Unsupervised Entity Linking with Guided Summarization and Multiple-Choice Selection

Young-Min Cho **Li Zhang** **Chris Callison-Burch**
University of Pennsylvania
{jch0, zharry, ccb}@seas.upenn.edu

Abstract

Entity linking, the task of linking potentially ambiguous mentions in texts to corresponding knowledge-base entities, is an important component for language understanding. We address two challenge in entity linking: how to leverage wider contexts surrounding a mention, and how to deal with limited training data. We propose a fully unsupervised model called SumMC that first generates a guided summary of the contexts conditioning on the mention, and then casts the task to a multiple-choice problem where the model chooses an entity from a list of candidates. In addition to evaluating our model on existing datasets that focus on named entities, we create a new dataset that links noun phrases from WikiHow to Wikidata. We show that our SumMC model achieves state-of-the-art unsupervised performance on our new dataset and on existing datasets.

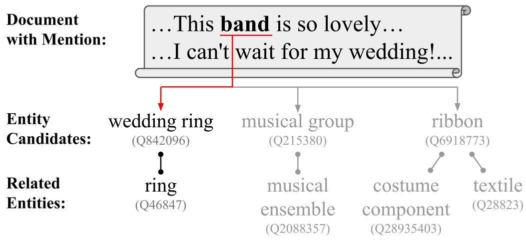

Figure 1: Example of an Entity Linking problem.

1 Introduction

Entity linking (EL) is an important Natural Language Processing (NLP) task that associates ambiguous mentions to corresponding entities in a knowledge base (KB, also called knowledge graph). EL is a crucial component of many NLP applications, such as question answering (Yih et al., 2015) and information extraction (Hoffart et al., 2011).

Although there have been significant and continuous developments of EL, most work requires sufficient labeled data and a well-developed KB (Zhang et al., 2021; Mulang' et al., 2020; van Hulst et al., 2020; Raiman and Raiman, 2018). However, many real-world applications, especially those in specific domains, suffer from scarcity of both training data and a fully-populated KB. Previous research has tackled this problem by learning EL models without data labeled entity links, but requires indirect supervision in the form of textual descriptions attached to entities in KBs, drawn from sources such as Wikipedia (Cao et al., 2017; Logeswaran et al., 2019). However, such descriptions may not be available in KBs in low-resource domains such as medicine or law. Thus, we focus on *fully unsupervised EL*, which only has access to the entities' names and their KB relations like `subclass-of` (Le and Titov, 2019; Arora et al., 2021).

One challenge of unsupervised EL is leveraging useful information from potentially noisy and misleading context (Pan et al., 2015). Specifically, a local context (the sentence containing the mention) may not be sufficient for disambiguating the target mention without the global context (other sentences in the document). For example, in Figure 1, the target mention '*band*' cannot be disambiguated solely with the local context "*This band is so lovely*", but needs to consider the global context that also includes "*I can't wait for my wedding*."

To address this problem, we introduce an unsupervised approach to EL that builds on the strengths of large neural language models like GPT-3 (Brown et al., 2020). We use zero-shot GPT-3 prompting for two sub-tasks. First, we perform **guided summarization**, which summarizes the input document conditioned on the target mention and outputs a condensed global context. Then, we cast EL to a **multiple-choice selection** problem where the model chooses an entity from a list of candidates. We refer to our unsupervised EL model as SumMC (**Sum**marization+**M**ultiple-**C**hoice).

