What more can Entity Linking do for Question Answering?

Naveen Raman
University of Maryland
nraman1@umd.edu

Pedro Rodriguez
University of Maryland
pedro@cs.umd.edu

Jordan Boyd-Graber
University of Maryland
jbg@umiacs.umd.edu

Abstract

We introduce a new NLP task–noun phrase linking (NPL)–which is a subset of entity linking and expands named entity linking (NEL) to link all noun phrases in a document to an external knowledge base. Our task is an expansion of NEL by linking not only named entities, but also references to named entities, and is distinct from coreference resolution in that references to unmentioned entities are also linked. Not only is this task more difficult, but performing well on this task would provide benefits to downstream systems, such as Question Answering systems (QA), which use entity linkers to assist with answering questions. By replacing these entity linkers with noun phrase linkers, the QA systems have more information, while shifting some of the difficulty of question answering to designing a good noun phrase linker. Our primary contribution is the introduction of the noun phrase linking task. To introduce NPL, we plan to collect an evaluation set based on annotating several QA datasets which we then use to compare NPL models, and estimate their effectiveness in improving end-to-end QA accuracy. This new entity linking task is more difficult than traditional entity linking, because of the difficulty connecting implicit references to named entities, and so requires a method to efficiently collect data. Our second contribution is that we develop an efficient method to collect annotation data by motivating domain experts to annotate and using human-in-the-loop annotation to assist annotators. Data collection is efficiently done by guiding human annotators towards examples where multiple entity linking models disagreed while maintaining accuracy on a gold set. We propose experiments to evaluate the effect of noun phrase linking on question answering systems, and also compare our new noun phrase linking systems against baseline coreference and entity linking systems. In summary, we introduce NPL, demonstrate a method to efficiently collect data, and propose experiments.

1 Introduction

We introduce the new task of noun phrase linking, which annotates all noun phrases in a document with its corresponding entry in a knowledge base (such as Wikipedia). This task can be viewed as a generalization of Named Entity Linking (NEL), which matches named entities with entries in an external knowledge base. The task is more difficult than NEL because of the difficulty in resolving non-explicit references. The task is also related to, but separate from, coreference resolution, as it also deals with references to entities not present in the document, but only deals with noun phrases that link to an external knowledge base.

Named entity linking is used in many downstream tasks, including question answering (QA) systems [23]. For example, in Figure 1, a question from Quizbowl (QB), a trivia competition, is annotated with entities, such as "one work by this author" linking to Novum Organum. Expanding from NEL to noun phrase linking gives QA systems more information and offloads some of the difficulty of question answering from QA systems to the noun phrase linkers. We focus on QB because Quizbowl questions have sophisticated noun phrases, because the queestions describe but don't explicitly mention named entities. Quizbowl questions average 21.2 entities per question, which is more than other datasets, such as TriviaQA which only has 2.2 [24]. Replacing noun phrases

Preprint. Under review.

> One work by this author (Novum Organum) uses printing, gunpowder, and the compass as symbols of personal ambition, national ambition, and the ambition of the human race to extend its grasp. This thinker (Francis Bacon) described three forms of false learning as “delicate”, “contentious”, and “fantastical” in categorizing the “distempers” that impede academic progress. This thinker (Francis Bacon) imagined a utopian university called Salomon’s House (Salomon’s house), and he (Francis Bacon) likened received systems of philosophy to stage plays that misrepresent the world, and thus labeled them “idols of the theatre”(Idola Theatari). This author of The New Atlantis (New Atlantis) established the doctrine of inductive, empirical methodology (Baconian method). For 10 points, name this 17th-century English philosopher (Francis Bacon) who wrote Novum Organum (Novum Organum) and spearheaded the Scientific Revolution (Scientific Revolution)
> **Answer:** Francis Bacon

