Event Linking: Grounding Event Mentions to Wikipedia

Anonymous ACL submission

Abstract

Comprehending an article requires understanding its constituent events. However, the context where an event is mentioned often lacks the details of this event. A question arises: how can the reader obtain more knowledge about this particular event in addition to what is provided by the local context in the article?

This work defines Event Linking, a new natural language understanding task at the event level. Event linking tries to link an event mention appearing in an article to the most appropriate Wikipedia page. This page is expected to provide rich knowledge about what the event mention refers to. To standardize the research in this new direction, we first formally define Event Linking task. Second, we collect a dataset for this new task. Specifically, we automatically gather training set from Wikipedia, and then create two evaluation sets: one from the Wikipedia domain, reporting the in-domain performance, and a second from the real-world news domain, to evaluate out-of-domain performance. Third, we propose EVELINK, the first-ever event linking system. Overall, as our analysis shows, Event Linking is a considerably challenging task requiring more effort from the community. Data and code will be publicly released.

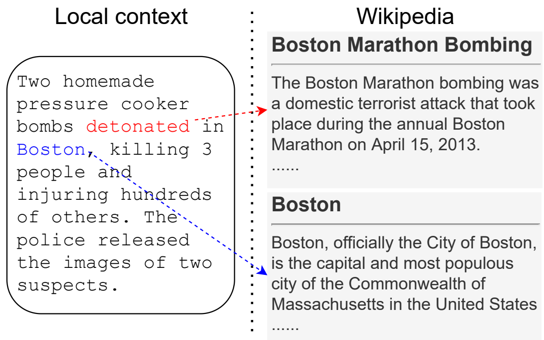

Figure 1: Examples of Event linking and Entity linking. The left side is the local context, and the right side contains Wikipedia pages. Entity linking model connects the entity "Boston" to the Wikipedia page "Boston", while event linking model links the event "detonated" to the Wikipedia page "Boston Marathon Bombing", which is more relevant to the local context.

1 Introduction

Grounding is a process of disambiguation and knowledge acquisition, and is an important task for natural language understanding. Entity linking, grounding entity mentions to a knowledge base (usually Wikipedia) (Bunescu and Pasca, 2006; Mihalcea and Csomai, 2007; Ratinov et al., 2011; Gupta et al., 2017; Wu et al., 2020), has been shown important in natural language understanding tasks, such as question answering (Yih et al., 2015). Despite the significant progress brought by entity linking, our following analysis will show that grounding entities may not provide enough background knowledge that is often needed to support text understanding. Consider the example, Figure 1; an entity linking model will link the entity "Boston" to the Wikipedia page "Boston" which introduces the history and culture of the city Boston. The information we can get from the page "Boston" is irrelevant to the local context. To really help understand this sentence, we need to link the event centered by the verb "detonated" to the Wikipedia page "Boston Marathon Bombing". We call this process that grounds events **Event Linking**.

In this paper, we formulate this Event Linking task for the first time and carefully design a benchmark dataset for this task. We automatically collect training data from the hyperlinks in Wikipedia, and create two evaluation sets to evaluate both in-domain and out-of-domain performance. For in-domain evaluation, the test data is also from hyperlinks in Wikipedia. To avoid models from overfitting, the test data is balanced with hard cases and easy cases determined by whether the event is seen in the training and by the similarity between the surface forms of event mentions and Wikipedia titles. For out-of-domain evaluation, we annotate real-world news articles across 20 years collected from New York Times. Considering the sparsity of events existed in Wikipedia, we also add "Nil" annotation to the test data, indicating that those events do not exist in Wikipedia, therefore, the model needs to tag them as "Nil".

Technically, we come up with an event linking model EVELINK that consists of two steps. The first step "candidate generation" uses a bi-encoder to narrow down the candidate space efficiently; the second step "candidate ranking" uses a more advanced cross-encoder to derive the matching degree between an event mention and a Wikipedia title. Both steps rely on a novel representation learning approach, which is our main technical contribution, for event mentions as well as Wikipedia titles. EVELINK achieves strong performance on seen events and easy cases in test data, and shows the difficulties of grounding unseen events and hard cases. We conduct a detailed analysis to better understand the task as well as the new model.

