News Citation Recommendation with Implicit and Explicit Semantics

Hao Peng[*][1] **Jing Liu**,[2] **Chin-Yew Lin**[2]

[1]School of EECS, Peking University, Beijing, 100871, China
[2]Microsoft Research, Beijing, 100080, China
penghao.pku@gmail.com {liudani, cyl}@microsoft.com

Abstract

In this work, we focus on the problem of news citation recommendation. The task aims to recommend news citations for both authors and readers to create and search news references. Due to the sparsity issue of news citations and the engineering difficulty in obtaining information on authors, we focus on content similarity-based methods instead of collaborative filtering-based approaches. In this paper, we explore word embedding (i.e., implicit semantics) and grounded entities (i.e., explicit semantics) to address the variety and ambiguity issues of language. We formulate the problem as a reranking task and integrate different similarity measures under the learning to rank framework. We evaluate our approach on a real-world dataset. The experimental results show the efficacy of our method.

1 Introduction

When an author writes an online news article, s/he often cites previously published news reports to elaborate a mentioned event or support his/her point of view. For the convenience of the readers, the editor usually associates the words with hyperlinks. Through the links the readers can directly access the referenced articles to know more details about the events. If there is no reference for a mentioned event, the readers may search the related news reports for further reading. Hence, it is valuable to have automatic news citation recommendations for authors and readers to create or search news references.

In this paper, we focus on the problem of *news citation recommendation*. As shown in Table 1, given a snippet of citing context (left), the task aims to retrieve a list of news articles (right) as references. This task differs from traditional recommendation tasks, e.g., citation recommendation for scientific papers, in that: (a) based on the statistics from our dataset, the number of references per news article is 4.56 on average, much less than the number of citations per academic paper (typically dozens); (b) the author-topic information is usually unavailable, since it is technically difficult to obtain author information from news articles. These differences make the collaborative filtering-based methods, which have been widely applied to paper citation recommendation, less available in our scenario. Therefore, in this paper we focus on content similarity-based methods to deal with the task of news citation recommendation.

Previous studies use string-based overlap (Xu et al., 2014), machine translation measures (Madnani et al., 2012), and dependency syntax (Wan et al., 2006; Wang et al., 2015) to model text similarity. More recent work focuses on neural network methods (Yin and Schütze, 2015; He et al., 2015; Hu et al., 2014; dos Santos et al., 2015; Lei et al., 2016). There are two major challenges rendering these approaches not suitable for this task: (i) the variety and (ii) the ambiguity of language. By variety, we mean that the same meaning may be expressed with different phrases. Taking the first row in Table 1 for example, *Vlaar* in the citing context refers to *Ron Vlaar*, a Dutch football player, who is referred to as *Dutch star* and *Netherlands international* in the cited article. By ambiguity, we mean that the same expression may have different meanings in different contexts. In the second example in Table 1, the mention *tiger* refers to *tiger the mammal*. By contrast, in "Detroit Tigers links: The Tigers are in trouble" for example, the word *Tiger* is the name of a team. In this paper, we explore both implicit and explicit semantics to ad-

[*] Work done during internship at Microsoft Research.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 388–398,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

Citing Context	Cited Article
. . . Manchester United and Arsenal have both been interested in Vlaar in the past, suggesting Southampton will have to fight hard to land him. . . .	**Man United and Arsenal on red alert as top Dutch star officially joins free agent list** The Netherlands international has joined the free agent list today and is no longer contractually obliged to remain at Villa Park. . . .
. . . Conservationists want the Bangladeshi government to step up and help save the tigers through greater administration and enforcement of anti-poaching laws, as Bangladesh does not legally protect tigers to the extent that other governments do, according to Inhabitat. . . .	**Bangladesh 's abundant tiger population has collapsed to just 100** In Bangladesh, a new census shows that tiger populations in the Sundarbans mangrove forest are more endangered than ever. The study, which used hidden cameras to track and record tigers, provides a more accurate update than previous surveys that used other methods. The year-long census, which ended this April, revealed only around 100 of the big cats remain in what was once home to the largest population of tigers on earth. . . .

Table 1: Two pair of news snippets. For readability concerns, we keep only the sentence associated with an anchor link in the citing part, and the **title** and lead paragraph of the cited part.

dress the above issues. Specifically, the implicit semantics can be obtained from the word embedding trained on large scale corpus, and the explicit semantics through linking entity mentions to the grounded entities in a knowledge base.

In this paper, we explore using both word embedding and grounded knowledge to model the relatedness between citing context and articles. We formulate the problem as a re-ranking task. We use learning to rank to integrate different similarity measures and evaluate the models on a real world dataset constructed from Bing News[1]. We further give quantitative analysis of the effects of word embedding and grounded entities in the task.

In summary, the main contributions of this paper are three-fold:

- We propose the task of *news citation recommendation* and construct a real-world dataset for this task.
- We utilize both word embedding based similarity measures and knowledge-based methods to tackle the problem. We formulate the problem as a re-ranking task and leverage learning to rank algorithm to integrate different similarity measures.
- We conduct extensive experiments on a large dataset. The results show the effectiveness of word embedding and grounded entities. We further quantitatively analyze how the implicit semantics from word embedding and explicit semantics from grounded knowledge benefit the task of interest.

