Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2245–2256
Marseille, 20-25 June 2022

OPENEL: An Annotated Corpus for Entity Linking and Discourse in Open Domain Dialogue

Wen Cui[1], Leanne Rolston[2], Marilyn A. Walker[1], Beth Ann Hockey[2]
[1]University of California, Santa Cruz
{wcui7, mawalker}@ucsc.edu
[2]LivePerson Inc.
{lrolston, bhockey}@liveperson.com

Abstract

Entity linking in dialogue is the task of mapping entity mentions in utterances to a target knowledge base. Prior work on entity linking has mainly focused on well-written articles such as Wikipedia, annotated newswire, or domain-specific datasets. We extend the study of entity linking to open domain dialogue by presenting the OPENEL corpus: an annotated multi-domain corpus for linking entities in natural conversation to Wikidata. Each dialogic utterance, in 179 dialogues over 12 topics from the original EDINA corpus, has been annotated for entities realized by definite referring expressions as well as anaphoric forms such as *he*, *she*, *it* and *they*. OPENEL thus supports training and evaluation of entity linking in open-domain dialogue, as well as analysis of the effect of using dialogue context and anaphora resolution in model training. It can also be used for fine-tuning a coreference resolution algorithm. To the best of our knowledge, this is the first substantial entity linking corpus publicly available for open-domain dialogue. We also establish baselines for named entity linking in open domain conversation using several existing entity linking systems. We find that the Transformer-based system, Flair + BLINK, has the best performance with a 0.65 F1 score. Our results show that dialogue context is extremely beneficial for entity linking in conversations, with Flair + BLINK achieving an F1 of 0.61 without discourse context. These results also demonstrate the remaining performance gap between the baselines and human performance, highlighting the challenges of entity linking in open-domain dialogue, and suggesting many avenues for future research using OPENEL.

Keywords: Entity Linking, Coreference, Discourse Modeling, Wikidata, Open-domain Dialogue

1. Introduction

Named entity recognition (NER), named entity linking (NEL) and discourse modeling (DM) are crucial aspects of natural language understanding (NLU) for open-domain dialogue systems, such as chatbots or socialbots. Unlike task-oriented dialogue systems, open-domain systems need to be able to converse on any topic, with the concomitant challenge of handling a much wider variety of named entities. NEL provides grounding of the named entities in a knowledge base (KB) (in this work Wikidata) by assigning a knowledge base ID to each named entity mention and anaphoric reference. This is demonstrated in the dialogue excerpt in Figure 1, where we see *The Other Woman* introduced by speaker A in A1, and then referred to in B1 as *that*. In A2, speaker A then refers to *Cameron Diaz* and *Leslie Mann* by the possessive referring expression *their*. The speakers continue, throughout the dialogue, to build on entities already added to the discourse model, as they comment further on entities introduced earlier in the dialogue, and introduce related entities, e.g. A3 and A4 introduce new movies, *Knocked Up* and *Big Daddy*, that *Leslie Mann* starred in.

NEL in general is challenging due to name variations. A named entity may have several surface forms, such as partial names, aliases and abbreviations. For example people may refer to the basketball player *Michael Jordan* by his partial name *Michael* or *Jordan*, or by his nickname *His Airness*. Organizations may be referred to by their full name or by an abbreviation, such as NBA for the *National Basketball Association*, or NFL for the *National Football League*.

A1: The Other Woman $_{Q14272676}$ is such a funny movie.
B1: That $_{Q14272676}$'s the one with Cameron Diaz $_{Q44380}$ and Leslie Mann $_{Q229011}$, right?
A2: That's right. Do you enjoy their $_{Q44380,\ Q229011}$ movies?
B2: Oh yes. Leslie Mann $_{Q229011}$ plays neurotic characters very well.
A3: You're right. She $_{Q229011}$ was great in Knocked Up $_{Q222800}$ as Debbie.
B3: I liked her $_{Q229011}$ in Big Daddy $_{Q509025}$, too. That was the first movie I ever saw her $_{Q229011}$ in.
A4: Really? It's been awhile since I watched an Adam Sandler $_{Q132952}$ movie.
B4: I love his $_{Q132952}$ movies. They always make me laugh.
A5: They are pretty funny. Happy Gilmore $_{Q1313063}$ is my favorite Adam Sandler $_{Q132952}$ movie.
B5: That $_{Q1313063}$'s a good one. The cameo by Bob Barker $_{Q381178}$ always cracks me up.

Figure 1: An annotated conversation from the Comedy domain in the Edina corpus, illustrating NEL on definite referring expressions and anaphora. Definite referring expressions are highlighted in blue while referring anaphora are highlighted in orange.

NEL in open-domain systems is inherently **more** challenging due to the open-ended nature of possible topics, and the fact that utterances tend to be short, informal

A1:You psyched about this coming NFL season?
B1:Oh yeah. Can't wait to see my **Giants** in action.

A2: What do you think about Nicki Minaj?
B2: My favorite is **Anaconda**.

Figure 2: potentially ambiguous reference resolved by dialogue context

and ambiguous. Typical open-domain dialogue topics include songs, albums and movies. These topics contain highly ambiguous entity names. *Her* is the name of a song and a movie. *Thank you* is the name of a song, an album and a TV show.

These challenges require the NEL system to leverage the dialogue context and entity information such as its type, alias and description. Figure 2 shows several cases where a reference is disambiguated by discourse context. The reference to the NFL in A1 helps determine that the entity mention *Giants* in B1 refers to the *New York Giants* (a football team) instead of the *San Francisco Giants* (a baseball team). Utterance B2 demonstrates how the mention of the singer *Nicki Minaj* disambiguates the entity mention *Anaconda*, where *Anaconda* could refer to the song, the 1997 horror film, the city in Montana, or the software.

Typically entity linking is a 3-stage pipeline system that performs mention detection, candidate generation, and entity disambiguation. Specifically, in B1 in Figure 2, mention detection identifies the span of the surface form *Giants*. Then candidate generation takes the mention *Giants* and generates plausible entity candidates from the KB such as *New York Giants*, *San Francisco Giants*, *Yomiuri Giants*, *Giants (the album)* and *Giants (the comic book)*. Finally, the entity disambiguation utilizes a trained ranker that takes the mention and the list of candidates and uses contextual information to rank the candidates. The desired result, the *New York Giants*, should then be promoted to the top of the candidate list.