To conclude, our contributions are three-fold: (i) We formulate the task Event Linking. (ii) We collect training data for this task, and design both in-domain and out-of-domain test data, with a balanced ratio of hard cases and easy cases to ensure the quality of the dataset. (iii) Our proposed approach EVELINK shows promising performance in experiments, and our in-depth analysis provides a better understanding of this new problem, the dataset, and the new approach.

2 Grounding Events in Wikipedia

Given an article (from news for example) and an event mention m in it, event linking tries to find a title t, from all the Wikipedia titles, to provide the best explanation of m. A correct title is defined as follows: as long as a Wikipedia page is about this event, or any subsection of the page introduces this event, we regard its title as the correct one.

We note that the term "event linking" has been used in the literature, e.g., (Nothman et al., 2012; Krause et al., 2016). However, these works are essentially performing cross-document event coreference: determine if a given event mention refers to another event mention (in the same or another document). We, on the other hand, link an event mention to a Wikipedia concept with a different purpose: acquiring external knowledge about the event which is often beyond what we can obtain from the local context. Our definition of event linking can not only improve the understanding of the article, but also pave the way for the intensively-studied event coreference and other relation identification problems studied in works on events.

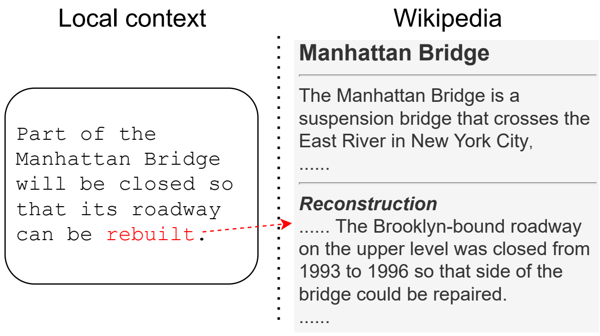

Figure 2: Example of event mentions that only exist in the subsection of a Wikipedia page. The event "rebuilt" of the bridge does not have its own page, but is mentioned in the subsection of the page "Manhattan Bridge".

Event Linking vs. Entity Linking Relatedness: (i) They both link an object (event/entity) from an article to Wikipedia; (ii) Some events, such as "World War II", are entities; in this case, two tasks are the same. Distinctions: (i) Entities are mostly consecutive text spans. Events, in contrast, are more structured objects, consisting of a trigger and a couple of arguments. Since arguments are mostly entities, events generally contain entities as the components. More complex structures in events make event linking more challenging; (ii) Except for some events that are also entities, generally speaking, events are information units of larger granularity. A better comprehension of events, such as through linking to Wikipedia, is expected to facilitate the text understanding more. (iii) Unlike entities with a large coverage in Wikipedia, many events do not have a record in Wikipedia. Considering the sparsity, we require models to tag event mentions that do not exist in Wikipedia as "Nil".

Challenges specific to Event Linking. (i) The correct title for some event mentions may not be unique. The same event could be introduced in several pages. For example, "Invasion of Poland" and "Occupation of Poland (1939–1945)" both introduce the event that German Army invaded Poland in 1939. How to decide the ground truth set and how to evaluate in this situation are not trivial.

(ii) Events may only exist in the subsection of

Figure 3: Example of hierarchical events. Event "draft" of Tom Brady is mentioned in the page "Tom Brady", and is also a sub-event of "2000 NFL Draft", which is again a sub-event of "National Football League Draft".

the Wikipedia page. Only a limited number of famous events have their own pages, while many other relatively infamous events only exist in the subsection of some pages. Considering the example in Figure 2, the event "rebuilt" of the Manhattan Bridge does not have its own Wikipedia page, but it is mentioned in the subsection "Reconstruction" of the page "Manhattan Bridge". Linking these events requires a model to understand the whole page instead of just encoding the first paragraph.