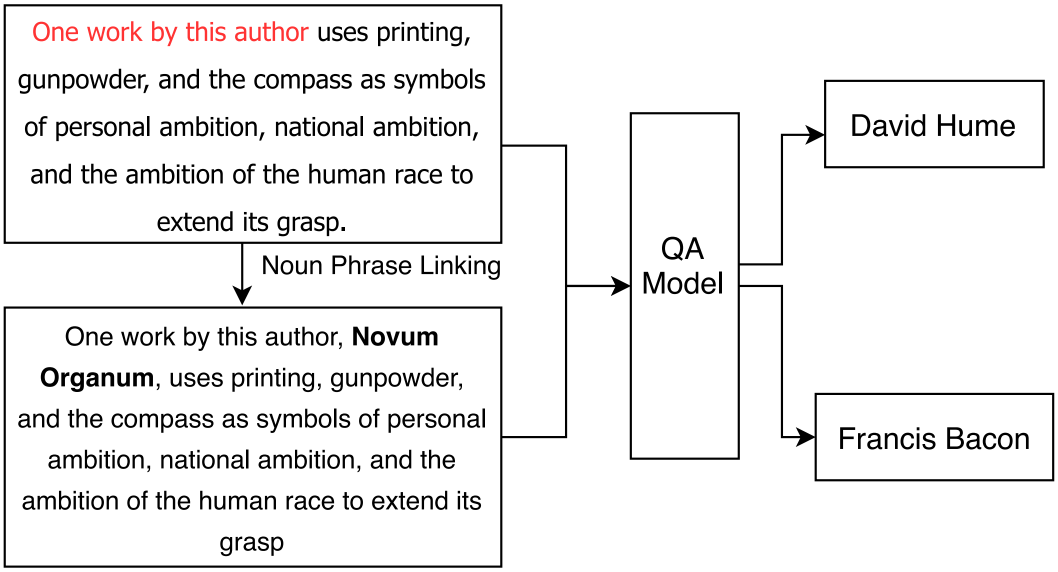

One work by this author uses printing, gunpowder, and the compass as symbols of personal ambition, national ambition, and the ambition of the human race to extend its grasp.
Noun Phrase Linking
One work by this author, **Novum Organum**, uses printing, gunpowder, and the compass as symbols of personal ambition, national ambition, and the ambition of the human race to extend its grasp
QA Model
David Hume
Francis Bacon

Figure 1: We show a QB question with a variety of hard and easy entities; entities that would be linked only by a noun phrase linker are in blue. Questions from this dataset are entity dense, and contain complex reference patterns, such as "doctrine of inductive, empirical methodology" linking to the Baconian method. In general, we validate the utility of noun phrase linking for question answering by substituting noun phrases with their links and observe accuracy gains. For example, replacing noun phrases in the first sentence changes the answer from the incorrect David Hume, to the correct Francis Bacon.

with their respective entities allows for question answering systems to gain information, which changes the answer from David Hume, to the correct answer of Francis Bacon.

Noun phrase linking is a more difficult task than named entity linking due to the difficulty of annotating non-explicit references, and thus presents a challenge when curating a dataset. This work describes a process to build a dataset of noun phrase annotations for challenging trivia questions like QB [19] in Figure 1, Jeopardy! [9], TriviaQA [13], and Quasar-T [5]. Although building datasets is expensive and time-consuming, this is mitigated and quality is improved by incorporating prior entity linking models into the annotation process [21, 6, 3, 17]. Our key insight and the hypothesis we test is that the bias introduced by exposing human annotators to machine predictions is small in comparison to the increase in annotation coverage and quality. To measure this, we plan to run an experiment where we vary the conditions in which the training data is annotated. Specifically, we plan to vary which entity linking or coreference models are used to assist annotator, and analyze the difference in linking accuracy by the type of models used.

We additionally plan to use human-in-the-loop annotation to suggest noun phrases to annotate along with the links for those phrases. We motivate experts to annotate, in this case, trivia competitors and organizers, which improves annotation quality.

After collecting data, we design experiments to evaluate state of the art NEL and coreference models on noun phrase data. We collect a gold set and evaluate performance on that data set to determine the difficulty of annotating noun phrases when compared to general entity linking. We develop a baseline noun phrase linking model based off of the dataset, which is trained through a human-in-the-loop process, and compare its performance against entity linkers and coreference models. We additionally develop experiments to determine the extent to which noun phrase linking assists with question answering.

