

[6] http://www.cipsc.org.cn/clp2012/bakeoff.html

[7] https://tac.nist.gov/

Name	Type	# D	# M	# M/# D	Difficulty	Language
MSNBC (Cucerzan, 2007)	news	20	658	32.90	Medium	English
ACE2004 (Ratinov et al., 2011)	news	57	253	4.44	Easy	English
DBpedia Spot. (Mendes et al., 2011)	news	58	330	5.69	Medium	English
AIDA TestB (Hoffart et al., 2011)	web	231	4,458	19.30	Medium	English
N3-RSS 500 (Röder et al., 2014)	RSS-feeds	500	524	1.05	Hard	English
Microposts (Rowe et al., 2014)	tweets	1,165	1,140	0.98	Hard	English
NLPCC 2013	tweets	441	826	1.87	Easy	Chinese
NLPCC 2014	tweets	263	607	2.31	Medium	Chinese
VoxEL-strict (Rosales-Méndez et al., 2018)	news	15	204	13.60	Easy	En/DE/ES/FR/IT
VoxEL-relaxed (Rosales-Méndez et al., 2018)	news	15	674	44.93	Easy	En/DE/ES/FR/IT
NTF (Chen et al., 2018)	headings	802	1,777	2.22	Easy	Chinese
CNDL (Chen et al., 2018)	short texts	236	341	1.44	Hard	Chinese
HQA (Chen et al., 2018)	questions	486	549	1.13	Hard	Chinese

Table 1: Statistics of Existing EL Corpora. # D and # M represent the number of documents and mentions respectively. # M/# D denotes the average number of mentions per document. Difficulty is measured according to our proposed difficulty metric in Section 3.

there are 8,845 mentions and 167 documents in evaluation data and 15,000 documents in the source data. Nonetheless, it requires that systems should not leverage topical coherence within each document and mentions are sparsely scattered in the documents, which impose restrictions on using collective EL solutions (Shen et al., 2015). Recently, (Chen et al., 2018) provides several new *short text* based Chinese EL datasets, including NTF (news titles and the first several sentences of news), CNDL (Chinese daily short language sentences) and HQA (hard question answering queries), which are detailed in Table 1.

To sum up, as shown in Table 1, by systematically examining the existing EL corpora from the perspectives of text type, scale and difficulty, it can be concluded that there is a lack of Chinese EL corpus constructed from long text. In addition, a well-defined measure for characterizing corpus difficulty is also of necessity, which can avoid constructing too many datasets with similar difficulties like the English counterpart.

EL Methods. Early works on EL tend to design a set of useful features to capture similarities between mentions and entities and rank the candidates merely in accordance to the semantic matching scores. Although methods of this kind (Mihalcea and Csomai, 2007; Dredze et al., 2010) can achieve good experimental results, semantic coherences within entities are neglected. Considering the deficiencies of previous solutions, collective EL methods (Hoffart et al., 2011; Alhelbawy and Gaizauskas, 2014; Pershina et al., 2015) are put forward. Most works assume mentions in the same document are semantically coherent, which should fit in the textual topic of the whole document.

Recent years have witnessed the emergence of neural networks and their promising performances on many natural language processing related tasks. He et al. (2013) are the first to introduce neural networks into EL framework, followed by Zwicklbauer et al. (2016) and Yamada et al. (2016), who strive to optimize mention, entity and word embeddings (inputs of neural networks). Other works, on the other hand, harness convolutional neural network (CNN) (Xue et al., 2019; Nguyen et al., 2016), recurrent neural network (RNN) (Gupta et al., 2017; Phan et al., 2017), attention mechanism (Ganea and Hofmann, 2017) and graph embeddings (Sevgili et al., 2019; Cao et al., 2018) to extract more effective features to model mention and entity representation. The representations are then leveraged for similarity and relatedness computation to determine the most possible candidate entity.

Compared with continuing advance in English-oriented EL task, Chinese EL is still in its infancy. This can be mainly ascribed to three aspects, namely, lack of up-to-date KBs, shortage of high-quality evaluation datasets, and difficulty posed by Chinese language processing. Thereby, a well-designed Chinese EL corpus can lay the foundation for future research on EL solutions. There is also an emerging tendency for the development of multi-lingual entity linkers (Moro et al., 2014; Pan et al., 2017; Raiman and Raiman, 2018), which have high robustness and can cope with EL problems in low-resource languages.