Table 1 provides an overview of the OPENELcorpus. Our goal in this paper is to advance research in NEL for open-domain dialogue systems by 1) releasing OPENEL, a high quality entity-enriched corpus of annotated dialogues with both NEL and anaphora annotations, as illustrated in Figure 1; 2) analyzing NEL and anaphora annotation quality (Section 3); 3) comparing existing NEL tools and establishing baselines; and 4) examining the domain coverage in NEL systems (Section 4.1 and Section 4.2). The corpus can be downloaded here `https://github.com/wenzi3241/OpenEL_corpus`.

2. Related Work

EL has been extensively studied in the past decade.

Corpus Properties: Overall	
conversations	179
conversational turns	2,570
topics	12
entity mentions (incl. anaphors)	2,263
entity mentions (excl. anaphors)	1,205
anaphors	1,058
unique entities	576
average conversation length (turns)	14.4
average mentions per entity	3.9
average mentions per conversational turn	0.9

Table 1: Counts and averages representing the composition of the corpus

NEL Corpora. There are many EL benchmark datasets constructed from written texts. For example, AIDA-CoNLL (Hoffart et al., 2011), ACE2004 (Ratinov et al., 2011), TAC-KBP2010 (Ji and Grishman, 2011), AQUAINT (Milne and Witten, 2008) and MSNBC (Cucerzan, 2007) are annotated news articles; Wiki-Disamb30 (Ferragina and Scaiella, 2010) and WNED-WIKI (Guo and Barbosa, 2018) are extracted from Wikipedia articles; and T-REx (Elsahar et al., 2018) uses DBpedia abstracts. There are also datasets that annotate noisy social media data such as Microposts2014 (Cano et al., 2014) and the Reddit Entity Linking dataset (Botzer et al., 2021).

However, open-domain chitchat differs from edited written texts and social media data in having utterances that are informal, multi-turn and highly ambiguous. To the best of our knowledge, as of this time there are only two other open-domain NEL datasets. The only other open-domain dialogue NEL corpus that we are aware of consists of ConEL (Joko et al., 2021), a much smaller corpus of 25 conversations sampled from Wizard of Wikipedia (Dinan et al., 2018) with only 33

Corpus Properties: By Topic			
	# Convs	**# Turns**	**% Entity**
music	33	20	0.60 / 0.40
pop	32	10	0.82 / 0.60
star wars	30	10	0.72 / 0.62
baseball	34	20	0.53 / 0.42
comedy	22	10	0.78 / 0.61
rap hiphop	5	10	0.76 / 0.68
action	4	10	0.72 / 0.55
basketball	4	20	0.68 / 0.60
horror	4	10	0.53 / 0.30
movies	3	20	0.80 / 0.55
nfl football	4	20	0.51 / 0.37
rock	4	10	0.78 / 0.63

Table 2: Number of conversations per topic, number of turns per conversation, proportion of turns with an entity (incl anaphors / excl anaphors).

unique annotated named entities. The occurrence of named entities is sparse, as pointed out by the authors, and a large portion of mentions are personal (i.e. my guitar) and conceptual (i.e tattoo). This corpus thus is not yet large enough to be useful as a corpus for training or testing NEL systems on such datasets.

Unfortunately, other work in this area sometimes do not release the data at all. For example, the MOC dataset (Shang et al., 2021), is a crowdsourced dataset with a total of 7,735 utterances across 8 popular topics. it is currently not publicly available and the annotation quality is unknown.

Existing NEL systems. Pipeline NEL systems (Hoffart et al., 2011; Ferragina and Scaiella, 2010; Mendes et al., 2011; Piccinno and Ferragina, 2014; van Hulst et al., 2020) solve mention detection, candidate generation and entity disambiguation as subtasks, whereas other systems (Durrett and Klein, 2014; Kolitsas et al., 2018; Li et al., 2020; De Cao et al., 2020; Broscheit, 2020; Prabhakar Kannan Ravi et al., 2021) jointly model these subtasks and exploit the interdependency among them. Various machine learning (ML) and deep learning (DL) techniques have been applied. Research is also being done into using zero-shot and few-shot approaches, leveraging pre-trained language models such as GPT3 (Brown et al., 2020), BERT (Devlin et al., 2018) and its variants, to perform NEL (Logeswaran et al., 2019; Li et al., 2020; Tang et al., 2021; Wu et al., 2019a).

The recent prompt-based learning diagram approach (Brown et al., 2020) using large pre-trained language models has been successfully applied to many NLP tasks. Prior work has first applied prompt-based learning to entity linking by leveraging BERT and appropriate prompts to achieve decent performance on NEL (Sun et al., 2021).

However, although there has been considerable research on NEL, few attempts have been made to tackle NEL in open-domain dialogue. Participants in the Alexa Prize Competition (Bowden et al., 2018; Curry et al., 2018) leverage existing tools and develop heuristic filtering with consideration of dialogue context. Other work (Shang et al., 2021) applies DL techniques and focuses on modeling dialogue context. It encodes context with BiLSTM (Chiu and Nichols, 2016) and BERT in the NER and NEL stages respectively, and shows significant gains in model performance.

3. The OpenELCorpus

We annotate a sample of the EDINA corpus (Krause et al., 2017) for named entities and named entity linking using Wikidata IDs. The EDINA corpus was collected through a “self-dialogue” approach that has Amazon Mechanical Turk workers create both sides of a conversation. Although this is clearly artificial data, this method was successful in producing quite natural conversations. The main advantages of the EDINA data for our work is that the conversations are fluent, have a wide variety of entity types, and have a high density of named entity mentions across a range of typical chat domains, which are shown in Table 2. The total number of entity mentions (excluding anaphors) is 1,205 and Each conversational turn has an average 0f 0.9 entity mentions, and the average mentions per named entity is 3.9. Our corpus consists of 179 conversations selected from 12 domains of the EDINA corpus. Corpus properties are given in Tables 1 and 2.

3.1. Annotation

The OpenELsubset of the EDINA corpus was first preprocessed using DBpedia Spotlight (Mendes et al., 2011), and then three experienced annotators verified or corrected spans and Wikidata IDs for each entity mention identified by Spotlight. Entity mentions or anaphors that were missed by Spotlight were identified by the annotators and tagged with an appropriate Wikidata ID. Figure 3 illustrates an excerpt of two turns of a conversation and their annotations.