In summary, we make three contributions: (1) Define the new problem of noun phrase annotating, and show that annotating noun phrases improves QA performance (2) Develop a method to collect noun phrase linking dataset using text from trivia datasets, (3) Propose experiments that evaluate current named entity linkers and coreference models on the noun phrase dataset, compare a baseline noun phrase model to entity linking and coreference, and evaluate the impact that noun phrase linking has upon question answering accuracy.

2 Noun Phrase Linking

We define the problem of Noun Phrase Linking and motivate the problem by discussing the improved performance in downstream tasks due to noun phrase linking. We additionally define guidelines for noun phrase linking and develop an interface for annotating documents using those guidelines in Section 3.

2.1 Noun Phrase Linking

Named entities refer to specific nouns such as people's names, and the name of places. Annotating named entities provides a gain in accuracy when augmented to QA systems, but named entities exclude certain noun phrases that could further assist QA systems. In particular, resolving anaphoric references, such as resolving "One work by this author" to "Novum Organum", provides a more difficult task because of the lack of a direct link between the noun phrase and the entity to be linked. Resolving anaphoric references could allow for a bigger gain in helping QA systems, as seen by changing the answer to the correct answer, Francis Bacon, when noun phrases are replaced by their referenced entity (Figure 1).

We define the task of noun phrase linking to be the union of annotating anaphoric references along with annotating named entities, and in effect, annotating all noun phrases in a document that link to a named entity. This task differs from the coreference and entity linking, as entities that are referred to, but not necessarily ever mentioned in the document, can be linked. For example, within the first sentence of Figure 1, Novum Organum is never mentioned. However, "One work by this author" refers to Novum Organum, and so would be annotated in Noun Phrase Linking, but not in either coreference or entity linking. We develop guidelines for linking noun phrases (Section 3), and plan to conduct experiments to determine the effect of annotating noun phrases upon QA performance (Section 4).

We extend traditional entity linking to noun phrase linking because traditional entity linkers have been shown to perform well on NEL tasks, and extending to noun phrase linking allows for a more challenging task. Current models do well in NEL and coreference resolution, both of which are tasks related to noun phrase linking. Thus, it's plausible, although imperfect, that future models may show improvement on downstream tasks. We also propose experiments that explore the effect that noun phrase linking has upon question answering accuracy (Section 4), to determine the extent to which noun phrase linking assists question answering systems.

2.2 Noun Phrase Linking Guidelines

We develop guidelines for determining which entities to link, and what to link them to. These guidelines will be used by annotators, along with examples, when determining what to annotate. We link text spans that are noun phrases and refer to a uniquely identifiable named entity in the knowledge base. For example, in Figure 1, we link "idols of the theatre" because it refers to a named entity. On the other hand, we don't link the word "symbols", despite the presence of a Wikipedia page for symbol, because smybol does not refer to a specific symbol, but rather the general word. If no Wikipedia page is present for a noun phrase, then we link it with "No Entity." The "No Entity" links are subdivided into "No Entity Character" and "No Entity Literature" for characters and works of literature respectively. Wikipedia is an incomplete knowledge base; rather than omit links simply because the correct entity does not exist, the entity should be linked, but assigned to a special null entity indicating its type.

3 Data Collection

The task of noun phrase linking is more difficult than NEL, due to the difficulty of finding indirect links. Because of this, we develop methods to efficiently collect data. The Quizzical Entity Linking (QUEL) dataset is annotated with an interface (Figure 2) that supports basic entity linking functionality (Section 3.1), configurable inclusion of machine-generated links to vary annotation conditions, and features to motivate expert annotators to participate in data collection (Section 3.5). Gold data will be collected by the authors of this paper (Section 3.4), and all other data will be collected by organizers of QB tournaments (Section 3.6). We plan to utilize human-in-the-loop annotation to help annotators and speed up the annotation process (Section 3.3) and assist users by pre-linking certain entities (Section 3.2).