3. Corpus

In this section, we first elaborate corpus construction process, followed by analysis of corpus properties and introduction of difficulty metric.

3.1. Corpus Construction

As is illustrated in Figure 2, the work flow of corpus construction initiates from mining news and commentaries from websites. Specifically, we crawl approximately 10,000 pieces of long texts from Sohu News [8] and China Newsweek [9], which cover five domains, namely, Sport, Travelling, Economy, Film Review and Politics. Nevertheless, the raw documents are of different length and consist of uneven number of mentions. To construct a long-text based corpus, we require that each document should be at least 350 words long and contain at least 10 mentions. Notably, the mentions represent named entities, and

[8] http://news.sohu.com/

[9] http://www.inewsweek.cn/

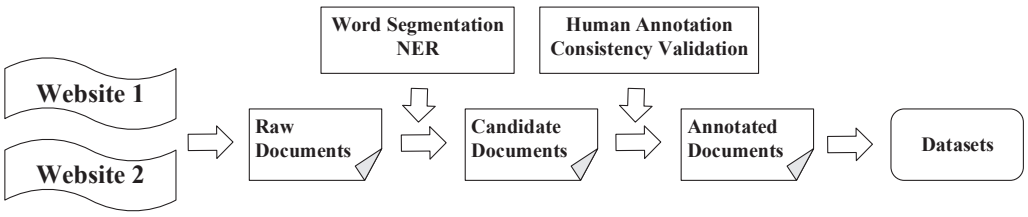

Figure 2: The Work Flow of Corpus Construction.

are defined extensionally: any name uniquely referring to one entity of a predefined class, e.g. a specific person or location (Ling et al., 2015).

Word Segmentation & NER. To generate candidate documents satisfying aforementioned two criteria, the article length threshold is set to 350 and Stanford word segmentation and named entity recognition tool are harnessed to roughly estimate the number of possible mentions in documents (Chang et al., 2008). Since Chinese word segmentation and named entity recognition techniques are still error-prone and cannot be fully trusted, we consider the number of recognized mentions as an indicator of document quality in a proportional manner, whereas they will not be used in following steps due to the low quality. We first filter out documents with length shorter than the threshold (350), and then rank the rest documents according to the amount of recognized mentions and select the top 100 documents from each domain (500 in total) as candidates for human annotation.

Manual Annotation. Volunteers (students working on NLP and familiar with the aforementioned domains) are invited to conduct dataset annotation, which involves two specific tasks: recognizing mentions in each document (**Mention Recognition**) and retrieving correct entity entries from Chinese Wikipedia and CN-DBpedia (**Mention Annotation**). Before annotation, we require them to read ACE (Automatic Content Extraction) Chinese Annotation Guidelines for Entities [10] to fully understand the concepts of mentions and entities. To ensure the annotators are well trained, a principled training procedure is adopted and the annotators are required to pass test tasks before annotating the dataset. And only carefully selected experienced annotators are kept (3 volunteers for each domain eventually).

Mention Recognition. According to the annotation results, the average consistency score for mention recognition is 92.8%, which represents the fraction of overlapped mentions that are recognized by different annotators over all recognized mentions. To resolve disagreements, all the volunteers are gathered for discussion and make final decisions by majority voting. As shown in Figure 3, there are disagreements concerning mention recognition of phrase *Nocioni and Deng*. Despite that in English, it is evident that this phrase comprises two mentions, in Chinese it is fairly confusing since *Deng* rarely refers to a foreigner's name, as told by Annotator 2 and 3, who consider it as a typo during annotation. This disagreement is further resolved after Annotator 1 points out that *Deng* actually refers to *Luol Deng* and his annotation also receives the most votes.

Figure 3: Example Showing the Difficulty of Corpus Annotation. A sport-related document is annotated by Annotator 1, 2 and 3, where there are inconsistencies during mention recognition and mention annotation. Note that for each domain, we have assigned annotators who have adequate background knowledge, and this example is to describe the difficulties during corpus annotation (which rarely happens since annotators are familiar with their responsible domains) and how we tackle the problems.

Mention Annotation. The average mention annotation consistency score is 97.3%, which is obtained by dividing the number of mentions which are annotated to the same entities by different volunteers, by the total number of mentions, and the differences are also settled by further discussion and majority voting. As shown in Figure 3, when annotating the corresponding entity of mention *Gordon*, annotator 2 wrongly labels it as *Aaron Gordon*, since both *Ben Gordon* and *Aaron Gordon* have strong connections with *McGrady*. This is further corrected after majority voting. Again this is a very rare case (considering the high consistency score), which on the other hand reflects that the corpus is of a certain level of difficulty.