2 Problem Formulation

In this section, we introduce the *news citation recommendation* problem and formulate it as a re-ranking task. We first introduce definitions that will be used through the rest of the paper:

Citing Context. *Citing context* is a sentence which contains an anchor text associated with a hyperlink. As shown in Table 1, the underlined words are associated with a hyperlink pointing to another news article, and the sentence (left) which contains the anchor is the citing context.

Cited Article. Given a piece of citing context, the article that the hyperlink links to is defined as its *cited article*. It is expected that a news article is well-structured, and its headline together with its lead paragraph gives a good brief description of the whole story (Kianmehr et al., 2009). In this paper, a news article can either be represented by its title and lead paragraph or by the passage as a whole. We conduct experiments under both of the two different settings.

Candidate Article Set. Considering efficiency, we follow the procedure adopted by many recommendation systems (Lei et al., 2016; Tan et al., 2015) and formulate the problem as a re-ranking task. In another word, given a citing context, we first use efficient retrieval methods with high recall to generate a list of articles as the *candidate article set*, and then run the system to get a re-ranked list.

News Citation Recommendation. Given a citing context, the task aims to construct an ordered list of news articles, top of which are most relevant to the context, and can serve as the cited articles.

[1] `https://www.bing.com/news`

3 Method

In this section, we first explain the similarity measures based on word embedding (implicit semantics) and grounded knowledge (explicit semantics) to deal with variety and ambiguity problems. Then we briefly introduce the baselines and the learning to rank framework.

3.1 Implicit Semantics for Variety

The distributed word representation by *word2vec* factors word distance and captures semantic similarities through vector arithmetic (Mikolov et al., 2013). In this work, we train a skip-gram model to bridge the vocabulary gap between context-article pairs. Previous work represents the documents with averaged vectors of words (Tang et al., 2014; Tan et al., 2015). However, this may lead to the loss of detailed information of the documents. In this paper, we adopt a different approach, explained below.

Word Mover's Distance (WMD). Kusner et al. (2015) combine distributed word representations with the earth mover's distance (EMD) (Rubner et al., 1998; Wan, 2007) to measure the distance between documents. They use the Euclidean distance between words' low dimensional representations as building blocks, and optimize a special case of the EMD to obtain the cumulative distance. More formally, let $X = \{(x_1, w_{x_1}), (x_2, w_{x_2}), \cdots, (x_m, w_{x_m})\}$ be the normalized bag-of-words representation for a citing context after removing stop-words, where word x_i appears w_{x_i} times (then normalized by the total count of words in X), $i = 1, 2, \cdots, m$. Similarly, we have the representation for a candidate article, $Y=\{(y_1, w_{y_1}), (y_2, w_{y_2}), \cdots, (y_n, w_{y_n})\}$. The WMD calculates the minimum cumulative cost by solving the linear programming problem below:

$$\min_{\mathbf{T}} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{T}_{ij} c_{ij}$$

$$s.t. \quad \sum_{j=1}^{n} \mathbf{T}_{ij} = w_{x_i}, \quad i = 1, 2, \cdots, m,$$

$$\sum_{i=1}^{m} \mathbf{T}_{ij} = w_{y_j}, \quad j = 1, 2, \cdots, n,$$

where $\mathbf{T} \in \mathbb{R}^{m \times n}$ is the transportation flow matrix, and c_{ij} indicates the distance between x_i and y_j. Here $c_{ij} = \|\textbf{vector}(x_i) - \textbf{vector}(y_j)\|$, where function $\textbf{vector}(w)$ returns the word vector of w. Then the distance is normalized by the total flow:

$$WMD(X, Y) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{T}_{ij} c_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{T}_{ij}}$$

3.2 Explicit Semantics for Ambiguity

News articles tend to be well written, and contain many named entity mentions. Making use of this property, we deal with the ambiguity problem by using grounded entities (explicit semantics). Given a context-article pair, we first recognize all named entity mentions on both sides and link them to knowledge bases (e.g., Wikipedia and Freebase), then use the following measures to model the similarity.

- **Entity Overlap.** Given a context–article pair, we consider two metrics, namely, $precision$ and $recall$, to measure their entity overlap. The $precision$ is defined as:

  $$precision = \frac{\textit{entity-overlap}(\textit{citing}, \textit{cited})}{\textit{entity-count}(\textit{citing})}$$

  and $recall$ as:

  $$recall = \frac{\textit{entity-overlap}(\textit{citing}, \textit{cited})}{\textit{entity-count}(\textit{cited})}$$

- **Embedding Based Matching.** We build two separate information networks for Wikipedia entities using (a) the anchor links on Wikipedia pages and (b) the Freebase entity graph (Bollacker et al., 2008). Then we apply Large-scale Information Network Embedding (LINE) (Tang et al., 2015) system[2] to the networks to embed the entities into low-dimensional spaces. We then measure the similarity by the minimized cosine distance between entities' on the citing and the cited side:

  $$minDIS_{citing} = \frac{1}{|citing|} \sum_{i=1}^{|citing|} \min_{y_j \in Y} (1 - cos(x_i, y_j)),$$

  and vice versa:

  $$minDIS_{cited} = \frac{1}{|citied|} \sum_{j=1}^{|cited|} \min_{x_i \in X} (1 - cos(x_i, y_j)),$$

  where X refers to the citing context, and $x_i \in X$ are the grounded entities in the citing part. Similar with Y and y_j.