3.1 Entity Linking Interface

To collect the QUEL dataset we built the interface in Figure 2. To annotate entity links, users: (1) select a text span, (2) search for the correct entity, and (3) confirm their choice. Annotators select entities from among all valid Wikipedia pages.[1] This process is iterative until the user is satisfied with the links in the question. Currently, we suggest entities for the user based on full-text search, matching their noun phrase to Wikipedia articles.

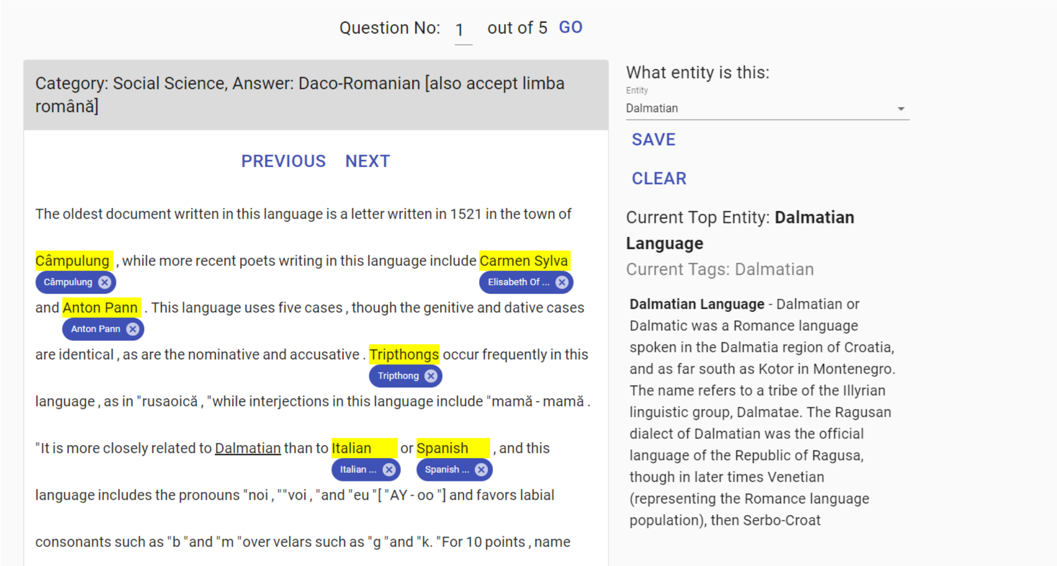
We plan to add annotations for subspans, otherwise known as nested entities. Examples of this include the entity "Washington crossing the Delaware"; the whole entity would be matched with the painting, while Washington would be linked to George Washington and Delaware would be linked to Delaware. We propose two methods of doing this

1. List of noun phrase suggestions, which includes nested noun phrases. The user simply has to annotate each noun phrase. This would reduce the time needed for annotation, as the user does not need to search for noun phrases.
2. Augment the current interface with nested spans, allowing users to tag a particular word or phrase as part of multiple entities. While simpler to implement, this option might be less user friendly.

3.2 Using other Entity Linkers

We utilize different experimental conditions to speed up the entity linking process by pre-populating entity links. Later (Section 4), we propose experiments to compare different experimental conditions to determine which condition would optimize entity linking accuracy and speed. Before a question is loaded, we assign the annotator an experimental condition that decides how entity links are pre-populated. In the first condition, no entity links are pre-populated so the question is annotated from scratch. The second condition pre-populates entity links with the output of one randomly selected entity linker. In the third condition, we use a named entity recognition system to display candidate mentions, but do not pre-link them to Wikipedia entities. The final condition pre-populates links that are predicted by two or more of the linkers. Later, we analyze the annotation differences on a shared set of questions as well as distributionally across the non-shared questions.

[1] We use the Wikipedia dump from 06/2020.