The Kappa inner-annotator agreement is substantial, at 0.64 and 0.72 for mention recognition and mention annotation, respectively.

Post-processing. It should be highlighted that we merely

[10] http://www.ldc.upenn.edu/Projects/ACE/

select 100 documents with relatively *higher quality* (appropriate document length and plenty of mentions) to constitute the final corpus, since the majority (>320) of documents contain approximately 15 or even less mentions, let alone *unique mentions*, and it might not be appropriate to include a document that merely contains 15 mentions, among which many are *repetitive*, into the whole dataset. Also, in this kind of document, EL solutions tend to achieve either too good or too bad results. Furthermore, the quality of dataset, instead of quantity, is the more important aspect. In this connection, within each domain, we sort the annotated documents according to the number of unique mentions in a descending order and keep the top ranked documents, such that the overall quality of the dataset can be well controlled. Note that the number of final documents within each domain is adjusted in accordance to its overall quality (which is unevenly distributed). In consequence, there are 10 Economy/Travelling related documents, 20 documents in the domains of Film and Politics, respectively, and 40 documents concerning Sport.

Corpus Properties. As is displayed in Table 2, there are 2,786 mentions in total, among which 181 are NIL mentions, meaning that their corresponding entities cannot be found in the target KB. Similar to most previous EL datasets, we do not investigate long-tail situations in this work. Each document is around 610 words long, containing approximately 28 mentions. Additionally, the annotated entities in CLEEK also cover a wide range of entity types, including person, location, organization, geo-political entities and facilities, etc.

3.2. Difficulty Measure

Guo et al. (2016) reveal that most existing datasets are biased towards popular entities and merely utilizing prior probability can achieve promising results. The prior probability is a statistic index, which represents the possibility of an entity being true given a specific mention name according to the statistical data across the web. For instance, regarding mention "Apple", based on the occurrences on the Internet, the possibility that it refers to entity "Apple company" will be higher than that of entity "Apple (fruit)", and this possibility is termed as prior probability.

Consequently, there is a pressing need to devise a measure for characterizing corpus difficulty. Although the performance of using prior probability can be regarded as an advisable measure, it neglects the semantic similarities between mentions and true entities. In other words, provided that a mention and its corresponding entity are close in the semantic space, the linking process could be easily realized via EL solutions based on neural networks and embeddings. Considering the on-going advancement of neural EL methods, we propose to adopt the accuracy score achieved by combining prior probability and embedding similarity to form a better indicator of EL corpus difficulty, represented as D.

Specifically, suppose the accuracy of merely using prior probability for EL on the corpus is P, and M is the accuracy score attained by only considering the cosine similarity between mention and entity name embeddings (averaged word embedding of the words in the name). Then $D = \alpha P + (1 - \alpha)M$ and the difficulty of each document can accordingly be denoted as $d = \alpha p + (1 - \alpha)m$, where the lower-case symbols represent document-wise indexes. Higher D or d values denote easier corpus or document. In our work, to balance the contributions made by two disparate indicators, we set $\alpha = 0.5$. The word embeddings are generated by training Word2vec (Mikolov et al., 2013) on language-specific Wikipedia dump.

To examine the usefulness of proposed difficulty metric, on the basis of document difficulty d, we divide the datasets into three groups, namely, easy, medium and hard. As is presented in Table 3, documents with medium difficulty occupy the largest share, which fits in real-life situation since news and commentaries tend to be understandable but not too explicit. Figure 4 displays examples of documents with different levels of difficulty, which also meets human perception of ambiguity/difficulty. The effectiveness of our proposed difficulty metric is further verified via experimental results in Table 4.

4. Experiment

In this section, we first introduce the baselines and competing methods that are implemented on CLEEK, followed by the description of experimental settings. In the end, the linking results on CLEEK and subsets with different degrees of difficulty are reported.

4.1. Baselines and Solutions

We evaluated two baselines, prior probability **Prior** and context similarity **Ctx**, and seven state-of-the-art EL solutions, **Babelfy**, **Pan**, **PR**, **PPRSim**, **REL-RW**, **NeuPL** and **PairLink** on CLEEK.