”text”:	”Mandy Moore used to sing. Did they ever record together?”,
”speaker-id”:	”A”,
”entities”:	[”annotator-1”: [”span”: [0, 10], ”surface-form”: ”Mandy Moore”, ”wikidata-id”: [”Q187832”], ”is-anaphora”: false, ”span”: [31, 34], ”surface-form”: ”they”, ”wikidata-id”: [”Q160009”, ”Q187832”], ”is-anaphora”: true], ”annotator-2”: [”span”: [0, 10], ”surface-form”: ”Mandy Moore”, ”wikidata-id”: [”Q187832”], ”is-anaphora”: false, ”span”: [31, 34], ”surface-form”: ”they”, ”wikidata-id”: [”Q160009”,”Q187832”], ”is-anaphora”: true], ”annotator-3”: [”span”: [0, 10], ”surface-form”: ”Mandy Moore”, ”wikidata-id”: [”Q187832”], ”is-anaphora”: false, ”span”: [31, 34], ”surface-form”: ”they”, ”wikidata-id”: [”Q160009”], ”is-anaphora”: true], ”ground-truth”: [”span”: [0, 10], ”surface-form”: ”Mandy Moore”, ”wikidata-id”: [”Q187832”], ”is-anaphora”: false, {”span”: [31, 34], ”surface-form”: ”they”, ”wikidata-id”: [”Q160009”, ”Q187832”], ”is-anaphora”: true]}}
”text”:	”I’m not sure, but I don’t think so.”,
”speaker-id”:	”B”,
”entities”:	”annotator-1”: [], ”annotator-2”: [], ”annotator-3”: [], ”ground-truth”: []

Figure 3: JSON formatted example in OpenEL.

For inter-annotator agreement, we report both pairwise Cohen’s Kappa and F-Scores (following (Deleger et al., 2012)) and percentage overlaps. These are summarized in Tables 3 and 4. Our agreement is considered to be nearly perfect as the average pairwise Cohen's Kappa of span and Wikidata ID agreement is 0.81.

Pairwise Annotator Agreement		
	Cohen Kappa	F-Score
Span & Wikidata ID	0.79	0.61
	0.82	0.67
	0.82	0.68
(average)	0.81	0.67
Span	0.83	0.73
	0.86	0.74
	0.84	0.75
(average)	0.84	0.74
Wikidata ID	0.85	0.76
	0.86	0.77
	0.89	0.83
(average)	0.87	0.79

Table 3: Pairwise annotator agreement: Cohen Kappa and F-scores

% Annotator Agreement		
all annotators agree		
	exact	overlapping
Span and Wikidata ID	0.67	0.75
Span	0.73	0.82
Wikidata ID	0.75	NA
2 or more annotators agree		
	exact	overlapping
Span & Wikidata ID	0.87	0.95
Span	0.90	0.99
Wikidata ID	0.89	NA

Table 4: Percent annotator agreement using strict and loose criteria across 2 dimensions: unanimous agreement vs 2 or more, and exact span match vs partial or overlapping span match.

There are limitations to NEL which were evident in the annotation process. Knowledge sources like Wikidata are neither infinite nor perfect. In the corpus, we had named entities for which there was no Wikidata ID and there were entities for which there was no unique Wikidata ID. An example of the latter is when a speaker asks about the "Yankees vs Redsox game". There is clearly a particular game that is being referenced (though this cannot be established without knowing exactly when the dialogue occurred), however since there is only a Wikidata ID for the season to which that game belonged, if the conversation went on to discuss this game as opposed to some other game in the same season, the season level Wikidata ID would not differentiate the two. In annotating the corpus we did not assign a Wikidata ID in these cases.

Annotation	**D-type**
(Star Wars topic)	
A1:Do you like the star wars movies $_{\mathit{Q22092344}}$?	
A2:Do you like the star wars movies $_{\mathit{Q462}}$?	
A3:Do you like the star wars $_{\mathit{Q462}}$ movies?	mention, link
(Star Wars topic)	
A1:The forrest planet $_{\mathit{Q832100}}$?	
A2:The forrest planet $_{\mathit{Q12180673}}$?	
A3:The forrest $_{\mathit{Q12180673}}$ planet?	anaphora, link
(Baseball topic)	
A1:Wouldn't think of missing it. I never miss that $_{\mathit{Q213417}}$ matchup.	
A2:Wouldn't think of missing it. I never miss that matchup $_{\mathit{Q213417}}$.	
A3:Wouldn't think of missing it. I never miss that matchup.	anaphora

Table 5: Annotation disagreement examples among three annotators. The annotation of mention is highlighted in yellow and the Wikidata ID is in *italics*. The last column **D-type** indicates the disagreement type.

3.2. Annotator Disagreement

The annotator agreement of our corpus is very high, however annotators did not completely agree on 13% of the annotations. The disagreements came in three aspects: 1) mention disagreement, 2) link disagreement and 3) anaphora disagreement. We show examples of these types of disagreement in Table 5. In the first example the mention *star wars movies* was linked differently: as Q22092344 (the movie series) and Q462 (the media franchise). There was also disagreement on the text span: the third annotator marked only *star wars* as the mention, while the other two marked the entire span *star wars movies*. The third example shows three-way span disagreement where two annotators chose different text spans while the third did not choose a span at all.

We postprocessed the annotations to clean up human errors. The postprocessing includes: 1) validating Wikidata IDs; 2) ignoring uncertain mentions or unlinkable entities; 3) removing non-entity tokens such as quotation marks, commas, etc; and 4) correcting missing mention annotation when the annotator annotated one mention but missed another one that is in the same utterance. Furthermore we created a anaphora list to identify anaphoric reference and indicated in the "is_anaphora" field in our data.