With a few exceptions (Ratinov et al., 2011; Cheng and Roth, 2013), the majority of EL work targets named entities, such as names of people

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9394 - 9401
December 7-11, 2022

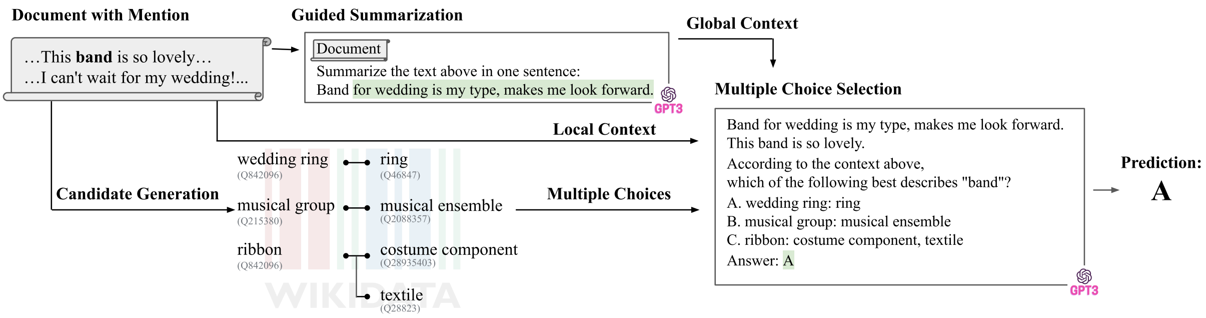

Figure 2: Pipeline of SumMC. Texts highlighted with green are machine generated.

and organizations (Mulang' et al., 2020; van Hulst et al., 2020), neglecting entities such as physical objects or concepts. To comprehensively evaluate our model, we create the first EL dataset on procedural texts, WikiHow-Wikidata, which links noun phrases from WikiHow[1] to Wikidata[2] entities (Vrandečić and Krötzsch, 2014).

Our SumMC model outperforms current state-of-the-art (SoTA) unsupervised EL models on our new WikiHow-Wikidata data, as well as existing benchmarks including AIDA-CoNLL (Hoffart et al., 2011), WNED-Wiki and WNED-Clueweb dataset (Guo and Barbosa, 2018). In addition, we also provide ablation studies to show the positive influence of generating guided summaries.[3]

2 Methodology

Fully unsupervised EL is the task that links a target mention from a given document to some entities in a KB without requiring any text data to be labeled with explicit links to the KB. The only available information in the KB is the names of the entities and the relations among them. In this paper, we follow previous work (Le and Titov, 2019; Arora et al., 2021) and use Wikidata as our target KB, which defines `instance-of` and `subclass-of` relations between entities. Wikidata can be seen as a knowledge graph with entities as nodes and relations as edges, and the popularity of an entity can be represented by its degree.

We now introduce SumMC, our proposed unsupervised EL model which consists of two instances of a generative language model. The first performs guided summarization by generating a summary of the document conditioned on a mention. The second casts EL to a multiple-choice selection problem and chooses an appropriate entity from a list of candidates generated by some heuristics. In our work, we use GPT-3 as the language model due to its superior performance on various NLP tasks (Brown et al., 2020).

Candidate Generation. Following previous work (Le and Titov, 2019; Arora et al., 2021), we first select all entities from Wikidata whose name or alias contains all tokens in a mention. Then, we narrow it down to the top 20 entities with the highest degree (in-degree + out-degree) in the KB. For each entity in the final list, we produce a textual representation by concatenating the names of all related entities. For example, the representation of the candidate *ribbon* in Figure 1 is *ribbon: costume component, textile*.

SumMC. The first application of GPT-3 performs a *guided summarization* of the input document. With zero-shot prompting, GPT-3 summarizes the texts using the prompt *"[D] Summarize the text above in one sentence: [M]"*, where *[D]* is the input document and *[M]* is the target mention. Here, we force GPT-3's summarization to start with the mention to ensure that the conditioned summary contains both the target mention and related global context. At this point, the generated summary serves as a global context while the sentence containing the mention serves as a local context, both of which help disambiguate the target mention.

The second application of GPT-3 casts the task to *multiple-choice selection* following many successful cases (Ouyang et al., 2022). With the two contexts, GPT-3 transforms EL to a multiple-choice question using the prompt *"According to the context above, which of the following best describes [M]?"*, followed by the representations of the mention *[M]*'s candidates as choices.