[2] https://github.com/tangjianpku/LINE

- **Wikipedia Evidence.** Given a context-article pair, we refer to world knowledge for supporting evidence. In particular, we first apply an entity linking system to detect the entity mentions on both sides and ground them into Wikipedia entries, each of which has its own description page. Second, we collect the descriptions for entities from the candidate article and extract as evidence those sentences containing entities from citing context. We refer to this evidence as *cited evidence*. For instance, the article in Table 8 contains grounded entity *Scottish National Party*. And in the description for it, there is a sentence containing the entity *Scotland* from the citing context: "The Scottish National Party (SNP) is a Scottish nationalist and social-democratic political party in *Scotland*." Thus we extract this sentence as cited evidence supporting this pair.

  We count the overlapping nouns between the citing context and the cited evidence to calculate $precision$ and $recall$,

$$precision = \frac{noun\text{-}overlap(context, cited\ evidence)}{noun\text{-}count(context)}$$

$$recall = \frac{noun\text{-}overlap(article, citing\ evidence)}{noun\text{-}count(article)}$$

3.3 Baselines

We design several baseline features for the two groups of features mentioned above:

- **TF-IDF Distance.** We use TF-IDF distance as a basic measure. The similarity is calculated with cosine distance based on TF-IDF vector representations for the text.
- **Ungrounded Mentions.** Note that entity overlap features also adapt to ungrounded mentions. The embedding-based matching features for ungrounded mentions are similar to those for grounded entities. The only difference is that here each mention is represented by the averaged vectors of all the words it contains. Wikipedia evidence is not feasible for ungrounded mentions.

Table 2 summarizes all the features we use. A cited article can either be represented by its headline+lead paragraph or as a whole. Therefore, we extract features under two different settings: (a) headlines and lead paragraphs only; (b) the full articles. Most of the features are extracted under both of the settings. However, feature 2 is much too computation-intensive and feature 7 needs POS-tagging as the preprocessing. Thus these two are only extracted under setting (a).

3.4 Learning to Rank Framework

Many different learning to rank algorithms have been proposed to deal with the ranking problem, including pointwise, pairwise, and listwise approaches (Xia et al., 2008). Listwise methods receive ranked lists as training samples, and directly optimize the metric of interest by minimizing the respective cost. And it has been reported that the listwise method usually achieves better performance compared to others (Qin et al., 2008; Cao et al., 2007). In this work, we use the linear model and apply coordinate ascent for parameter optimization.

4 Experiments

4.1 Data Collection

We collect one month's news articles from Bing News. The citing context set consists of all the sentences associated with anchor link(s). For each piece of citing context, its cited article is extracted through its hyperlink. If there are multiple links associated with the context, only the first one is considered. We pair each citing context and its cited article as a ground truth sample. We further label as ground truths those articles sharing the same title as the cited article. This is rather reasonable since a single passage may have multiple reprints by difference sources. On average, there are 2.20 ground truth cited articles for each citing context in the dataset.

In order to focus only on news events, we filter out those pairs whose hyperlinks are associated with three words or less (usually names for persons or places, and lead to definition pages). We also discard those samples whose citing contexts contain or are exactly the same as the titles of the cited articles. For example, "READ MORE: The stories you need to read, in one handy email" links to an article titled "The stories you need to read, in one handy email".

The dataset is preprocessed with Stanford CoreNLP toolkit (Manning et al., 2014), including sentence splitting, tokenizing for whole passages, and POS-tagging for titles and lead paragraphs. We use the JERL system by Luo et al. (2015) for entity detection and grounding. It recognizes entity mentions and links them to Wikipedia entries.

	Feature	Full Article?	# of features	Description
Dealing with Variety				
1	WMD	n	1	Word vector based earth mover's distance.
Dealing with Ambiguity				
2	Grounded Entity Overlap	y	4	Precision and recall for grounded named entities.
3	Embedding-based Matching	y	16	Minimized matching distance with LINE vectors.
4	Wikipedia Evidence	n	2	Precision and recall for evidence from Wikipedia.
Baselines				
5	TF-IDF	y	2	The cosine distance with TF-IDF.
6	Ungrounded Mention Overlap	y	4	Precision and recall for ungrounded mentions.
7	Embedding-based Matching	y	4	Minimized matching distance with averaged vectors.
Total			33	

Table 2: A list of all features used in the experiments. The third column indicates whether the corresponding feature is extracted from the full articles. If not, it's extracted only from the headlines and lead paragraphs.