Prior probability. As is mentioned in Section 3., prior probability represents the possibility of an entity being true given a specific mention according to the statistical data across the web. We obtained the statistical information via the frequency dictionary introduced in Section 4.2. The prior probability of entity e_m for mention m is calculated by dividing the frequency value of e_m for m by the overall frequency value of all candidate entities for m.

Context similarity. Context similarity ranks the candidate entities according to the text similarity between mention context and entity description. Following recent works, we harnessed long short-term memory (LSTM) to capture semantics in text and calculate context similarity.

The framework of calculating context similarity is illustrated in Figure 5, which comprises three LSTM units, and they are harnessed to model the representations of mention's left context, mention's right context and entity description, respectively. Then the *mention representation* is generated by concatenating the max-pooling results of the two LSTMs for mention, while the *entity representation* is composed of the entity embedding, as well as the max-pooling result of the entity description LSTM. Notably, the entity description is derived from the first paragraph of its corresponding Wikipedia page with a given text length. Eventually, *mention representation* and *entity representation* are concatenated and forwarded to the two fully-connected layers so as to generate the final similarity

Property	Overall	Sport	Economy	Film	Travelling	Politics
Number of Documents	**100**	40	10	20	10	20
Number of Mentions	**2,786**	1,345	253	599	242	347
Number of Linkable Mentions	**2,605**	1,228	235	585	233	324
Mentions per Document	**27.86**	33.63	25.30	29.95	24.20	17.35
Average Document length	**609.7**	597.8	553.2	657.5	700.3	568.7
Number of Sentences	**1,293**	517	105	283	140	248

Table 2: Corpus Properties.

Category	# Docs	# Avg. M	# Avg. Len	Condition
Easy	32	27.31	592.6	$\{document \mid 0.6 \leq d(document) \leq 1\}$
Medium	49	26.43	602.1	$\{document \mid 0.4 < d(document) < 0.6\}$
Hard	19	32.47	657.9	$\{document \mid 0 \leq d(document) \leq 0.4\}$

Table 3: Datasets Divided by Document Difficulty. The specific approach of calculating d can be found in Section 3.2.. **# Docs**, **# Avg. M** and **# Avg. Len** denote the number of documents, the average number of mentions per document and the average document length respectively.

score. Noteworthy is that word embeddings, which are the inputs of LSTM units, along with entity embedding, are obtained via a joint training process elaborated in Section 4.2. The final mention-entity similarity score is denoted as s, and g (ground truth) is set to 0/1 if the candidate entity is true/wrong. The objective of training is to minimize the loss:

$$L(s, g) = (1 - g)\log(1 - s) + g\log(s). \quad (1)$$

Competing methods. Aside from two baselines, we also implemented seven state-of-the-art EL methods: (1) **Babelfy** (Moro et al., 2014): a graph-based EL system using BabelNet semantic network for disambiguation, which supports multi-lingual EL; and (2) **Pan** (Pan et al., 2017): an EL system performing a series of KB mining methods to achieve the task of identifying name mentions, assigning a coarse-grained or fine-grained type to each mention, and linking it to a KB, given a piece of text in any language; and (3) **NeuPL** (Phan et al., 2017): an EL system employing LSTM and attention mechanism for entity disambiguation and also incorporating a simple but effective and significantly fast linking algorithm to improve linking accuracy; and (4) **PR** (Alhelbawy and Gaizauskas, 2014): a collective disambiguation approach using a graph model and PageRank algorithm for candidate entities ranking and linking; and (5) **PPRSim** (Pershina et al., 2015): a novel graph-based disambiguation approach based on Personalized PageRank that combines local and global evidence for disambiguation and effectively filters out noise introduced by incorrect candidates; and (6) **REL-RW** (Guo and Barbosa, 2014): an EL system harnessing the notion of semantic similarity rooted in Information Theory and capturing global coherence via random walks on the disambiguation graph induced by choice of entities for each mention; and (7) **PairLink** (Phan et al., 2019): this work introduces MINTREE, a new tree-based objective for the problem of entity disambiguation, and designs Pair-Linking (**PairLink**), a novel iterative solution for the MINTREE optimization problem.

4.2. Experimental Settings

Dataset. Regarding evaluation benchmark, we utilized CLEEK (annotated to Chinese Wikipedia). It is noted that six out of nine aforementioned baselines do not need training data, while **Ctx**, **NeuPL** and **PairLink** require training corpus constructed from Wikipedia (detailed later). As thus, we used all the documents for testing. Nevertheless, if needed, this dataset can also be easily split into training/validation/test sets.