(iii) Events have a hierarchical structure. Events at larger granularity consist of many sub-events, and these sub-events may have their own Wikipedia pages, or just be mentioned in the pages of the large events. Ideally, the model should always link the event mention to the most appropriate page. If the sub-event page exists, then link to the sub-event page. Otherwise, link to the page of the large event. However, the term "appropriate" here could be unclear because of the event hierarchy. As Figure 3 shows, the Wikipedia page "Tom Brady" is most specific to the event "drafted". On the other hand, draft of "Tom Brady" is a sub-event of "2000 NFL Draft", which is further a sub-event of "National Football League Draft". Annotators prefer to link this event to "Tom Brady", while Wikipedia hyperlinks link the event to "National Football League Draft". The hierarchy of events makes the standard of the correct title inconsistent.

Evaluation of Event Linking. For any event mention, a system is expected to label it with the correct Wikipedia page or a "Nil" tag. Accuracy is adopted as the official evaluation metric.

3 Data Construction

We collect training data and in-domain test data from Wikipedia automatically, and manually annotate an evaluation set in the news domain for out-of-domain evaluation purpose.

3.1 Wikipedia

We first collect all hyperlinks $(hypertext, title)$ in Wikipedia text, which links a hypertext to a Wikipedia title. Then, we map the FreeBase type of Wikipedia titles to FIGER types (Ling and Weld, 2012), and all titles with a type "Event" are regarded as event titles. All the hypertexts linked to these event titles are regarded as event mentions.

Because same event mentions in one Wikipedia page are hyperlinked only once, and editors tend to hyperlink more nominal mentions than verb mentions, verb mentions are highly limited in Wikipedia. To balance the size of verbs and nominals, we use SpaCy Part-of-Speech model[1] to keep all verb mentions, and sample the same size of nominals. To prevent models from overfitting, we design hard and easy cases for verbs and nominals:

Verbs: We classify each verb mention mainly by whether the surface form (S) of the verb is seen in training data, and whether the gold event title (T) is seen in training data. If both S and T are seen in training data, we call it **Seen Event**. If T is seen in training data, but S is new, we call it **Unseen Form**. If T is never seen in training data, we call it **Unseen Event**. Under this setting, "Seen Event" is regarded as easy cases, and the other two are hard cases. Because of the limited size of verb mentions, all the event titles with fewer than or equal to 5 verb mentions are used as "Unseen Event".

Nominals: We classify each nominal mention mainly by its surface form similarity to the gold title. We calculate the Jaccard similarity between the nominal mention and the gold title by taking 3 grams of the surface form. If the similarity is lower than 0.1, we think it is a **hard nominal**; otherwise, it is an **easy nominal**. Then we sample same numbers of hard and easy cases.

3.2 New York Times

We sample 2,500 lead paragraphs from The New York Times Annotated Corpus (Sandhaus, 2008),

[1] https://spacy.io/usage/linguistic-features#pos-tagging

	Event mention in local context	Wikipedia title
Wiki	At the start of the wartime 1940s , he had four releases.	World War II
	Henry Louis Gates, a black Harvard University professor who was arrested after police mistakenly thought he was breaking into his own home in Cambridge, Massachusetts.	Henry Louis Gates arrest controversy
	Ibrox hosted four Scotland games in the first phase, starting with a 1994 World Cup qualifier against Portugal in October 1992.	1994 FIFA World Cup qualification
NYT	The move is in retaliation for efforts by China to get around American limits on imports by shipping goods through other nations	China–United States trade war
	A man who killed his former wife, a bartender and a cook in 1984 was executed by injection early today.	Godinez v. Moran
	A 45-year-old fashion photographer was shot and killed in his West Village apartment yesterday morning, the police said.	Nil

Table 1: Data examples. The upper part is data collected from Wikipedia hyperlinks. The lower part is annotated New York Times (NYT) paragraphs. Event mentions are highlighted in red.