[1]https://www.wikihow.com/Main-Page
[2]https://www.wikidata.org/wiki/Wikidata:Main_Page
[3]The code and data are available at https://github.com/JeffreyCh0/SumMC

3 WikiHow-Wikidata Dataset

Most work on EL has targeted named entities, especially in the news. To account for more diverse entities in different styles of texts, we create a human-annotated dataset called **WikiHow-Wikidata** that links noun phrases in procedural texts to Wikidata. The research revolving around entities in procedural texts have long received much attention in the community (Dalvi et al., 2018; Zhang et al., 2020; Tandon et al., 2020; Zhang, 2022), without existing large-scale datasets of entity links in such a style of texts.

To create the dataset, we first extract 40,000 articles from the WikiHow corpus (Zhang et al., 2020) detailing everyday procedures. To select mentions to link, we choose the top 3 most-frequently-occurring nouns from each article using a part-of-speech tagger, assuming that most mentions in a document share the same word sense (Gale et al., 1992). Then, we ask students from a university in the U.S. to manually link these mentions to some Wikidata entity. Finally, to measure and control annotation quality, we manually annotate a subset of examples beforehand as control questions. Details about our data collection process, interface, and measures for quality control can be found in Appendix B. Eventually, WikiHow-Wikidata consists of 11,287 triples of a WikiHow article, a target mention, and a Wikidata entity.

4 Experiments

We evaluate our SumMC model along with other strong baselines on some widely used EL datasets and our WikiHow-Wikidata dataset.

4.1 Models

τ**MIL-ND**: Le and Titov (2019) introduced the first EL model that did not require an annotated dataset. Their model casts the EL task to a binary multi-instance learning (Dietterich et al., 1997) problem along with a noise-detecting classifier.

Eigentheme: Arora et al. (2021) created Eigentheme, the current state-of-the-art among fully unsupervised EL models. By representing each entity with its graph embedding, the model identifies a low-rank subspace using SVD on the embedding matrix and ranks candidates by the distance to this hyperplane.

To analyze the effect of using global context in our SumMC model, we report the evaluation results using three variations.

Dataset	Mentions #Easy	#Hard	#Not-found	#Documents
WikiWiki	2,727 (24%)	8,560 (76%)	0	7,097
AIDA-B	2,555 (57%)	1,136 (25%)	787 (18%)	230
WNED-Wiki	2,731 (41%)	1,475 (22%)	2,488 (37%)	318
WNED-Cweb	4,667 (42%)	3,056 (28%)	3,317 (30%)	320

Table 1: Statistics of datasets showing distributions of mention difficulty.

SumMC: Our proposed model integrates GPT-3 guided summarization and multiple-choice selection models. We use the Curie model for summarization conditioned on the target mention and the Davinci model for multiple-choice. As discussed before, both global and local contexts are provided.

–Guide: This is an ablated version of SumMC that generates summaries without being conditioned on the target mention. While both global and local contexts are provided, the global context is not guaranteed to be related to the target mention.

–Sum: This is another ablated version that does not generate summaries of a whole document but directly performs multiple-choice selection, given only with the local context of the mention.

4.2 Dataset

We choose AIDA-CoNLL-testb (AIDA-B), WNED-Wiki, and WNED-Clueweb (WNED-Cweb) to measure models' performance on disambiguating named entities and use our WikiHow-Wikidata (WikiWiki) dataset for evaluating on noun phrases.

Following previous settings (Tsai and Roth, 2016; Guo and Barbosa, 2018; Arora et al., 2021), we report micro precision@1 (P@1) and categorize each mention into 'easy' and 'hard' by whether the candidate entity with the highest degree in the knowledge graph is the correct answer. Performance on 'hard' mention is important since it shows the model's ability on highly ambiguous mentions. 'Not-found' is for mentions whose candidate list does not contain the correct answer. 'Overall' performance is reported considering all mentions, including 'Not-found' by treating it as a false prediction. The distribution of each dataset is shown in Table 1.