We apply majority vote to determine the ground-truth. The ground-truth label therefore is determined by at least two annotators' agreement on the exact same span and Wikidata ID. The data is in JSON format and each

Setting	System Input	Output of the last turn
(Topic=Rap-hiphop, System=Flair + BLINK)		
UTT.	B: My favorite is ***Anaconda***.	(Anaconda, Colt Anaconda, Q1112001)
DIA.	A: What do you think about Nicki Minaj? B: My favorite is ***Anaconda***	(Anaconda, Anaconda, Q17485058) ✓
DIS.	A: What do you think about Nicki Minaj? B:My favorite is ***Anaconda***.	(Anaconda, Anaconda, Q17485058) ✓
(Topic=Baseball, System=WAT)		
UTT.	B: Looks like ***Posey***'s all recovered from his concussion.	(Posey, Posey County, Q6307475)
DIA.	A: Now if Belt will just start hitting . . . B: Dude, if he can keep getting on base by taking walks, I don't care if he never gets another home run. A: A walk's as good as a single. I'll take it. B: You hate to see all that power go to waste, though. A: Yeah. Maybe Bonds could work with him. Give him some pointers. B: ***Posey***'s all recovered from his concussion.	(Posey, James Posey, Q717793)
DIS.	A: Now if Belt will just start hitting . . . B: Dude, if Belt can keep getting on base by taking walks, I don't care if Belt never gets another home run. A: A walk's as good as a single. I'll take it. B: You hate to see all that power go to waste, though. A: Yeah. Maybe Bonds could work with Belt . Give Belt some pointers. B: Looks like ***Posey***'s all recovered from his concussion.	(Posey, Buster Posey, Q971912) ✓

Table 6: Examples of NEL system inputs and outputs on the last utterances in the UTTERANCE (UTT.), DIALOGUE (DIA.) and DISCOURSE (DIS.) settings. Ground-truth anpahoras are replaced with their mentions in the DISCOURSE setting (highlighted in orange). The system output is in the format of (mention, entity title, Wikidata ID). Correct outputs are indicated with a ✓mark. For illustration purpose, we assign A and B to indicate turn exchanges, however they are not a part of the input. Also the dialogue is shortened due to space limitation.

entry represents one conversational turn. The index represents the conversation id joined with utterance id. See an example of the JSON format in Figure 3. Note that two turns are captured in this example. The first, attributed to speaker A, contains entities, while the response, attributed to speaker B, does not. Each annotator's list of entities is represented followed by the majority vote "ground-truth" used in the experiments. Each item in an annotator's entity list is one entity mention with "span", "surface_form", "wikidata_id" and "is_anaphora" fields. Note that the "wikidata_id" field is a list of IDs since anaphora such as *they* could refer to multiple entities.

4. Experiments

4.1. Setup

Prior work (Shang et al., 2021; Lazic et al., 2015; Shang et al., 2021) has shown that dialogue context and entity discourse improve model performance in NER and NEL related tasks. Therefore we test selected NEL systems in three settings in terms of how much context the system has access to. Sample dialogue snippets with ground-truth labels for anaphora and NEL are shown in the first column of Table 6.

- **UTTERANCE**. The input to each system is only one utterance as shown in Table 6), therefore it does not have access to any dialogue context.
- **DIALOGUE**. Unlike performing NEL on written articles or news, a dialogue system has access only to previous context but not any content that comes later. To replicate this situation, the input to each system is all previous dialogue context and the utterance itself.
- **DISCOURSE**. Similar to the DIALOGUE setting, the system has access to all previous dialogue context. Additionally, we resolve anaphoras by substituting them with their actual entity mentions. In the example shown in Table 6, the anaphoras "he" and "him" are replaced with the mention "Belt" (highlighted in orange).

			strict match			weak match		
Model	**#**	**Metric**	**UTT.**	**DIA.**	**DIS.**	**UTT.**	**DIA.**	**DIS.**
	1	P	0.44	†0.43 (-2.3%)	†0.43 (-2.3%)	0.47	†0.46 (-2.1%)	†0.47 (0.0%)
Spotlight	2	R	0.45	†0.44 (-2.2%)	†0.44 (-2.2%)	0.47	†0.46 (-2.1%)	†0.47 (0.0%)
	3	F1	0.44	†0.43 (-2.3%)	†0.43 (-2.3%)	0.47	†0.46 (-2.1%)	†0.47 (0.0%)
	4	P	0.29	0.31 (+6.9%)	0.31 (+6.9%)	0.36	0.38 (+5.6%)	0.38 (+5.6%)
WAT	5	R	0.36	0.50 (+38.9%)	0.51 (+41.7%)	0.44	0.62 (+40.9%)	0.62 (+40.9%)
	6	F1	0.33	0.38 (+15.2%)	0.39 (+21.2%)	0.40	0.47 (+17.5%)	0.47 (+17.5%)
	7	P	0.50	0.57 (+14.0%)	0.57 (+14.0%)	0.57	0.66 (+15.8%)	0.66 (+15.8%)
*REL	8	R	0.39	0.45 (+15.4%)	0.45 (+15.4%)	0.44	0.52 (+18.2%)	0.52 (+18.2%)
	9	F1	0.44	0.50 (+13.6%)	0.51 (+15.9%)	0.50	0.58 (+16.0%)	0.58 (+16.0%)
*Flair	10	P	0.57	0.60 (+5.3%)	0.61 (+7.0%)	0.66	0.70 (+6.1%)	0.71 (+7.6%)
+	11	R	0.49	0.51 (+4.1%)	0.52 (+6.1%)	0.57	0.59 (+3.5%)	0.60 (+5.3%)
BLINK	12	F1	**0.53**	**0.55** (+3.8%)	**0.56** (+5.7%)	**0.61**	**0.64** (+4.9%)	**0.65** (+6.6%)

Table 7: NEL evaluation results in UTTERANCE (UTT.), DIALOGUE (DIA.) and DISCOURSE (DIS.) settings using all of the conversations in OpenEL. Numbers in parenthesis are the relative improvement compared to the UTTERANCE setting within the same metric. And numbers started with † are not statistically significant (p-value $>$ 0.05 with approximate randomization tests) compared to the UTTERANCE setting. The best F1-score evaluated using different settings and matching methods are highlighted in bold. Models indicated by * are DL-based models.

4.2. Competing Systems

We evaluate several well-known traditional ML and DL-based NEL models on our corpus OpenEL and compare their results. We consider publicly available models that do not require extra fine-tuning steps or modifications. [1]

- **DBpedia Spotlight** (Mendes et al., 2011), is a well-established NEL tool that represents the target KB DBpedia in a Vector Space Model. It then links mention by cosine similarity between the context vector of the mention and the candidate in the KB. We set the confidence threshold to 0.5 as suggested in the documentation. Since Spotlight links to DBpedia KB, we mapped DBpedia entities to Wikidata IDs by sending SPARQL queries to DBpedia[2] and manually creating the mapping if the query fails.