	Train	Dev	Test	
	Wiki	Wiki	Wiki	NYT
Verb	33,213	8,346	9,633	1,319
Seen Event	-	1,814	2,913	0
Unseen Form	-	2,585	3,828	75
Unseen Event	-	3,947	2,892	435
Nil	-	-	-	809
Nominal	33,213	8,346	9,633	443
Hard	-	4173	4817	244
Easy	-	4173	4817	15
Nil	-	-	-	184
Total	66,426	16,692	19,266	1,762

Table 2: Wikipedia and New York Times (NYT) data statistics. NYT is only for testing.

which contains New York Times articles from 1987 to 2006. We first use an off-the-shelf verb and nominal SRL model[2] to extract event mention candidates, and then we use Amazon Mechanical Turk to annotate the corresponding Wikipedia title of the predicted mention candidates. To ensure the quality of the annotation, we design our annotation process in two rounds:

First round. Annotators need to answer whether they think the predicted mention is an event or not. If they think it is an event, then they need to find the corresponding Wikipedia title, otherwise submit "Nil". Each mention is annotated by three annotators. If all of them submit "Nil", we include this event mention as a "Nil" example in the final test data. *To prevent annotators from simply submitting "Nil", 10% of the event mentions are the relatively easy cases from the Wiki data and we know their answers. We randomly insert them into the input data for AMturk (i.e., annotators are unaware of that) to evaluate the accuracy of the annotator. Only the annotation from annotators with an accuracy higher than 90% will be accepted.*

Second round. This round verifies the annotated results in the first round. Each mention with the annotated title is verified by another three annotators, and they need to read the page, and figure out whether it introduces the mention. If the majority vote for "yes", we include it in the final test data.

Table 1 lists some data examples, and Table 2 shows detailed statistics.

3.3 Domain Analysis

Event linking in the news domain is more challenging than that in the Wikipedia domain because of the following reasons:

(i) News articles describe an event at a different granularity as how Wikipedia does, usually with more details. For example, here is a piece of news about "Iraq_War": "A contractor working for the American firm Kellogg Brown & Root was wounded in a mortar attack in Baghdad." The event "wounded" here is a very small event in Iraq War, but it is what daily news would report. On the other hand, the event mention that links to "Iraq_War" in Wikipedia domain is: "When touring in Europe, the US went to war in Iraq." The different granularity in representing events makes the task slightly different in two domains. Event linking in Wikipedia domain is more like event coreference, while event linking in news domain is mixed with more sub-event relation extraction.

(ii) As analyzed in Section 2, event linking is challenging because some event mentions may only exist in the subsection of the correct page, and the correct title is not consistent because of the event hierarchy. However, these problems mainly happen in the news domain. First of all, the Wikipedia

[2] https://cogcomp.seas.upenn.edu/page/demo_view/SRLEnglish

hyperlinked mentions usually have their own pages instead of just existing in subsections. In news domain, we annotate events that only exist in subsections of a Wikipedia page. Second, in Wikipedia domain, the gold title of same event mentions is usually consistent. For example, all of the event mentions "drafted" of football players link to "National Football League Draft" instead of the page of the specific player. However, the annotation standard of NYT is not always consistent with Wikipedia hyperlinks. For example, annotators would link event mentions about sports player draft to the page of the specific player instead of the general concept page "National Football League Draft". These problems make data annotation and model evaluation in news domain very challenging.

Because of the reasons claimed above, we think that, for some cases in news domain, the correct answer is multiple titles instead of just one title. Ranking the annotated title to the second place may be because the top one is also correct. To relax the evaluation metric here, for news domain, we also report the number of Accuracy@5, which means that if the annotated title is ranked in the top 5 candidates, we think it is correct.

4 Model

In this section, we propose EveLink as the first event linking model. We first introduce the representation of event mentions and event titles in Section 4.1, and then introduce the model architecture in Section 4.2.

4.1 Event Representation

A key difference between entity and event is that the context of an entity is more diverse than the context of an event. For example, when the entity "China" is mentioned in a sentence, it is unclear what entities or what events probably would also be mentioned together. However, if a verb like "invade" is used to represent the event "Battle of France" in a sentence, it is very likely that entities like "Germany", "Italy" and "France" will also be mentioned. This shows that an event is defined by its arguments, and these arguments, with a large chance, will also be mentioned in the local context because the verb itself cannot refer to any event. Given this observation, we think that the entities in the local context of the event mention should overlap with the entities in the correct Wikipedia page, and these entities can be used to help the model better represent events.