Category: Social Science, Answer: Daco-Romanian [also accept limba română]

The oldest document written in this language is a letter written in 1521 in the town of Câmpulung , while more recent poets writing in this language include Carmen Sylva and Anton Pann . This language uses five cases , though the genitive and dative cases are identical , as are the nominative and accusative . Tripthongs occur frequently in this language , as in "rusaoică , "while interjections in this language include "mamă - mamă . "It is more closely related to Dalmatian than to Italian or Spanish , and this language includes the pronouns "noi , ""voi , "and "eu "["AY - oo "] and favors labial consonants such as "b "and "m "over velars such as "g "and "k. "For 10 points , name

What entity is this:

Entity: Dalmatian

Current Top Entity: **Dalmatian Language**

Current Tags: Dalmatian

Dalmatian Language - Dalmatian or Dalmatic was a Romance language spoken in the Dalmatia region of Croatia, and as far south as Kotor in Montenegro. The name refers to a tribe of the Illyrian linguistic group, Dalmatae. The Ragusan dialect of Dalmatian was the official language of the Republic of Ragusa, though in later times Venetian (representing the Romance language population), then Serbo-Croat

Figure 2: We show our annotation interface currently, which has the ability to select text spans and tag them with an entity. We plan to suggest noun phrases to annotate in the future, and also plan to allow users to annotate nested spans.

3.3 Human-in-the-Loop Annotation

Prior research has shown that human-in-the-loop annotation for entity linking tasks can speed up the process [14]. To do this, we plan to recommend which entities a particular noun phrase might be linked to via a model, so that our annotation system assists users with determining what noun phrased are linked to. At the moment, we plan to build two simple baseline models. The first model recognizes noun phrases based off of n-grams, though we could also use some type of BERT based embedding [4]. The second model uses these noun phrases and links them to a Wikipedia page. The model will deliver a confidence score for a list of entities associated with each noun phrase, and we let the user select among these options to assist them with entity linking. With this, annotators can focus on annotating lower confidence entities. As the model improves, our suggestions will also improve, and annotators will be able to annotate faster. We additionally use information from user annotations to determine which noun phrases can be pre-annotated, saving annotators time in finding noun phrases.

3.4 Gold Annotations

Prior to scaling our data collection, we annotated a gold set of ten development set questions in QB; in the future, we plan to annotate one hundred development set questions in QB, TriviaQA, and SearchQA. For gold annotation, we—the authors—plan to doubly annotate each question from scratch. We plan to iteratively annotate twenty-five questions at a time before checking for annotation disagreements. On disagreement, we plan to either identify the mistake, identify unclear guidelines, or identify genuinely ambiguous cases. To create the final gold set, we plan to exclude ambiguous cases and reach a consensus on disagreements. To determine inter-rater reliability, we plan to use kappa scores [16].

3.5 Motivating Experts to Annotate

Instead of crowd workers, we work with the QB community, where incentives are aligned. Within this community, it is valuable for players to know the distribution of topics and entities to help them study. It is similarly helpful for question writers to know the distribution of question topics so that they can design tournaments with a diverse collection of entities. To support the QB community, we plan to build two features. First, we plan to add a *tournament view* that aggregates information from all questions in a tournament, such as the distribution of topics, which entities were mentioned, and the types of entities mentioned. This allows tournament organizers to know which entities are under and over-represented when writing questions. Second, we plan to build an interface where users can search for questions based on the entities mentioned, types of the entities, and topic area. This allows users to study based on particular entities, and find out the context that a particular entity appears in. Using expert trivia competitors instead of crowd workers is better, due to the skill level of these competitors. Their annotations would accurately identify entities, and annotate them, allowing for a higher quality dataset, that is also annotated faster.

3.6 Quality Control

In addition to aligning incentives, we also plan to control annotation quality through multiple annotations and test examples. We only plan to use questions that were annotated at least twice, and we measure inter-annotator agreement using kappa scores [16]. Additionally, we plan to annotate two questions per packet, which we use as canaries to detect under-performing annotators. If the same user annotates too many canaries incorrectly, we disregard all their annotations.