- **WAT** (Piccinno and Ferragina, 2014) is another prominent NEL tool based on TagMe (Ferragina and Scaiella, 2010) and links to Wikipedia entities. A support vector machine (SVM) is trained with hand-crafted features for MD, while a voting scheme based on PageRank is proposed for ED. It also employs an additional annotation pruning step that is trained on another SVM. We used its web service[3] together with the Wikipedia API[4] to obtain the Wikidata ID from the Wikipedia entity. The confidence threshold of 0.3 achieved the best F1 on our tests and therefore were used throughout the experiments.

- **REL** (van Hulst et al., 2020) is a pipeline system which uses Flair (Akbik et al., 2018) for NER and a multi-layer perceptron (MLP) with a max-margin loss for ED. It is trained on AIDA with pre-trained word embedding and features such as context similarity, coherence measures and mention relations. We used their API service[5] for testing. The confidence threshold was tuned and set to be 0.2 for all the experiments.

- **Flair + BLINK** BLINK (Wu et al., 2019b) is a zero-shot Transformer based ED system that is fine-tuned on BERT to encode mentions, context and entity description with a linear layer for scoring. It achieved SOTA performance on the benchmark dataset TAC-KBP2010 (Ji and Grishman, 2011). To perform end-to-end NEL, we employed BLINK with Flair for NER [6].

4.3. Evaluation Metrics

Micro-averaged precision, recall and F1-measure are commonly used to evaluate NEL systems. We com-

[1] We also tried GENRE (De Cao et al., 2020) but did not get decent results. We suspected that it depends on customized dictionaries of mentions and candidates.

[2] `http://wikidata.dbpedia.org/OnlineAccess`

[3] `https://services.d4science.org/web/tagme/wat-api`

[4] `https://en.wikipedia.org/wiki/Wikipedia:Finding_a_Wikidata_ID`

[5] `https://github.com/informagi/REL`

[6] The BLINK model is published here `https://github.com/facebookresearch/BLINK`

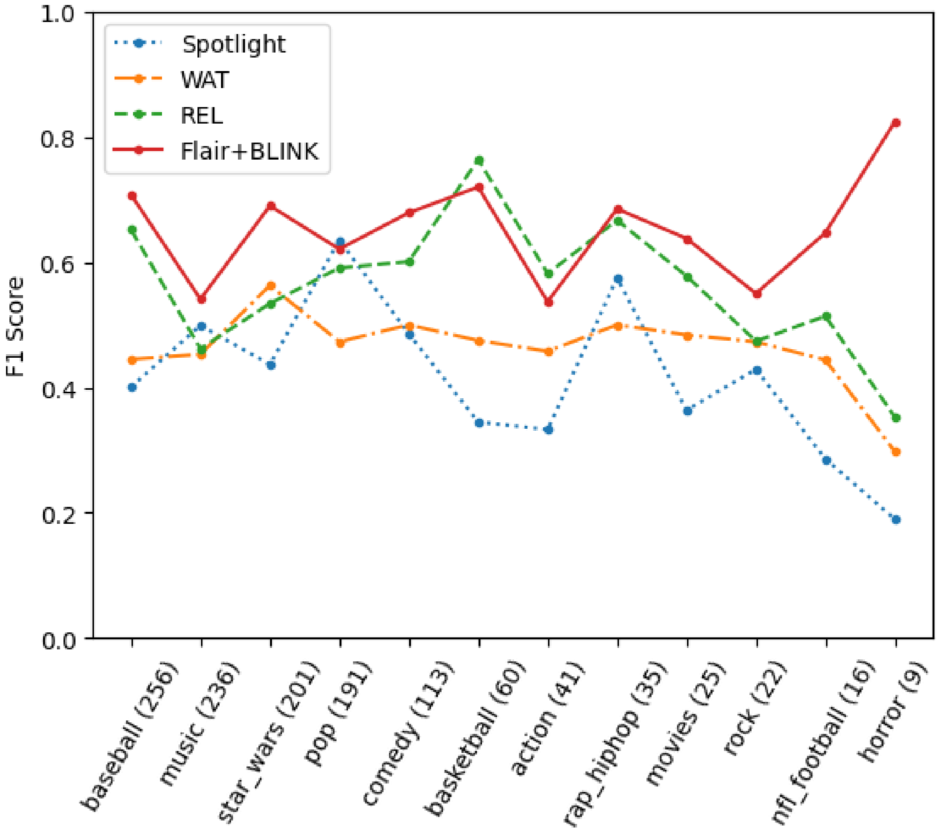

Figure 4: Comparison of F1-score on 12 topics in DISCOURSE setting. Scores are calculated based on weak match. The support of gold entities for each topics are in the parenthesis on the label of X-axis. And topics are sorted in descending order in terms of support.

pare the predicted outputs to the ground-truth labels using both weak and strong matching approaches. Strong matching requires exact matching of text span boundaries *and* a correct Wikidata ID while weak matching accounts for partial span overlap.

4.4. Results & Discussions

We ran baseline systems using all the conversations in OpenEL. Table 7 presents the performance of four baseline systems in all three settings introduced in Section 4.1. From the table, we can see that the Transformer-based system Flair+BLINK performs the best across all metrics (Row 12) with the highest 0.56 F1-score for strict match and 0.65 F1-score for weak match. It is pre-trained on massive data and captures more sophisticated feature representations of NEL. However there still remains a big performance gap compared to its almost SOTA performance on other NEL benchmarks (Wu et al., 2019a; Shen et al., 2021).

Table 7 also suggests the effectiveness of dialogue context in the decision making process of NEL systems. All systems except for Spotlight have a substantial performance gain (numbers in parenthesis as shown in Table 7) that varies from 3.8% to 17.5% in terms of F1-score by feeding the previous dialogue. Another performance boost up to 6% is observed by resolving anaphoras in the dialogue context. WAT has a relative ∼ 40% improvement in recall (Row 5) benefiting from its feature of exploiting context surrounding the mention. REL and Flair+BLINK both use Flair for NER, whereas REL has a higher relative F1-score improvement in both DIALOGUE and DISCOURSE settings (Row 9 and 12).