Frequency Dictionary. A frequency dictionary records mentions and their possible corresponding entities with frequency values, which is harnessed to obtain prior probability and generate candidate entities. In this work, the frequency dictionary was derived from Wikipedia dump [11], which was also considered as the local KB. In specific, we obtained all the anchor texts in the dump, along with their links, and replaced the links with corresponding entity names, thus creating the frequency dictionary.

Joint Embedding Training. In this work, similar to state-of-the-art EL solutions (Phan et al., 2017; Yamada et al., 2016), we harnessed a joint embedding approach which mapped entities and words to the same continuous vector space, and similar entities and words were placed close to each other. Concretely, the joint training process was achieved by utilizing python package Gensim [12], where the dimension was 100, iteration was set to 3, and window size was 5. Eventually, there were 6,363,417,735 items with embedding values.

LSTM Network Settings. Regarding the parameter values in the neural network, the window size for the mention's left and right context was 20, while for entity description, the size was 100. Note that we considered segmented tokens as basic units and zero paddling was leveraged if the

[11] We utilized Chinese Wikipedia dump on 01-Dec-2017: `https://dumps.wikimedia.org/zhwiki/20171201/`

[12] `https://radimrehurek.com/gensim/`

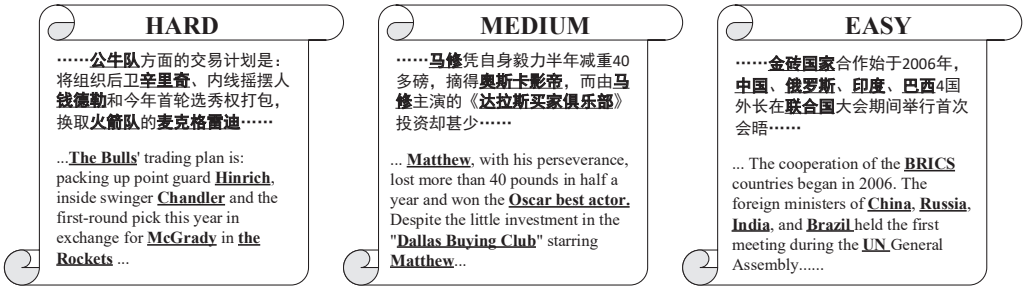

Figure 4: Examples of Documents with Different Levels of Difficulty Generated by Our Proposed Measure. Evidently, mentions in the Hard document are rather ambiguous, as *Hinrich*, *Chandler* can refer to many different entities. In contrast, Easy document contains very obvious mentions such as *BRICS*, *UN* and the country names. This verifies that our proposed measure can well characterize the difficulty of a document and in turn the whole dataset.

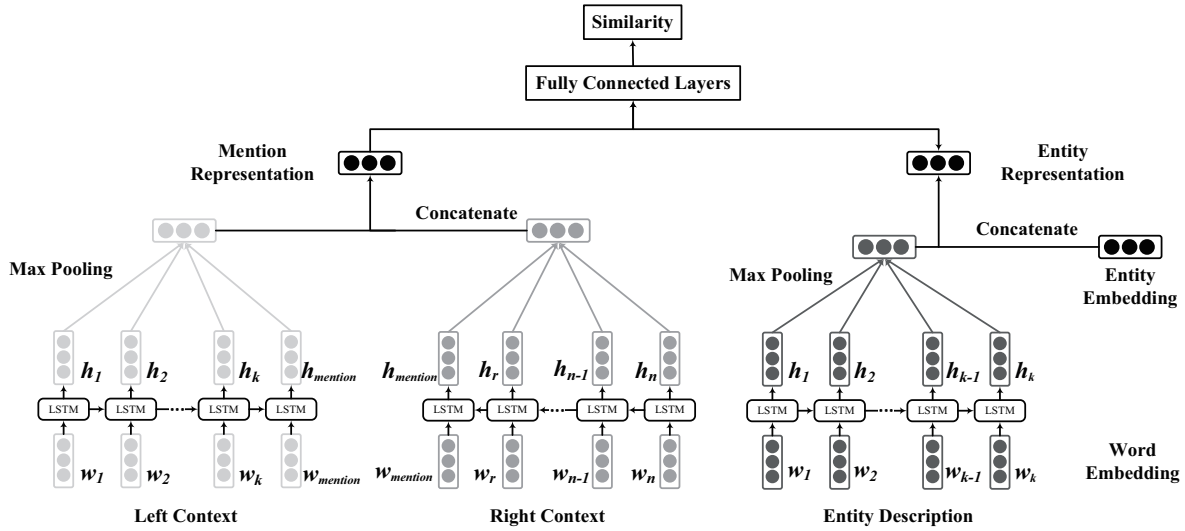

Figure 5: Context Similarity via LSTMs.