4 Proposed Experiments

We design three experiments to evaluate different aspects of noun phrase linking. Our first experiment compares prior entity linkers and coreference annotators on the noun phrase task, to determine if prior solutions can be used to solve the problem. We design an experiment to determine if using prior entity linkers to assist annotators with improves the precision or recall of annotations. After data collection, we design an experiment to assess the degree to which noun phrase linking assists question answering systems. We also plan to develop a simple model for noun phrase linking, and compare it with a baseline coreference+named entity linking model on the collected dataset.

4.1 Comparison of prior Entity Linkers

To evaluate current entity linkers on QA tasks, we first plan to characterize the generalizability of NEL models trained on AIDA and TAC 2010 to text from QA tasks. We measure this directly by comparing the predictions of TAGME, BLINK [22], and [11] to a gold set of noun phrase annotations on one hundred questions from the development and test sets of QB, TriviaQA, and SearchQA. We compare the precision and recall for each of these on the gold set, both only considering named entities, and also considering all noun phrases. We additionally plan to augment entity linkers with co-reference models to see if this provides a gain in precision or recall.

4.2 Noun Phrase Annotation effect on QA models

To determine the effect of noun phrase annotation on QA models, we run an experiment on QB questions. We plan to do this by replacing entities with their linked Wikipedia page title and evaluating performance through QANTA, as in Figure 1. We consider our initial gold dataset to consist of 20 randomly chosen QB questions, and we replace noun phrases with their corresponding entity. We compare the three aforementioned linkers in addition to the gold linking set, which annotates noun phrases. To evaluate QA accuracy, we use QANTA to predict each sentence in each question and compute the accuracy percentage. This is to demonstrate whether resolving noun phrases has any effect upon downstream performance. To motivate the experiment, we replace noun phrases for one sentence in a question (Figure 1) and find that it changes the answer from the incorrect David Hume, to the correct Francis Bacon. Our metric is the accuracy of the QANTA QA system with the replaced entities.

4.3 Entity Linker effect on annotations

We plan to analyze annotation quality by comparing annotations from a set of one hundred questions. These questions will be annotated five times: once by us for gold annotations and once for each of the four experimental conditions (Section 3.1). Annotations will be compared with standard entity linking metrics, such as precision and recall, with annotations treated as model predictions. Additionally, we plan to compare inter-rater reliability through kappa scores. Our goal is to determine how best to improve the task of entity linking to make it easier for annotators, without sacrificing accuracy.

4.4 Model comparison

We develop a noun phrase model that is improved during human-in-the-loop data collection. The model is based off of n-grams, and is split into two parts; the first finds noun phrases, and the second links these noun phrases to Wikipedia. We train and compare this model to a baseline coreference and named entity linking model, to determine whether models can be developed to perform well on the noun phrase linking task. We perform cross validation on our collected data, as to not mix the train and test sets, to compare the two models. We measure both the precision and recall for retrieving noun phrases.

5 Related Work

Our work fits within the larger context of entity linking and question answering systems. In particular, we define a new version of entity linking that expand upon named entity linking. Despite the popularity of entity linking, there is little consensus amongst practitioners on how to precisely define Entity Linking [15]. This results in there being a variety of different ways to link entities, which depend on what the entities are used for [20]. Many versions of entity linking build upon named entity linking and develop a more difficult task, such as multilingual entity linking [18]. Other versions of entity linking include Wikification, which has Wikipedia as the external knowledge base [2]. Similarly, co-reference has been jointly accomplished with named entity recognition and entity linking [10]. Noun phrase linking is also related to implicit entity recognition, which has previously been studied in the context of references between tweets [12].

Entity linking has been used for a variety of applications, including question answering, such as the EARL system [7, 8], which relies upon performing entity and relation linking at the same time. Additionally, entity linking is used along with knowledge graphs in order to answer questions [24]. Entity linking is also used for text understanding when used along with BERT [1]. Because entity linking is used for so many applications, developing a noun phrase linking dataset could potentially be useful for many downstream tasks.