Since REL explicitly represents mention relations as latent variables proposed by (Le and Titov, 2018) in the ED stage, having access to dialogue context results in a better mention representation as well as linking. On the other hand, Flair+BLINK benefits from its architecture that captures long-distance dependency in language. In contrast, Spotlight shows slight F1-score decline in both DIALOGUE and DISCOURSE settings (Row 3) due to limitations of cosine similarity.

Table 6 shows sample outputs from the NEL systems. It shows that the addition of the context referring to *Nicki Minaj*, helps Flair + BLINK successfully links *Anaconda* to the correct song, while the absence of context links the term to a gun. In the second example, without any context, WAT links *Posey* to a county. Whereas in the DIALOGUE setting, by giving dialogue context, the system links it to a basketball player, *Posey James*.

Finally, in the DISCOURSE setting where references are substituted for anaphoras, it is linked to the correct baseball player *Buster Posey*.

In open-domain conversations, the topics human participants can talk about are unrestricted. Therefore it requires the NEL systems to be able to perform linking for entities in any topic domain. To evaluate the topic coverage of these systems, we compare the F1-score on all 12 topics in the DISCOURSE setting, shown in Figure 4. We see that the performance of each system varies on different topics. WAT and Flair+BLINK systems have relatively consistent performance across all topics. Both Spotlight and REL have weakness in the NFL football topic. The overall best model, Flair+BLINK, has its lowest F1-scores on the music, rock and action movies topics, which may be due to the nature of recognizing album names, song names or movie names since they tend to be highly ambiguous. See Tables 8, 9 and 10 in the Appendix for more detailed evaluation in other metrics and settings.

5. Conclusion

In this paper, we present and make publicly available the OPENEL corpus, the first large-scale corpus of open-domain dialogues annotated for NEL and anaphora with high annotator agreement. We tested and compared existing NEL systems including ML and DL-based methods on our corpus. We demonstrated the effectiveness of using dialogue context and anaphora resolution in open-domain NEL. We also showed the existing NEL systems with a performance gap between open-domain dialogues and human performance, which highlights the challenges of NEL in such settings. We plan to extend our corpus by annotating more conversations from other topics in EDINA and other conversational datasets. We envision our corpus as a good source of studying challenging problems, such as entity linking, anaphora resolution and knowledge-grounded dialogue generations, in the context of open-domain dialogue systems.

6. Bibliographical References

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649.

Botzer, N., Ding, Y., and Weninger, T. (2021). Reddit entity linking dataset. *Information Processing & Management*, 58(3):102479.

Bowden, K. K., Wu, J., Oraby, S., Misra, A., and Walker, M. (2018). Slugnerds: A named entity recognition tool for open domain dialogue systems. *arXiv preprint arXiv:1805.03784*.

Broscheit, S. (2020). Investigating entity knowledge in bert with simple neural end-to-end entity linking. *arXiv preprint arXiv:2003.05473*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Cano, A. E., Rizzo, G., Varga, A., Rowe, M., Stankovic, M., and Dadzie, A.-S. (2014). Making sense of microposts:(# microposts2014) named entity extraction & linking challenge. In *Ceur workshop proceedings*, volume 1141, pages 54–60.

Chiu, J. P. and Nichols, E. (2016). Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 708–716.

Curry, A. C., Papaioannou, I., Suglia, A., Agarwal, S., Shalyminov, I., Xu, X., Dušek, O., Eshghi, A., Konstas, I., Rieser, V., et al. (2018). Alana v2: Entertaining and informative open-domain social dialogue using ontologies and entity linking. *Alexa Prize Proceedings*.

De Cao, N., Izacard, G., Riedel, S., and Petroni, F. (2020). Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*.

Deleger, L., Li, Q., Lingren, T., Kaiser, M., Molnar, K., et al. (2012). Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*, volume 2012, page 144. American Medical Informatics Association.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. (2018). Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Durrett, G. and Klein, D. (2014). A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the association for computational linguistics*, 2:477–490.

Elsahar, H., Vougiouklis, P., Remaci, A., Gravier, C., Hare, J., Laforest, F., and Simperl, E. (2018). T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ferragina, P. and Scaiella, U. (2010). Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628.

Guo, Z. and Barbosa, D. (2018). Robust named entity disambiguation with random walks. *Semantic Web*, 9(4):459–479.

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*, pages 782–792.

Ji, H. and Grishman, R. (2011). Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 1148–1158.

Joko, H., Hasibi, F., Balog, K., and de Vries, A. P. (2021). Conversational entity linking: Problem definition and datasets. *arXiv preprint arXiv:2105.04903*.

Kolitsas, N., Ganea, O.-E., and Hofmann, T. (2018). End-to-end neural entity linking. *arXiv preprint arXiv:1808.07699*.

Krause, B., Damonte, M., Dobre, M., Duma, D., Fainberg, J., Fancellu, F., Kahembwe, E., Cheng, J., and Webber, B. (2017). Edina: Building an open domain socialbot with self-dialogues. *arXiv preprint arXiv:1709.09816*.

Lazic, N., Subramanya, A., Ringgaard, M., and Pereira, F. (2015). Plato: A selective context model for entity resolution. *Transactions of the Association for Computational Linguistics*, 3:503–515.

Le, P. and Titov, I. (2018). Improving entity linking by modeling latent relations between mentions. *arXiv preprint arXiv:1804.10637*.

Li, B. Z., Min, S., Iyer, S., Mehdad, Y., and Yih, W.-t. (2020). Efficient one-pass end-to-end entity linking for questions. *arXiv preprint arXiv:2010.02413*.

Logeswaran, L., Chang, M.-W., Lee, K., Toutanova, K., Devlin, J., and Lee, H. (2019). Zero-shot entity linking by reading entity descriptions. *arXiv preprint arXiv:1906.07348*.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8.

Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518.

Piccinno, F. and Ferragina, P. (2014). From tagme to wat: a new entity annotator. In *Proceedings of the first international workshop on Entity recognition & disambiguation*, pages 55–62.

Prabhakar Kannan Ravi, M., Singh, K., Onando Mulang, I., Shekarpour, S., Hoffart, J., and Lehmann, J. (2021). Cholan: A modular approach for neural entity linking on wikipedia and wikidata. *arXiv e-prints*, pages arXiv–2101.

Ratinov, L., Roth, D., Downey, D., and Anderson, M. (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.