number of tokens was below window size. Before being forwarded to the LSTM network, the tokens were first replaced by their embedding vectors. In addition, the hidden state size of LSTM units was 96 and the output size for first fully connected layer was set to 200. The activation function for fully connected layer was $\tanh$ and Adam was harnessed as the optimizer. We set the number of epochs and batch size to 30 and 128 respectively.

For training the neural network, the training corpus was also derived from Wikipedia. We regarded the anchor texts as mentions and the entities that anchor texts referred to were considered as the true entities. Due to the large size of Wikipedia and the limited computational resources, we merely took into account the aggregated candidate entities over the dataset and 100 mentions were generated for each entity. Moreover, negative sampling strategy was harnessed, and five negative samples were created for each correct sample, which was achieved by substituting the correct entity with other entities in the mention's candidate entities. This training corpus was also utilized for **NeuPL** and **PairLink**.

Evaluation Metric. We used overall *Accuracy* as evaluation metric, which represents the fraction of correctly linked mentions over all mentions.

Implementation Details of Existing Methods. Both **Babelfy** [13] and **Pan** [14] offer well-wrapped APIs, and we directly used them to obtain linking results. We implemented **NeuPL** and **PairLink** by using the codes kindly provided by authors, whereas we replaced their candidate entities generation strategy with ours since their strategy was only applicable to English. Regarding **PR**, **PPRSim** and **REL-RW**, we reproduced them with the optimal settings reported in respective papers, although all of their candidate entities generation processes were replaced with ours, since again, their strategies were targeted at English.

4.3. Results

Candidate Entities Generation. Despite the fact that entity ranking method is crucial to the overall EL performance, its previous step, candidate entities generation, determines the upper bound of the linking accuracy. Concretely, chances are that in the candidate generation step, the candidate entities generated for some mentions do not contain the true entity, thus leading to wrong linking results in spite of following steps.

[13] http://babelfy.org/

[14] http://blender02.cs.rpi.edu:3300/elisa_ie/api

Dataset	**Babelfy**	**Pan**	*Upp. Bound*	**Prior**	**Ctx**	**PR**	**PPRSim**	**REL-RW**	**PairLink**	**NeuPL**
CLEEK	0.393	0.241	*0.753*	0.627	0.545	0.638	**0.655**	0.649	0.642	0.630
Hard	0.300	0.193	*0.574*	0.441	0.378	0.447	**0.475**	0.473	0.447	0.440
Medium	0.382	0.280	*0.754*	0.622	0.558	0.640	0.651	**0.652**	0.649	0.636
Easy	0.600	0.373	*0.871*	0.759	0.641	0.764	**0.782**	0.763	0.764	0.752
Sport	0.423	0.192	*0.782*	0.652	0.581	0.660	**0.673**	0.666	0.665	0.648
Economy	0.460	0.328	*0.796*	0.681	0.596	0.685	0.672	0.672	**0.723**	0.693
Film Review	0.236	0.120	*0.597*	0.504	0.453	0.506	0.526	**0.542**	0.505	0.505
Travelling	0.485	0.330	*0.828*	0.631	0.476	0.652	**0.734**	0.657	0.665	0.652
Politics	0.448	0.515	*0.840*	0.710	0.593	0.750	0.750	**0.753**	0.725	0.728

Table 4: Linking Accuracy and the Upper Bound Value. The methods using their own candidate generation strategies are placed on the left of *Upper Bound* column, while the methods on the right are based on our candidate generation strategies. The best results are in boldface and the second-best are underlined.