We use each mention's text span as an ungrounded mention, and its corresponding Wikipedia ID as a grounded entity. For instance, in Table 8, the detected text span *Westminster* is an ungrounded mention, and it's grounded to the entry *Parliament of the United Kindom*.

4.2 Selecting Candidates

Given a citing context, we construct its candidate article set with the top 200 articles retrieved by TF-IDF distance. In the experiments, approximately 92.61% of the ground truth cited articles appear in the candidate sets. We discard those that do not. We further randomly split the remaining 33318 pairs into training/validation/test sets with the proportion of 3:1:1.

For each training pair, we randomly sample 5 articles from its candidate article set (excluding ground truth) and pair them with the citing context as negative samples. According to Tan et al. (2015), the number of negative samples does not significantly affect the linear learning to rank model's performance. During validation and testing, all of the 200 candidates are taken into account.

4.3 Experimental Setup

In the experiments, we set the TF-IDF as the baseline, and incrementally add different groups of features to the system.

The word embedding is pretrained with skip-gram model (Mikolov et al., 2013) on Wikipedia corpus and then fine-tuned using the method proposed in Wieting et al. (2015) on PPDB (Ganitkevitch et al., 2013). The embedding fine-tuned with paraphrase pairs can better capture the semantic relatedness of different phrase. In the experiments, we observe a $1\% - 2\%$ improvement by the fine-tuned word representations compared to vanilla skip-gram vectors.

We use the linear model in RankLib[3] for the learning to rank implementation. Coordinate ascent is used for parameter optimization. The model is trained to directly optimize the evaluation metrics, Precision@1, Precision@5, NDCG@5 and MAP, respectively.

For NDCG@5 measure, we set a binary relevance score, i.e., the scores equal to 1 for ground truths, 0 for negative samples.

4.4 Experimental Results

Table 3 gives the performance of the baselines and the systems using different groups of features on test and validation sets. The results show that WMD brings a consistent improvement over its TF-IDF baseline, and so do grounded entities compared to ungrounded mentions.

Individually added to the TF-IDF baseline, WMD has the largest performance boost, followed by grounded entity features. Besides, the additional information from grounded entity knowledge helps the model outperform the ungrounded mentions, with a consistent margin of 1.0%-2.0% NDCG@5.

We further compare the performance of the models when using features from headlines+lead paragraphs only and those from full passages. As shown in Table 3, the former brings much better performance on each metric compared to the latter.

It's worth noting that there are ground truths mis-labeled as irrelevant in the dataset. A primary

[3] `https://sourceforge.net/p/lemur/wiki/RankLib/`

id	Features	Precision@1		Precision@5		NDCG@5		MAP	
		Test	Dev	Test	Dev	Test	Dev	Test	Dev
	Headline + Lead Para								
1	TF-IDF	42.61	42.21	19.84	19.78	52.72	52.22	53.50	53.06
2	+ Ungrounded Mentions	43.67	43.04	19.45	19.30	53.84	53.26	54.46	54.12
3	+ Grounded Entities	44.52	44.02	20.84	20.51	55.99	55.0	56.55	56.09
4	+ Ungrounded+Grounded	43.93	44.05	20.2	19.66	55.99	55.17	56.52	56.13
5	TF-IDF + WMD	45.94	45.84	21.11	21.62	57.20	57.50	58.12	58.34
6	+ Ungrounded Mentions	46.44	46.63	21.05	21.56	57.61	57.80	58.55	58.78
7	+ Grounded Entities	**47.63**	**47.5**	**21.96**	**22.07**	58.52	58.41	**60.01**	**59.83**
8	+ Ungrounded+Grounded	47.23	46.84	21.58	21.56	**59.01**	**58.91**	59.88	59.66
	Full Article								
9	TF-IDF	49.3	48.11	23.33	23.06	60.51	59.54	60.71	59.73
10	+ Ungrounded Mentions	50.46	50.42	23.81	23.67	61.97	61.73	62.6	61.94
11	+ Grounded Entities	51.42	50.27	23.91	23.78	63.26	62.09	63.23	62.15
12	+ Ungrounded+Grounded	51.46	50.23	23.85	23.74	62.94	62.48	63.15	63.02
13	TF-IDF + WMD	52.31	51.82	23.87	24.04	63.71	63.99	64.08	63.62
14	+ Ungrounded Mentions	53.26	53.3	23.98	24.16	64.57	64.29	64.52	64.37
15	+ Grounded Entities	**54.12**	**53.29**	24.37	24.05	65.29	64.48	65.32	64.53
16	+ Ungrounded+Grounded	54.04	53.21	**24.52**	**24.33**	**65.56**	**65.11**	**65.35**	**64.56**

Table 3: Experimental results in percentage on the dataset collected from Bing News.

id	Features	NDCG@5 on $\mathbf{S}$	NDCG@5 on $\tilde{\mathbf{S}}$
	Headline + Lead Para		
1	TF-IDF	52.77	56.28
2	+ Ungrounded Mentions	53.34	56.86
3	+ Grounded Entities	55.03	58.57
4	+ Ungrounded+Grounded	55.18	58.86
5	TF-IDF + WMD	56.51	60.13
6	+ Ungrounded Mentions	57.04	60.82
7	+ Grounded Entities	57.4	61.47
8	+ Ungrounded+Grounded	**58.05**	**61.78**