Shang, M., Wang, T., Eric, M., Chen, J., Wang, J., Welch, M., Deng, T., Grewal, A., Wang, H., Liu, Y., et al. (2021). Entity resolution in open-domain conversations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 26–33.

Shen, W., Li, Y., Liu, Y., Han, J., Wang, J., and Yuan, X. (2021). Entity linking meets deep learning: Techniques and solutions. *IEEE Transactions on Knowledge and Data Engineering*.

Sun, Y., Zheng, Y., Hao, C., and Qiu, H. (2021). Nspbert: A prompt-based zero-shot learner through an original pre-training task–next sentence prediction. *arXiv preprint arXiv:2109.03564*.

Tang, H., Sun, X., Jin, B., and Zhang, F. (2021). A bidirectional multi-paragraph reading model for zero-shot entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13889–13897.

van Hulst, J. M., Hasibi, F., Dercksen, K., Balog, K., and de Vries, A. P. (2020). Rel: An entity linker standing on the shoulders of giants. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2197–2200.

Wu, L., Petroni, F., Josifoski, M., Riedel, S., and Zettlemoyer, L. (2019a). Scalable zero-shot entity linking with dense entity retrieval. *arXiv preprint arXiv:1911.03814*.

Wu, L., Petroni, F., Josifoski, M., Riedel, S., and Zettlemoyer, L. (2019b). Zero-shot entity linking with dense entity retrieval. corr abs/1911.03814 (2019). *arXiv preprint arxiv:1911.03814*.

Appendix

		Overall	baseball	music	star wars	pop	comedy	basketball	action	rap hiphop	movies	rock	nfl football	horror
strict match														
Spotlight	P	0.44	0.36	0.46	0.42	0.56	0.49	0.48	0.30	0.47	0.38	0.23	0.26	0.08
	R	0.45	0.34	0.50	0.41	0.60	0.48	0.40	0.32	0.60	0.44	0.23	0.31	0.11
	F	0.44	0.35	0.48	0.42	0.58	0.49	0.44	0.31	0.53	0.41	0.23	0.29	0.10
WAT	P	0.29	0.23	0.27	0.47	0.28	0.40	0.24	0.32	0.32	0.30	0.35	0.27	0.12
	R	0.36	0.33	0.32	0.43	0.33	0.36	0.40	0.39	0.57	0.32	0.32	0.44	0.44
	F	0.33	0.27	0.29	0.45	0.30	0.38	0.30	0.35	0.41	0.31	0.33	0.33	0.19
REL	P	0.50	0.43	0.54	0.48	0.73	0.60	0.40	0.51	0.44	0.50	0.44	0.37	0.43
	R	0.39	0.45	0.29	0.34	0.41	0.48	0.42	0.46	0.46	0.40	0.32	0.44	0.33
	F	0.44	0.44	0.38	0.40	0.53	0.53	0.41	0.49	0.45	0.44	0.37	0.40	0.38
Flair + BLINK	P	0.57	0.50	0.61	0.60	0.68	0.65	0.40	0.51	0.53	0.68	0.44	0.61	0.50
	R	0.49	0.55	0.41	0.51	0.44	0.55	0.43	0.51	0.60	0.60	0.36	0.69	0.44
	F	0.53	0.52	0.49	0.55	0.53	0.59	0.42	0.51	0.56	0.64	0.40	0.65	0.47
weak match														
Spotlight	P	0.47	0.40	0.48	0.43	0.61	0.49	0.48	0.33	0.51	0.38	0.41	0.26	0.08
	R	0.47	0.39	0.52	0.42	0.65	0.48	0.40	0.34	0.66	0.44	0.41	0.31	0.11
	F	0.47	0.39	0.50	0.43	0.63	0.49	0.44	0.33	0.57	0.41	0.41	0.29	0.10
WAT	P	0.36	0.28	0.32	0.55	0.38	0.42	0.40	0.42	0.35	0.33	0.40	0.27	0.15
	R	0.44	0.41	0.37	0.51	0.43	0.38	0.67	0.51	0.63	0.36	0.36	0.44	0.56
	F	0.40	0.34	0.34	0.53	0.41	0.40	0.50	0.46	0.45	0.35	0.38	0.33	0.24
REL	P	0.57	0.50	0.57	0.56	0.76	0.62	0.67	0.51	0.47	0.65	0.44	0.37	0.43
	R	0.44	0.52	0.30	0.40	0.43	0.50	0.70	0.46	0.49	0.52	0.32	0.44	0.33
	F	0.50	0.51	0.40	0.47	0.55	0.55	0.68	0.49	0.48	0.58	0.37	0.40	0.38
Flair + BLINK	P	0.66	0.58	0.66	0.71	0.76	0.73	0.66	0.54	0.60	0.73	0.56	0.61	0.75
	R	0.57	0.65	0.44	0.61	0.49	0.62	0.72	0.54	0.69	0.64	0.46	0.69	0.67
	F	0.61	0.61	0.53	0.66	0.60	0.67	0.69	0.54	0.64	0.68	0.50	0.65	0.71
	Sup.	1205	256	236	201	191	113	60	41	35	25	22	16	9

Table 8: NEL evaluation per topic of UTTERANCE setting. The last row (Sup.) is the support of ground-truth entities.