In our work, three strategies, i.e., mention regularization, frequency dictionary and Wikipedia functional pages, were utilized to boost the *Upper Bound* value produced by candidate entities generation. The *Upper Bound* index denotes the fraction of mentions which contain correct entity in their candidate entities, over all mentions. Mention regularization helps remove useless punctuations and formalize mention expressions, while frequency dictionary includes the possible candidates given a specific mention. Since we utilized Wikipedia dump as the local KB, its disambiguation pages and redirect pages were harnessed to retrieve candidate entities. In short, for mention m, we first cleaned its surface form by using mention regularization strategy, and then retrieved its candidate entities according to the frequency dictionary and Wikipedia functional pages. More details can be found in (Zeng et al., 2018).

The specific empirical performance is reported in Table 4. Although the candidate entities generation strategies can improve *Upper Bound* value, it merely reaches 75.3% in the overall dataset, and it represents the *utmost Accuracy* value that can be attained by **Prior**, **Ctx**, **PR**, **PPRSim**, **REL-RW**, **NeuPL** and **PairLink**.

EL Results. The EL results are reported in Table 4. First of all, it is not hard to observe that, both **Babelfy** and **Pan** attain quite poor results, with the overall accuracy standing at 39.3% and 24.1% on CLEEK respectively. The inferior outcome, to a certain degree, can be attributed to the deficiency of generating candidate entities. Specifically, according to the observation of results, the candidate entities generation strategies for the two approaches fail to generate candidate entities for the majority of mentions. Also, the candidate entities generation strategies in them cannot be replaced by other relatively superior methods, e.g., the strategies we utilized, since both **Babelfy** and **Pan** merely offer APIs. More importantly, the results also reveal that the effectiveness of multi-lingual entity linkers on a specific language is in fact doubtful, hence it is more appropriate for them to serve as back-up solutions if there are no EL systems on the specified language.

As for other competitors, we utilized aforementioned candidate entities generation strategies, and the *Upper Bound* results are reported in the corresponding column. The two baselines, **Prior** and **Ctx**, attain overall accuracies at 62.7% and 54.5% respectively, which proves that CLEEK is of a certain degree of difficulty. In addition, **PR**, **PPRSim**, **REL-RW**, **PairLink** and **NeuPL** outperform the baselines, verifying the effectiveness of existing methods, among which **PPRSim** yields the most promising outcome on CLEEK and most subsets. Noteworthy is that the gaps among the results would widen if better entity generation strategies were harnessed to attain higher *Upper Bound*, since low *Upper Bound* restrains the linking performance of these methods. The results over different domains also indicate the superiority of existing solutions over baselines and the ambiguity of documents in the corpus.

With regard to the results over subsets with different degrees of difficulty, both *Upper Bound* and overall linking accuracy decline over 10% when the difficulty climbs to a higher stage. For instance, **PairLink** only achieves linking accuracy at 44.7% in the hard segment of CLEEK, while this value surges to 76.4% on the easy fraction. In all, the experiment results not only validate the usefulness of CLEEK, but also reveal that the difficulty measure can well characterize the ambiguity of documents and in turn the overall corpus difficulty.

5. Conclusion

We construct a Chinese corpus of multi-domain long-text for EL, CLEEK, to make up for the shortage of high-quality evaluation benchmarks in Chinese EL. Moreover, in order to avoid the development of EL datasets on which simple baselines can achieve promising results, a corpus difficulty measure is proposed to characterize the quality of EL corpora. To validate the usefulness of CLEEK, we report and compare the results of two baselines and seven state-of-the-art EL systems. The evaluation results on CLEEK and the subsets divided according to difficulty metric reveal that our proposed corpus is of high quality with documents in different levels of difficulty, and the difficulty measure can well characterize the ambiguity of documents.

Acknowledgment. This work was partially supported by NSFC under grants Nos. 61872446, 61902417, and 71971212, and NSF of Hunan province under grant No. 2019JJ20024.

6. Bibliographical References

Alhelbawy, A. and Gaizauskas, R. J. (2014). Graph ranking for collective named entity disambiguation. In *Proceedings of the 52nd Annual Meeting of the Association*

for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers, pages 75–80.

Cao, Y., Hou, L., Li, J., and Liu, Z. (2018). Neural collective entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 675–686.

Chang, P., Galley, M., and Manning, C. D. (2008). Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation, WMT@ACL 2008, Columbus, Ohio, USA, June 19, 2008*, pages 224–232.

Chen, L., Liang, J., Xie, C., and Xiao, Y. (2018). Short text entity linking with fine-grained topics. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 457–466.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pages 708–716.

Dredze, M., McNamee, P., Rao, D., Gerber, A., and Finin, T. (2010). Entity disambiguation for knowledge base population. In *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, pages 277–285.