Table 4: Experimental results in percentage on $\mathbf{S}$ and $\tilde{\mathbf{S}}$. $\mathbf{S}$ is a randomly constructed subset of the test set, and $\tilde{\mathbf{S}}$ is obtained by manually labeling samples in $\mathbf{S}$.

reason is that news sites sometimes individually publish different reports on a certain event. And the articles don't necessarily share the same title. To see how this affects the model, we randomly build a subset $\mathbf{S}$ of the test set and manually label the selected samples, which gives $\tilde{\mathbf{S}}$[4]. Table 4 compares the model's performance on $\mathbf{S}$ and $\tilde{\mathbf{S}}$ under Headline+Lead paragraph setting. There is a consistent improvement of NDCG@5 score on $\tilde{\mathbf{S}}$ compared to that on $\mathbf{S}$. Besides that, on manually labeled data, the model's performance across different feature settings is almost in accord with that on the full test set. These results show that there are indeed mis-labeled ground truths in the dataset, but they have little influence when comparing different groups of features.

5 Analysis

In this section, we give detailed win-loss analysis for the models trained with NDCG@5 metric under headlines+lead paragraphs setting. Specifically, given two systems with different feature configurations, we compare their performance on each test sample. The results are shown as a heatmap in Figure 1. X and Y axises indicate the identifiers for each feature group, following those in Table 3. For example, the data point at $(5, 1)$ indicates that the inclusion of WMD brings better ranking scores to TF-IDF on 18.4% of the test samples; and as a trade off, it lowers the scores on 11.4% of the samples. We also observe that grounded entities brings gain to 15.9% of the samples, and loss for 9.6% of them. On average, two different groups of features disagree on 26.4% of the test samples.

We further give several mis-predictions by the model using certain groups of features, and illustrate how they are corrected by the inclusion of others (or the other way round). By misprediction, we mean that no ground truth cited article appears in the top 5 predictions of the returned list.

5.1 Dealing with Variety

Table 5 shows a mis-prediction by TF-IDF, but corrected after including WMD.

TF-IDF distance favors the high-score match-

[4] Manually labeling all of the dev and test samples would be too time consuming, and we leave it to future work.

ing keywords *approval* and *rating* between the citing context and mis-predicted article. On the other hand, distributed word representations factor the distances between word pairs, which helps to capture their semantic closeness, e.g., (*Argentines*, *Argentina*, Cosince distance: 0.210), (*poll*, *election*, 0.020), and (*increasingly*, *growing*, 0.286). WMD helps to bridge the vocabulary gap between the citing context and the cited article.

On the other hand, though not often, the use of distributed representation can also create mistakes. Table 6 gives an example where the inclusion of the WMD feature changes a correct prediction by TF-IDF into a mistake. By analyzing the WMD's transportation flow matrix **T**, we find that the used word embedding relates *MP* to *minister*, and *publicly* to *government*. More curiously, persons' names are very similar in its semantic space: (*Davies*, *Stephen*, 0.602), and (*Davies*, *Harper*, 0.635). A possible reason could be that both of the two names are very common, and thus the cooccurrence-based representation learning method is not able to distinguish them. This also justifies our use of grounded entities as additional information: from the Wikipedia description for entity *Stephen Harper*, the system might be able to find out that he actually serves in Canadian government, not in the UK's nor in the Welsh.

5.2 Dealing with Ambiguity

Entity grounding helps by resolving the ambiguity e.g., alias, abbreviation, of the entity mentions.

As shown in Table 7, *tiger* refers to *the mammal* in the ground truth pair. However, the same word refers to *Detroit Tigers the team* in the mis-predicted article. This ambiguity is resolved when the mention is grounded to its Wikipedia entry. In another example shown in Table 8, ungrounded mention *SNP*, though detected, contributes little to supporting the ground truth pair. However, when it's grounded to the entry *Scottish National Party*, the system leverages world knowledge and relates it to the mention of *Scotland* in the citing context.

The inclusion of grounded entity information may also lead to mistakes, many of which are due to the limited performance of the entity recognition and disambiguation system. We'd like to discuss another kind of error here, shown in Table 9. In the citing context, *The Daily Telegraph* is a newspaper published in the UK. It has little to do with the involved event except for reporting it. However, the system favors a farmers' story which actually happened in the UK. We find that this contributes a lot to the system's errors when including grounded entities. We leave it to future work to figure out how to deal with this issue.