Event mentions: To represent event mentions in local context, we first use an off-the-shelf Named Entity Recognition model [3] trained on 18-type OntoNotes dataset (Weischedel et al., 2013) to extract the entities around the event. We simply define the context window by 500 characters around the event mention. After predicting all the entities e_i with their type t_i, we represent the event mentions by:

$$r_1 = [\mathrm{CLS}]\ \mathrm{ctxt}_l\ [\mathrm{M}_s]\ \mathrm{m}\ [\mathrm{M}_e]\ \mathrm{ctxt}_r \quad (1)$$

$$r_2 = [\mathrm{t1}_s]\ \mathrm{e}_1\ [\mathrm{t1}_e]\ \cdots [\mathrm{t}_{n_s}]\ \mathrm{e}_n\ [\mathrm{t}_{n_e}] \quad (2)$$

$$r_m = r_1\ [\mathrm{SEP}]\ r_2\ [\mathrm{SEP}] \quad (3)$$

where m, ctxt_l, ctxt_r, e_i are tokens of event mention, the context on the left of the mention, the context on the right of the mention and predicted entities. $[\mathrm{M}_s]$ and $[\mathrm{M}_e]$ are special tokens to tag the start and end of the event mention. $[\mathrm{t}_{i_s}]$ and $[\mathrm{t}_{i_e}]$ are special tokens to tag the start and end of the entity whose type is t_i. r_m is the final representation of event mentions.

Title: To represent Wikipedia titles, since important entities are already hyperlinked in the page contents, we take the first ten hyperlinked spans as entities, and represent the title by:

$$r_3 = [\mathrm{CLS}]\ \mathrm{title}\ [\mathrm{TITLE}]\ \mathrm{description} \quad (4)$$

$$r_4 = \mathrm{h}_1\ [\mathrm{SEP}]\ \mathrm{h}_2\ [\mathrm{SEP}]\ \cdots\ [\mathrm{SEP}]\ \mathrm{h}_n \quad (5)$$

$$r_t = r_3\ [\mathrm{SEP}]\ r_4\ [\mathrm{SEP}] \quad (6)$$

where title, h_i and $\mathrm{description}$ are tokens of the title, hyperlinked spans, and the content of the Wikipedia page. We simply take the first $2,000$ characters as the description. [TITLE] is the special token to separate the title and the description. r_t is the final representation of Wikipedia titles.

4.2 Model Architecture

Similar to Wu et al. (2020), we first use an bi-encoder architecture to efficiently generate candidates, and use a cross-encoder architecture, which requires more computations, to rank the candidates.

Candidate Generation We use an bi-encoder architecture to train the candidate generation model. We use two independent BERT transformers (Devlin et al., 2019) to encode the representation of

[3] https://cogcomp.seas.upenn.edu/page/demo_view/NEREnglish

Models	Verb				Nominal			Verb + Nominal
	Seen	Unseen Form	Unseen	Overall	Hard	Easy	Overall	
BLINK-Entity	88.91	69.64	62.31	73.37	67.93	95.20	81.57	77.42
BLINK-Event	80.88	85.84	84.54	83.95	79.39	89.10	84.24	84.10
EVELINK	**93.99**	**92.74**	**93.91**	**93.47**	**89.79**	**95.52**	**92.65**	**93.06**

Table 3: Recall on Wikipedia Test. "Seen" means both the surface forms of the mention and the gold title are seen in training. "Unseen Form" means the surface form of the mention is new, but the gold title is seen in training. "Unseen" means that the gold title is unseen in training. BLINK-Entity is the original BLINK model trained on entity linking dataset. BLINK-Event is trained on the new event linking dataset. More details in Section 5