Ganea, O. and Hofmann, T. (2017). Deep joint entity disambiguation with local neural attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2619–2629.

Guo, Z. and Barbosa, D. (2014). Robust entity linking via random walks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 499–508.

Guo, Z. and Barbosa, D. (2016). Robust named entity disambiguation with random walks. *Semantic Web*, (Preprint):1–21.

Gupta, N., Singh, S., and Roth, D. (2017). Entity linking via joint encoding of types, descriptions, and context. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2681–2690.

He, Z., Liu, S., Li, M., Zhou, M., Zhang, L., and Wang, H. (2013). Learning entity representation for entity disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers*, pages 30–34.

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. (2011). Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 782–792.

Ji, H., Nothman, J., Dang, H. T., and Hub, S. I. (2016). Overview of tac-kbp2016 tri-lingual edl and its impact on end-to-end cold-start kbp. *Proceedings of TAC*.

Ling, X., Singh, S., and Weld, D. S. (2015). Design challenges for entity linking. *TACL*, 3:315–328.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: shedding light on the web of documents. In *Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011*, pages 1–8.

Mihalcea, R. and Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6-10, 2007*, pages 233–242.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

Moro, A., Cecconi, F., and Navigli, R. (2014). Multilingual word sense disambiguation and entity linking for everybody. In *Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, ISWC 2014, Riva del Garda, Italy, October 21, 2014.*, pages 25–28.

Nguyen, T. H., Fauceglia, N., Rodriguez-Muro, M., Hassanzadeh, O., Gliozzo, A. M., and Sadoghi, M. (2016). Joint learning of local and global features for entity linking via neural networks. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2310–2320.

Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., and Ji, H. (2017). Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1946–1958.

Pershina, M., He, Y., and Grishman, R. (2015). Personalized page rank for named entity disambiguation. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 238–243.

Phan, M. C., Sun, A., Tay, Y., Han, J., and Li, C. (2017). Neupl: Attention-based semantic matching and pair-linking for entity disambiguation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 1667–1676.

Phan, M. C., Sun, A., Tay, Y., Han, J., and Li, C. (2019).

Pair-linking for collective entity disambiguation: Two could be better than all. *IEEE Trans. Knowl. Data Eng.*, 31(7):1383–1396.

Raiman, J. and Raiman, O. (2018). Deeptype: Multilingual entity linking by neural type system evolution. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5406–5413.

Ratinov, L., Roth, D., Downey, D., and Anderson, M. (2011). Local and global algorithms for disambiguation to wikipedia. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 1375–1384.

Röder, M., Usbeck, R., Hellmann, S., Gerber, D., and Both, A. (2014). N^3 - A collection of datasets for named entity recognition and disambiguation in the NLP interchange format. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 3529–3533.

Rosales-Méndez, H., Hogan, A., and Poblete, B. (2018). Voxel: A benchmark dataset for multilingual entity linking. In *The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings, Part II*, pages 170–186.

Matthew Rowe, et al., editors. (2014). *Proceedings of the 4th Workshop on Making Sense of Microposts co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 7th, 2014*, volume 1141 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Sevgili, Ö., Panchenko, A., and Biemann, C. (2019). Improving neural entity disambiguation with graph embeddings. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 2: Student Research Workshop*, pages 315–322.

Shen, W., Wang, J., and Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Trans. Knowl. Data Eng.*, 27(2):443–460.

Xu, B., Xu, Y., Liang, J., Xie, C., Liang, B., Cui, W., and Xiao, Y. (2017). Cn-dbpedia: A never-ending chinese knowledge extraction system. In *Advances in Artificial Intelligence: From Theory to Practice - 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017, Arras, France, June 27-30, 2017, Proceedings, Part II*, pages 428–438.

Xue, M., Cai, W., Su, J., Song, L., Ge, Y., Liu, Y., and Wang, B. (2019). Neural collective entity linking based on recurrent random walk network learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5327–5333.

Yamada, I., Shindo, H., Takeda, H., and Takefuji, Y. (2016). Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 250–259.

Zeng, W., Tang, J., and Zhao, X. (2018). Entity linking on chinese microblogs via deep neural network. *IEEE Access*, 6:25908–25920.

Zwicklbauer, S., Seifert, C., and Granitzer, M. (2016). Robust and collective entity disambiguation through semantic embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 425–434.