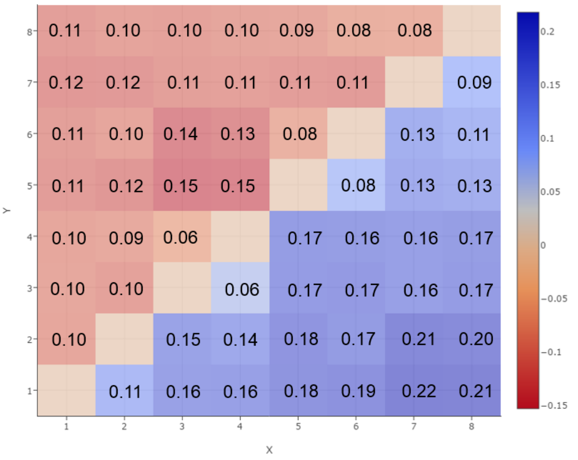

Figure 1: Heatmap for win-loss analysis results. Point (x, y) indicates how much feature x wins (loses if negative) against y.The X and Y axises indicate the identifiers for each feature group, following those in Table 3.

6 Related Work

This section reviews three lines of related work: (i) document recommendation, (ii) pharaphrase identification, (iii) question retrieval.

6.1 Document Recommendation

Existing literature mainly focuses on content-based methods and collaborative filtering (Adomavicius and Tuzhilin, 2005). There are studies trying to recommend documents based on citation contexts, either through identifying the motivations of the citations (Aya et al., 2005), or through the topical similarity (Ritchie, 2008; Ramachandran and Amir, 2007). On the other hand, Mcnee et al. (2002) leverage multiple information sources from authorship, paper-citation relations, and co-citations to recommend research papers.

Combining the context-based approaches and collaborative filtering, Torres et al. (2004) and Strohman et al. (2007) report better performance. Tang and Zhang (2009) use the Restricted Boltzmann Machine to model citations for placeholders, and Tan et al. (2015) integrate multiple features to recommend quotes for writings.

In the news domain, context-based approaches are presumably favorable due to the fact that the articles are relatively content-rich and citation-sparse. Previous studies manage to utilize information retrieval techniques to recommend news articles given a seed article (Yang et al., 2009; Bogers and van den Bosch, 2007).

Sides	–	Samples
citing	–	An earlier poll showed Argentines are also increasingly happy with her performance as President, putting her approval rating at almost 43%, up from 31% in September.
cited	Ground Truth	**Kirchner's Growing Popularity Could Skew Argentine Election** As Argentina gears up for a presidential election in October, the approval ratings of the current president, Cristina Kirchner, are improving and her rising reputation could affect the results of the election to replace her.
	Top-1 Prediction	**Bill Shorten's Approval Rating Falls in Wake of Royal Commission** The opposition leader gained approval from only 27% of the voters surveyed, while 52% disapproved.

Table 5: A mis-prediction by TF-IDF corrected by the inclusion of WMD.

Sides	–	Samples
citing	–	Chris Grayling was responding to a question from Gower Conservative MP Byron Davies about the regeneration investment fund for Wales "and the underselling of a large amount of publicly owned property".
cited	Ground Truth	**Wales land deal leaves taxpayers 15m short** A Welsh government spokesperson said there were conflicting valuations.
	Top-1 Prediction	**Conservative MP compares Stephen Harper government to Jesus, inspiring hilarious #CPCJesus tweets** Is it time we started referring to Prime Minister Stephen Harper as "Our Lord and Saviour"?

Table 6: A correct prediction by TF-IDF but then changes into a mistake when including WMD.

Sides	–	Samples
citing	–	Conservationists want the $\underset{\text{Government of Bangladesh}}{\underline{\text{Bangladeshi government}}}$ to step up and help save the tigers through greater administration and enforcement of anti-poaching laws, as $\underset{\text{Bangladesh}}{\underline{\text{Bangladeshi}}}$ does not legally protect tigers to the extent that other governments do, according to Inhabitat.
cited	Ground Truth	**$\underset{\text{Bangladesh}}{\underline{\textbf{Bangladeshi's}}}$ abundant tiger population has collapsed to just 100** In $\underset{\text{Bangladesh}}{\underline{\text{Bangladeshi}}}$, a new census shows that tiger populations in $\underset{\text{Sundarbans}}{\underline{\text{the Sundarbans}}}$ mangrove forest are more endangered than ever. The study, which used hidden cameras to track and record tigers, provides a more accurate update than previous surveys that used other methods.
	Top-1 Prediction	**$\underset{\text{Detroit Tigers}}{\underline{\textbf{Detroit Tigers}}}$ links: The $\underset{\text{Detroit Tigers}}{\underline{\textbf{Tigers}}}$ are in trouble** After losing three straight games prior to $\underset{\text{Major League Baseball All-Star Game}}{\underline{\text{All-Star break}}}$, the the $\underset{\text{Detroit Tigers}}{\underline{\text{Tigers}}}$ don't have much more time to waste if they want to stay in contention.

Table 7: A mis-prediction by TF-IDF corrected by the inclusion of grounded entity features. The linked Wikipedia entries are indicated below the underlined entity mentions.