Models	Verb				Nominal			Verb + Nominal
	Seen	Unseen Form	Unseen	Overall	Hard	Easy	Overall	
BLINK-Entity	64.13	48.56	45.92	52.48	46.79	88.27	67.53	60.00
BLINK-Event	77.72	69.78	62.72	70.06	62.59	82.29	72.44	71.25
EVELINK	**91.11**	**79.60**	**77.87**	**82.56**	**75.57**	**89.60**	**82.58**	**82.57**

Table 4: Accuracy on Wikipedia Test.

event mentions r_m and Wikipedia titles r_t, and use the output of the two [CLS] tokens in r_m and r_t as the event mention vector v_m and the title vector v_t. Then, we maximize the dot product between the vectors of event mentions v_m and the correct title v_t in a batch with randomly selected negatives. At inference time, representations of all the titles are cached, and for each event mention, we calculate the dot products between its representation and the representation of all the titles, and titles with higher scores will become candidates.

Candidate Ranking For each event mention, we take 30 candidates from the candidate generation model as the training data for the ranking model, and use a cross-encoder architecture to train the candidate ranking model. We concatenate the representation of event mentions r_m and titles r_t, use one BERT transformer to encode the concatenated representation, and use the output of the [CLS] token as the final vector v.Then we maximize the dot product between the vector v of the concatenated representation with the correct title and an additional linear layer W from other 29 candidates.

5 Experiments

We report the in-domain performance on Wikipedia test set in 5.1 and the out-of-domain performance on news data in 5.2, and conduct an analysis in 5.3. Implementation details are in Appendix A.

Baselines Since there is no existing event linking system, we have to compare with previous entity linking systems. In this paper, we mainly compare our system with one of the SOTA entity linking model BLINK (Wu et al., 2020). To make a fair comparison, BLINK has the following two setups:

BLINK-Entity: Since a large portion of event mentions are nominals, which is also a kind of entity, it would be interesting to see how a SOTA entity linking system performs for event linking. Therefore, we test the BLINK pretrained specific to entity linking in (Wu et al., 2020) directly. Please note that the size of entity linking training data is 9 million, which is much larger than the size of event linking training data 66k.

BLINK-Event: It adopts the same algorithm with the original BLINK system, but is trained on our event linking training set.

For all the experiments, BLINK-Entity retrieves 10 candidates from candidate generation, and both BLINK-Event and EVELINK retrieves 100 candidates from candidate generation. These numbers are tuned on development data.

5.1 In-domain experiment on Wikipedia

In this section, we evaluate EVELINK on the Wikipedia test set as the in-domain performance. We report the recall of candidate generation in Table 3, and the accuracy of candidate ranking in Table 4. As shown in Table 3 and Table 4, EVELINK outperforms baseline models by a large margin, 8.96 points in Recall and 11.32 points in Accuracy. EVELINK also achieves a high performance on seen verbs and easy nominals, around 90 accuracy, but a relatively low performance on other hard cases, which leaves a large space for future

Models	Verb				Nominal			Verb + Nominal
	Seen	Unseen Form	Unseen	Overall	Hard	Easy	Overall	
BLINK-Entity	-	4.00	6.67	6.27	7.79	60.00	10.81	7.80
BLINK-Event	-	35.14	37.39	37.06	37.45	75.00	39.77	37.97
EVELINK	-	**52.70**	**59.40**	**58.43**	**51.03**	**93.75**	**53.68**	**56.83**

Table 5: Recall on New York Times data. Because "Nil" mentions do not have the Wikipedia title, the Recall is only evaluated on the mentions that exist in Wikipedia.

Models	Verb			Nominal			Verb + Nominal	
	Unseen Form	Unseen	Overall	Hard	Easy	Overall	Accu@5	Accu@1
BLINK-Entity	1.33	2.76	2.55	4.92	33.33	6.56	11.44	3.90
BLINK-Event	17.57	5.28	7.06	11.11	37.50	12.74	17.04	8.97
EVELINK	**25.68**	**10.78**	**12.94**	**13.17**	**56.25**	**15.83**	**27.57**	**13.91**

Table 6: Accuracy on New York Times data without Nil. Only event mentions that exist in Wikipedia are given. Accu@5 means the correct title is ranked top 5. Accu@1 means the correct title is top 1.

works to further improve.