		Overall	baseball	music	star wars	pop	comedy	basketball	action	rap hiphop	movies	rock	nfl football	horror
strict match														
Spotlight	P	0.43	0.36	0.46	0.43	0.56	0.49	0.38	0.30	0.47	0.33	0.25	0.26	0.17
	R	0.44	0.35	0.49	0.42	0.60	0.48	0.32	0.32	0.60	0.40	0.23	0.31	0.22
	F	0.43	0.36	0.47	0.43	0.58	0.48	0.34	0.31	0.53	0.36	0.24	0.29	0.19
WAT	P	0.31	0.28	0.30	0.43	0.29	0.39	0.20	0.31	0.32	0.35	0.28	0.26	0.16
	R	0.50	0.51	0.50	0.53	0.45	0.57	0.38	0.41	0.71	0.52	0.41	0.44	0.67
	F	0.38	0.36	0.37	0.47	0.35	0.47	0.26	0.35	0.44	0.42	0.33	0.33	0.26
REL	P	0.57	0.52	0.62	0.52	0.77	0.62	0.44	0.63	0.54	0.50	0.44	0.47	0.62
	R	0.45	0.56	0.34	0.38	0.44	0.50	0.47	0.58	0.57	0.40	0.32	0.56	0.56
	F	0.50	0.54	0.44	0.44	0.56	0.55	0.46	0.61	0.56	0.44	0.37	0.51	0.59
Flair + BLINK	P	0.60	0.56	0.66	0.65	0.71	0.64	0.41	0.49	0.56	0.68	0.50	0.61	0.62
	R	0.51	0.61	0.40	0.54	0.44	0.55	0.45	0.49	0.63	0.60	0.41	0.69	0.56
	F	0.55	0.58	0.50	0.59	0.54	0.59	0.43	0.49	0.59	0.64	0.45	0.65	0.59
weak match														
Spotlight	P	0.46	0.41	0.48	0.44	0.62	0.49	0.38	0.33	0.51	0.33	0.45	0.26	0.17
	R	0.47	0.40	0.51	0.43	0.65	0.48	0.32	0.34	0.66	0.40	0.41	0.31	0.22
	F	0.46	0.40	0.49	0.44	0.64	0.48	0.34	0.33	0.57	0.36	0.43	0.29	0.19
WAT	P	0.38	0.35	0.37	0.51	0.38	0.42	0.35	0.40	0.36	0.41	0.38	0.33	0.18
	R	0.62	0.64	0.61	0.62	0.60	0.61	0.68	0.54	0.80	0.60	0.55	0.56	0.78
	F	0.47	0.45	0.46	0.56	0.47	0.49	0.47	0.46	0.50	0.48	0.44	0.42	0.30
REL	P	0.66	0.62	0.66	0.61	0.81	0.68	0.75	0.63	0.65	0.65	0.56	0.47	0.62
	R	0.52	0.67	0.36	0.45	0.46	0.54	0.78	0.58	0.69	0.52	0.41	0.56	0.56
	F	0.58	0.65	0.46	0.52	0.59	0.60	0.76	0.61	0.67	0.58	0.47	0.51	0.59
Flair + BLINK	P	0.70	0.65	0.70	0.77	0.80	0.72	0.69	0.54	0.64	0.68	0.61	0.61	0.88
	R	0.59	0.71	0.43	0.64	0.49	0.62	0.75	0.54	0.71	0.60	0.50	0.69	0.78
	F	0.64	0.68	0.54	0.70	0.61	0.67	0.72	0.54	0.68	0.64	0.55	0.65	0.82
	Sup.	1205	256	236	201	191	113	60	41	35	25	22	16	9

Table 9: NEL evaluation per topic of DIALOGUE setting. The last row (Sup.) is the support of ground-truth entities.

		Overall	baseball	music	star wars	pop	comedy	basketball	action	rap hiphop	movies	rock	nfl football	horror
strict match														
Spotlight	P	0.43	0.36	0.46	0.43	0.56	0.49	0.38	0.30	0.47	0.33	0.25	0.26	0.17
	R	0.44	0.35	0.50	0.42	0.59	0.48	0.32	0.32	0.60	0.40	0.23	0.31	0.22
	F	0.43	0.36	0.48	0.43	0.57	0.49	0.34	0.31	0.53	0.36	0.24	0.29	0.19
WAT	P	0.31	0.28	0.29	0.44	0.29	0.40	0.20	0.31	0.33	0.35	0.30	0.28	0.16
	R	0.51	0.52	0.49	0.53	0.46	0.57	0.38	0.41	0.71	0.52	0.46	0.50	0.67
	F	0.39	0.36	0.37	0.48	0.36	0.47	0.26	0.35	0.45	0.42	0.36	0.36	0.26
REL	P	0.57	0.53	0.62	0.54	0.77	0.62	0.44	0.60	0.54	0.50	0.44	0.47	0.38
	R	0.45	0.56	0.34	0.40	0.45	0.50	0.47	0.56	0.57	0.40	0.32	0.56	0.33
	F	0.51	0.54	0.44	0.46	0.56	0.55	0.46	0.58	0.56	0.44	0.37	0.51	0.35
Flair + BLINK	P	0.61	0.57	0.66	0.65	0.73	0.66	0.41	0.49	0.58	0.64	0.50	0.61	0.62
	R	0.52	0.63	0.40	0.54	0.45	0.56	0.45	0.49	0.63	0.56	0.41	0.69	0.56
	F	0.56	0.60	0.50	0.59	0.56	0.60	0.43	0.49	0.60	0.60	0.45	0.65	0.59
weak match														
Spotlight	P	0.47	0.41	0.48	0.44	0.62	0.49	0.38	0.33	0.51	0.33	0.45	0.26	0.17
	R	0.47	0.40	0.52	0.43	0.65	0.48	0.32	0.34	0.66	0.40	0.41	0.31	0.22
	F	0.47	0.40	0.50	0.44	0.64	0.49	0.34	0.33	0.57	0.36	0.43	0.29	0.19
WAT	P	0.38	0.34	0.36	0.51	0.39	0.43	0.36	0.40	0.36	0.41	0.39	0.34	0.18
	R	0.62	0.64	0.61	0.62	0.60	0.60	0.70	0.54	0.80	0.60	0.59	0.62	0.78
	F	0.47	0.45	0.45	0.56	0.47	0.50	0.47	0.46	0.50	0.48	0.47	0.44	0.30
REL	P	0.66	0.63	0.66	0.63	0.81	0.68	0.75	0.60	0.65	0.65	0.56	0.47	0.38
	R	0.52	0.67	0.36	0.46	0.47	0.54	0.78	0.56	0.69	0.52	0.41	0.56	0.33
	F	0.58	0.65	0.46	0.53	0.59	0.60	0.76	0.58	0.67	0.58	0.47	0.51	0.35
Flair + BLINK	P	0.71	0.68	0.71	0.76	0.81	0.74	0.69	0.54	0.66	0.68	0.61	0.61	0.88
	R	0.60	0.74	0.44	0.63	0.50	0.63	0.75	0.54	0.71	0.60	0.50	0.69	0.78
	F	0.65	0.71	0.54	0.69	0.62	0.68	0.72	0.54	0.69	0.64	0.55	0.65	0.82
	Sup.	1205	256	236	201	191	113	60	41	35	25	22	16	9

Table 10: NEL evaluation per topic of DISCOURSE setting. The last row (Sup.) is the support of ground-truth entities.