Sides	–	Samples
citing	–	With activities at $\underset{\text{Parliament of the United Kingdom}}{\underline{\text{Westminster}}}$ challenging a narrow view of nationalism, and a planned charm offensive across $\underset{\text{IrelandUnited Kingdom relations}}{\underline{\text{the UK and Ireland}}}$, it is that the party intends to significantly expand its reach beyond $\underset{\text{Scotland}}{\underline{\text{Scotland}}}$.
cited	Ground Truth	**$\underset{\text{Scottish National Party}}{\underline{\textbf{SNP}}}$ launches bid to extend influence beyond $\underset{\text{Scotland}}{\underline{\textbf{Scotland}}}$** First Minister of Scotland and $\underset{\text{Scottish National Party}}{\underline{\text{SNP leader}}}$ $\underset{\text{Nicola Sturgeon}}{\underline{\text{Nicola Sturgeon}}}$ worked hard to reassure voters in the election campaign.
	Top-1 Prediction	**$\underset{\text{Apple Pay}}{\underline{\textbf{Apple Pay}}}$ $\underset{\text{United Kingdom}}{\underline{\textbf{UK}}}$ launch confirmed for mid-July** Leaked documents from retailers suggest a launch date early next week.

Table 8: A mis-prediction by TF-IDF+ungrounded mention features corrected by the TF-IDF+grounded entity features. The linked Wikipedia entries are indicated below the underlined entity mentions.

Sides	–	Samples
citing	–	According to a UK Telegraph report, the government is now forcing farmers and food United Kindom The Daily Telegraph manufacturers to sell anywhere from 30-100% of their products to the state , as opposed to stores and supermarkets.
cited	Ground Truth	**Venezuelan farmers ordered to hand over produce to state** Venezuela As Venezuela's food shortages worsen, the president of the country's Food Industry Venezuela Food Industry Chamber has said that authorities ordered producers of milk, pasta, oil, rice, sugar and flour milk oil rice sugar to supply their products to the state stores.
	Top-1 Prediction	**Welsh farmers launch #NoLambWeek price campaign** United Kindom Fed-up Welsh farmers are encouraging others to withhold their fat lambs for a week in United Kindom protest at the current slump in the UK lamb trade.

Table 9: A correct prediction by TF-IDF but then changes into a mistake when including grounded entity features. The linked Wikipedia entries are indicated below the underlined entity mentions.

6.2 Paraphrase Identification

Several hand-crafted features have proven helpful in modeling sentence/phrase similarity, e.g., string-based overlap (Xu et al., 2014), machine translation measures (Madnani et al., 2012), and dependency syntax (Wan et al., 2006; Wang et al., 2015). Using the combination and discriminative re-weighting of the mentioned features, Ji and Eisenstein (2013) manage to obtain more competitive results.

More recent work has switched the focus onto neural methods. Socher et al. (2011) recursively encode the representations of sentences by the compositions of words. Convolutional neural nets (LeCun et al., 1998; Collobert and Weston, 2008) are also exploited in the tasks of paraphrase identification and sentence matching (Yin and Schütze, 2015; He et al., 2015; Hu et al., 2014).

Story link detection (SLD) is a similar task which aims to classify whether two news stories discuss the same event. Farahat et al. (2003) leverage part of speech tagging technique as well as task-specific similarity measures to boost the system's performance. Shah et al. (2006) show that entity based document representation is a better choice compared to word-based representations in SLD. In our scenario, the query is typically a piece of context sentence instead of an entire article. Therefore, we find that document level methods yield sub-optimal performance when used to model the similarity of citing context and the articles. Besides, due to the fact that there might be multiple reports for a single event, we consider it reasonable to formulate our problem into a ranking task instead of classification.

6.3 Question Retrieval

The key problem in question retrieval lies in modeling questions' similarity. Machine translation techniques (Jeon et al., 2005) and topic models (Duan et al., 2008) have been utilized by previous works. An alternative is representation learning. Zhou et al. (2015) use category-based meta-data to learn word embeddings. dos Santos et al. (2015) and Lei et al. (2016) obtain superior performance over hand-crafted features with CNN.

News articles are more well-written than most documents in QA communities, which results in the feasibility of high-quality entity detection and grounding.

7 Discussions

In this paper, we propose a novel problem of *news citation recommendation*, which aims to recommend news citations for references based on a citing context. We develop a re-ranking system leveraging implicit and explicit semantics for content similarity. We construct a real-world dataset. The experimental results show the efficacy of our approach.

8 Acknowledgments

This research is partially supported by National Basic Research Program of China under Grant No. 2015CB352201, National Natural Science Foundation of China under Grant No. 61502014, and China Post-doctoral Foundation under Grant No. 2015M580927.

References

Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.*, pages 734–749.

Selcuk Aya, Carl Lagoze, and Thorsten Joachims. 2005. Citation classification and its applications. In *Proceedings of the International Conference on Knowledge Management*, pages 287–298.

Toine Bogers and Antal van den Bosch. 2007. Comparing and evaluating information retrieval algorithms for news recommendation. In *Proceedings of the 2007 ACM Conference on Recommender Systems*, RecSys '07, pages 141–144.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, pages 129–136.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *ICML*.

Cícero Nogueira dos Santos, Luciano Barbosa, Dasha Bogdanova, and Bianca Zadrozny. 2015. Learning hybrid representations to retrieve semantically equivalent questions. In *ACL-IJCNLP 2015*, pages 694–699.