5.2 Out-of-domain experiment on News

In this section, we evaluate EVELINK on the NYT test set as the out-of-domain performance. In Table 5, we evaluate the recall of candidate generation. Because "Nil" mentions do not have correct titles in Wikipedia, we only evaluate the the recall of event mentions that exist in Wikipedia. Though the recall of EVELINK is much higher than the recall of other baseline models (56.83 vs. 37.97), the recall drops significantly compared with the recall on Wikipedia test set (56.83 vs. 93.06). In Table 6, we only evaluate the accuracy on the event mentions that exist in Wikipedia, which is the same setting as the experiments in Wikipedia domain, and again the accuracy drops significantly from 82.57 to 13.91. Even if we accept 5 predictions instead of just one to solve the multiple correct titles problem, the Accuracy@5 is 27.57, which is still low. Detailed error analysis is in Section 5.3. In Table 7, we evaluate the accuracy of all the event mentions, including Nil. Because we do not have Nil examples in training data and development data, we simply predict all the event mentions with probability lower than 50 to "Nil", and leave better solutions to future works.

5.3 Analysis

In this section, we do an analysis for our approach EVELINK. We wonder following questions:

$\mathcal{Q}_1$: Where the gain comes from, compared with the BLINK system?
We do ablation study in Table 8. Explicitly adding entities to the event representation boosts the performance by 10.14 accuracy on Wikipedia test data and 2.73 accuracy on NYT data. Adding entity types further improves the performance by 1.18 accuracy on Wikipedia test data and 2.21 accuracy on NYT data.

$\mathcal{Q}_2$: Error patterns of EVELINK
We collect several error patterns that are common in both domains, and patterns that are mostly in news domains. Error patterns of both domains:

(i): Repeating events. In the errors, we find many repeating events, like award ceremonies or sports games, that would happen every several years, and the model usually cannot find the correct year of the event if the year is not explicitly mentioned in the context. For example:

> In 1995, his debut season, Biddiscombe made two appearances, ··· The following year he earned a Rising Star nomination for his performance···

In this example, the gold event is "1996 AFL Rising Star", and the prediction is "1998 AFL Rising Star", though there is a temporal hint (the following year of 1995 is 1996) to indicate that the correct answer should be the award in 1996. There are many similar errors when linking awards or games, which shows that a deeper understanding of time is necessary for future works.

(ii): Unrelated context. EVELINK replies on the surrounding entities to link the event mentions, however, the context is not always related and surrounding entities cannot help linking. For example:

Models	Verb				Nominal				Verb+Nominal	
	Unseen Form	Unseen	Nil	Overall	Hard	Easy	Nil	Overall	Accu@5	Accu@1
BLINK-Entity	2.7	1.15	79.85	49.51	1.23	25.00	63.04	27.77	57.26	44.04
BLINK-Event	12.16	1.61	90.85	56.94	4.53	37.50	88.59	40.63	58.45	52.84
EVELINK	**16.21**	**3.67**	**91.7**	**58.38**	**7.82**	**50.00**	86.41	**41.99**	**59.70**	**54.26**

Table 7: Accuracy on New York Times data with Nil. We simply predict all the mentions with a probability lower than 50 to Nil.

Models	Wiki Test	NYT (no Nil)	NYT
EVELINK	82.57	13.91	54.26
- type	81.39	11.70	55.96
- entities	71.25	8.97	52.84

Table 8: Ablation Study of EVELINK

> Returning to his country at the end of the conflict and another begun, Barinaga rejected an offer from Athletic Bilbao, moving to Real Madrid instead.

In this example, the gold event is "World War II", but the prediction is "1939–40 La Liga". All the entities, like "Barinaga", "Athletic Bilbao" and "Real Madrid", are about football, which is unrelated to the war. To link to the correct page, the model needs to know the second conflict of Barinaga's country, which indicates that only using the local context maybe not enough.

Error patterns specific to news domain:

(i): Subsection events. Some events do not have their own pages, and are only introduced in the subsections of other pages. For example:

> The Philippine government lifted its five - year ban on the return of Imelda Marcos today and said the widow of the late President Ferdinand Marcos was free to come home from exile in the United States.