6 Conclusion

We introduce the new problem of noun phrase annotation, which is a generalization of NEL. We find that introducing noun phrase annotation may be useful in downstream tasks such as question answering. However creating a dataset for noun phrase annotation is a difficult task. We counter this problem by developing a human-in-the-loop method to efficiently annotate questions and motivate experts to annotate questions by assisting them with studying and directing tournaments. To explore the difficulty of noun phrase annotation, we propose experiments that compare the performance of NEL and coreference linkers on our noun phrase annotation dataset. We additionally design experiments to compare entity linking with multiple configurations in order to determine how best to assist users when entity linking. We finally design experiments to determine the effect that noun phrase annotation has upon question answering; we do this by comparing the accuracy of QA models when replacing entities with their entry in the knowledge base. Our next steps are to proceed forward with data collection, and to run the experiments on the noun phrase dataset.

References

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

[2] X. Cheng and D. Roth. Relational inference for wikification. In *Proceedings of Empirical Methods in Natural Language Processing*, 2013.

[3] P. Dasigi, N. F. Liu, A. Marasović, N. A. Smith, and M. Gardner. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proceedings of Empirical Methods in Natural Language Processing*, 2019.

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

[5] B. Dhingra, K. Mazaitis, and W. W. Cohen. Quasar: Datasets for question answering by search and reading. July 2017.

[6] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the Association for Computational Linguistics*, 2019.

[7] M. Dubey, D. Banerjee, D. Chaudhuri, and J. Lehmann. EARL: joint entity and relation linking for question answering over knowledge graphs. In *International Semantic Web Conference*, pages 108–126. Springer, 2018.

[8] M. Dubey, S. Dasgupta, A. Sharma, K. Höffner, and J. Lehmann. Asknow: A framework for natural language query formalization in sparql. In *European Semantic Web Conference*, pages 300–316. Springer, 2016.

[9] M. Dunn, L. Sagun, M. Higgins, V. U. Güney, V. Cirik, and K. Cho. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.

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

[11] N. Gupta, S. Singh, and D. Roth. Entity linking via joint encoding of types, descriptions, and context. In *Proceedings of Empirical Methods in Natural Language Processing*, 2017.

[12] H. Hosseini. Implicit entity recognition, classification and linking in tweets. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1448–1448, 2019.

[13] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the Association for Computational Linguistics*, 2017.

[14] J.-C. Klie, R. E. de Castilho, and I. Gurevych. From zero to hero: Human-In-The-Loop entity linking in low resource domains. In *Proceedings of the Association for Computational Linguistics*, 2020.

[15] X. Ling, S. Singh, and D. S. Weld. Design challenges for entity linking. *Transactions of the Association for Computational Linguistics*, 3:315–328, 2015.

[16] M. L. McHugh. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282, 2012.

[17] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the Association for Computational Linguistics*, 2020.

[18] J. Raiman and O. Raiman. Deeptype: Multilingual entity linking by neural type system evolution. In *AAAI*, 2018.

[19] P. Rodriguez, S. Feng, M. Iyyer, H. He, and J. Boyd-Graber. Quizbowl: The case for incremental question answering. *arXiv preprint arXiv:1904.04792*, 2019.

[20] H. Rosales-Méndez, B. Poblete Labra, and A. Hogan. What should entity linking link? 2018.

[21] E. Wallace, P. Rodriguez, S. Feng, and J. Boyd-Graber. Trick me if you can: Human-in-the-loop generation of adversarial question answering examples. In *Transactions of the Association for Computational Linguistics*, 2019.

[22] L. Wu, F. Petroni, M. Josifoski, S. Riedel, and L. Zettlemoyer. Zero-shot entity linking with dense entity retrieval. 2019.

[23] I. Yamada, R. Tamaki, H. Shindo, and Y. Takefuji. Studio ousia's quiz bowl question answering system. *arXiv preprint arXiv:1803.08652*, 2018.

[24] C. Zhao, C. Xiong, X. Qian, and J. Boyd-Graber. Complex factoid question answering with a free-text knowledge graph. In *Proceedings of the World Wide Web Conference*, 2020.