Huizhong Duan, Yunbo Cao, Chin-Yew Lin, and Yong Yu. 2008. Searching questions by identifying question topic and question focus. In *ACL*, pages 156–164.

Ayman Farahat, Francine Chen, and Thorsten Brants. 2003. Optimizing story link detection is not equivalent to optimizing new event detection. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 232–239.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: the paraphrase database. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, pages 758–764.

Hua He, Kevin Gimpel, and Jimmy Lin. 2015. Multi-perspective sentence similarity modeling with convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1576–1586.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In *Advances in Neural Information Processing Systems*, pages 2042–2050.

Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee. 2005. Finding similar questions in large question and answer archives. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 84–90.

Yangfeng Ji and Jacob Eisenstein. 2013. Discriminative improvements to distributional sentence similarity. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 891–896.

Keivan Kianmehr, Shang Gao, Jawad Attari, M. Mushfiqur Rahman, Kofi Akomeah, Reda Alhajj, Jon Rokne, and Ken Barker. 2009. Text summarization techniques: Svm versus neural networks. In *Proceedings of the 11th International Conference on Information Integration and Web-based Applications Services*, pages 487–491.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *ICML*, pages 957–966.

Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE.*

Tao Lei, Hrishikesh Joshi, Regina Barzilay, Tommi S. Jaakkola, Kateryna Tymoshenko, Alessandro Moschitti, and Lluís Màrquez i Villodre. 2016. Semi-supervised question retrieval with gated convolutions. In *NAACL HLT 2016*.

Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. Joint entity recognition and disambiguation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 879–888.

Nitin Madnani, Joel R. Tetreault, and Martin Chodorow. 2012. Re-examining machine translation metrics for paraphrase identification. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, pages 182–190.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Sean M. Mcnee, Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K. Lam, Al M. Rashid, Joseph A. Konstan, and John Ried. 2002. On the recommending of citations for research papers. *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pages 116–125.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tao Qin, Xu-Dong Zhang, Ming-Feng Tsai, De-Sheng Wang, Tie-Yan Liu, and Hang Li. 2008. Query-level loss functions for information retrieval. *Inf. Process. Manage.*, 44:838–855.

Deepak Ramachandran and Eyal Amir. 2007. Bayesian Inverse Reinforcement Learning. *Proceedings of the 20th International Joint Conference on Artical Intelligence*, 51:2586–2591.

Anna Ritchie. 2008. *Citation Context Analysis for Information Retrieval*. Ph.D. thesis.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 1998. A metric for distributions with applications to image databases. In *Proceedings of the Sixth International Conference on Computer Vision*.

Chirag Shah, W. Bruce Croft, and David Jensen. 2006. Representing documents with named entities for story link detection (sld). In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pages 868–869.

Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems*, pages 801–809.

Trevor Strohman, W. Bruce Croft, and David Jensen. 2007. Recommending citations for academic papers. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 705–706.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2015. Learning to recommend quotes for writing. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2453–2459.

Jie Tang and Jing Zhang. 2009. A discriminative approach to topic-based citation recommendation. In *Advances in Knowledge Discovery and Data Mining*, volume 5476, pages 572–579.

Xuewei Tang, Xiaojun Wan, and Xun Zhang. 2014. Cross-language context-aware citation recommendation in scientific articles. In *Proceedings of the 37th International ACM SIGIR Conference on Research &#38; Development in Information Retrieval*, pages 817–826.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *WWW*. ACM.

R. Torres, S.M. McNee, M. Abel, J.A. Konstan, and J. Riedl. 2004. Enhancing digital libraries with techlens. In *Digital Libraries, 2004. Proceedings of the 2004 Joint ACM/IEEE Conference on*, pages 228–236.

Stephen Wan, Mark Dras, Robert Dale, and Cecile Paris. 2006. Using dependency-based features to take the 'para-farce' out of paraphrase. In *Proceedings of the Australasian Language Technology Workshop 2006*, pages 131–138.

Xiaojun Wan. 2007. A novel document similarity measure based on earth mover's distance. *Inf. Sci.*, pages 3718–3730.

Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. 2015. Syntax-based deep matching of short texts. In *IJCAI*, pages 1354–1361.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *TACL*, 3:345–358.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. Listwise approach to learning to rank: Theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1192–1199.

Wei Xu, Alan Ritter, Chris Callison-Burch, William B. Dolan, and Yangfeng Ji. 2014. Extracting lexically divergent paraphrases from twitter. *TACL*, 2:435–448.

Yin Yang, Nilesh Bansal, Wisam Dakka, Panagiotis Ipeirotis, Nick Koudas, and Dimitris Papadias. 2009. Query by document. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 34–43.

Wenpeng Yin and Hinrich Schütze. 2015. Convolutional neural network for paraphrase identification. In *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 901–911.

Guangyou Zhou, Tingting He, Jun Zhao, and Po Hu. 2015. Learning continuous word embedding with metadata for question retrieval in community question answering. In *ACL-IJCNLP 2015*, pages 250–259.