In this example, the return of Imelda Marcos is introduced in the subsection "Return from exile (1991–present)" of the page "Imelda Marcos". However, we only use the first 2,000 characters of the page contents to represent the title "Imelda Marcos", which has no information about the return from exile. A document-level representation may be a potential solution for future works.

(ii): Sub-events. Some events are sub-events of other larger events. For example:

> Stepping in at the 11th hour, Hillary Rodham Clinton will campaign in Florida on Saturday for her brother, Hugh Rodham, in his bid for a United States Senate seat.

This event is a sub-event of "1994 United States Senate election in Florida", and this event is also mentioned in the contents of the page. However, the names in the local context do not overlap with the names in the first paragraph of the page.

In this work, we discuss many challenges of the task in different domains, but EVELINK cannot address all of them. We leave them to future works.

6 Related Work

"Event Linking" Nothman et al. (2012) use the term "event linking" to define the task that links the event mention to the place it is first mentioned, which is a sub-task of cross-document event coreference. We define "event linking" by grounding mentions to Wikipedia.

Model Architecture Humeau et al. (2019) and Wu et al. (2020) use a bi-encoder architecture to train the candidate generation model, and a cross-encoder architecture to train the candidate ranking model for entity linking. Considering the structure of events that entities do not have, EVELINK extends their model by adding entity information to the event mention representation.

Event Representation Vyas and Ballesteros (2021) use similar methods to add entity attributes to the entity representation as our method of adding entities to the event representation. The difference is that we also add entity type information to the representation by using special tokens to indicate the start and end of entities in different types.

Data Eirew et al. (2021) collect training data from Wikipedia hyperlinks for event coreference, while we use similar methods to collect data for event linking. In this work, we use the FIGER type of the title to find event titles, while Eirew et al. (2021) use the Wikipedia infobox.

7 Conclusion

In this work, we formulate Event Linking, a challenging but essential task, with a carefully designed Wikipedia dataset and NYT test set, and propose an event linking model EVELINK for future works.

References

Razvan Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Alon Eirew, Arie Cattan, and Ido Dagan. 2021. WEC: Deriving a large-scale cross-document event coreference dataset from Wikipedia. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2498–2510, Online. Association for Computational Linguistics.

Nitish Gupta, Sameer Singh, and Dan Roth. 2017. Entity linking via joint encoding of types, descriptions, and context. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2681–2690, Copenhagen, Denmark. Association for Computational Linguistics.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*.

Sebastian Krause, Feiyu Xu, Hans Uszkoreit, and Dirk Weissenborn. 2016. Event linking with sentential features from convolutional neural networks. In *Proceedings of CoNLL*, pages 239–249.

Xiao Ling and Daniel S Weld. 2012. Fine-Grained Entity Recognition. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking documents to encyclopedic knowledge. In *Proc. of the ACM Conference on Information and Knowledge Management (CIKM)*.

Joel Nothman, Matthew Honnibal, Ben Hachey, and James R. Curran. 2012. Event linking: Grounding event reference in a news archive. In *Proceedings of ACL*, pages 228–232.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384, Portland, Oregon, USA. Association for Computational Linguistics.

Evan Sandhaus. 2008. The new york times annotated corpus ldc2008t19. *Web Download*.

Yogarshi Vyas and Miguel Ballesteros. 2021. Linking entities to unseen knowledge bases with arbitrary schemas. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 834–844, Online. Association for Computational Linguistics.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1321–1331, Beijing, China. Association for Computational Linguistics.

A Implementations

We use 4 Nvidia RTX A6000 48GB GPUs for model training and evaluation. For both candidate generation model and candidate ranking model, we train 10 epochs with learning rate $1e^{-5}$, and use BERT-large-uncased as the pretrained language model (Devlin et al., 2019). The maximum tokens of both event mention representation and Wikipedia title representation are 256.

B Data Annotation

We require all the annotators from Amazon Mechanical Turk to be English speaker, and with an acceptance rate higher than 95%.