5 Results and Discussion

We show our results in Table 2. Our SumMC model achieves significantly better results than other unsupervised EL models in all evaluation datasets. Specifically, SumMC has a strong performance on

	WikiHow-Wikidata			AIDA-B			WNED-Wiki			WNED-Clueweb		
	Overall	Easy	Hard	Overall	Easy	Hard	Overall	Easy	Hard	Overall	Easy	Hard
τMIL-ND	-	-	-	0.45	0.70	0.19	0.13	-	-	0.27	-	-
Eigentheme	0.50	0.61	0.53	0.62	**0.86**	0.50	0.44	**0.82**	0.47	0.41	**0.77**	0.29
SumMC (ours)	**0.76**	**0.62**	**0.80**	**0.64**	0.80	**0.71**	**0.47**	0.81	**0.65**	**0.48**	0.75	**0.60**
Improvement over SoTA	+0.26	+0.01	+0.27	+0.02	-0.06	+0.21	+0.03	-0.01	+0.18	+0.07	-0.02	+0.31

Table 2: Performance comparison across SoTA models. Result is reported with Precision@1. We get results of τMIL-ND and Eigentheme on public datasets from Arora et al. (2021). 'Overall' shows result considering 'Not-found' mentions.

		–Guide	–Sum
WikiWiki	Easy	-0.02	-0.01
AIDA-B	Easy	-0.02	-0.03
WNED-Wiki	Easy	-0.01	-0.07
WNED-Cweb	Easy	-0.02	-0.03
Average	**Easy**	**-0.02**	**-0.04**
WikiWiki	Hard	-0.01	-0.00
AIDA-B	Hard	-0.04	-0.08
WNED-Wiki	Hard	-0.01	-0.06
WNED-Cweb	Hard	-0.01	-0.02
Average	**Hard**	**-0.02**	**-0.04**

Table 3: Ablation study showing the effects on our SumMC model by removing the mention condition on summary or the global context.

'hard' mentions. In comparison, Eigentheme, the current SoTA model, has slightly higher scores on 'easy' mentions on most datasets but performs worse on 'hard' mentions.

Comparison with Previous Models. Overall, SumMC achieves 63% precision, while Eigentheme scores 47%. Although SumMC has 1% less precision on 'easy' cases (75% vs. 76%), it outperforms Eigentheme on 'hard' cases by 26% (73% vs. 47%). Eigentheme assumes that gold entities in a document are topically related (Arora et al., 2021). It captures global context only using the relations between mentions while neglecting the texts in the document. However, this assumption might not always hold. Our model, in contrast, removes this assumption by producing a guided summary of texts in the document.

Effect of Global Context. We show the results of our ablation study in Table 3. On all datasets, SumMC outperforms the variation without having the summary guided by the mention (–Guide), which outperforms the variation without summarization (–Sum). This result shows the efficacy of not only using summaries as global contexts, but also forcing the summaries to contain information about the mention. Indeed, in many cases, we find that the mention might not be central to the document so that a standard summary might contain noise or insufficient signal for disambiguating the mention.

Interestingly, we observe that the performance gap between variations on WikiHow-Wikidata is relatively small. We speculate that WikiHow's instructional sentences are usually self-explanatory, so the local context often provides enough information to disambiguate the mention.

Effect of Multiple-Choice Selection. Using similarity measures to link a mention to an entity is one of the most successful EL methods (Pan et al., 2015). We also examine this approach using Sentence-BERT(Reimers and Gurevych, 2019) and cosine similarity instead of the multiple-choice selection model. As a result, it has only 42% P@1 on AIDA-B dataset. The text-based embedding approach might not be practical in our setting because entity candidates can only be represented by minimal texts, making text embedding unstable.

Error Analysis. In some cases, common sense is required to disambiguate mentions. For example, "Japan" in an article about a soccer tournament should be linked to the entity "Japan national football team" instead of the country "Japan." The correct answer can be inferred from the term "Asian Cup" in the text. However, our model fails such a case when the word 'soccer' is not included in the context.

Currently, each of our multiple choices is a concatenation of the target entity and its related entities based on two KB relations: `instance-of` and `subclass-of`. However, these might be insufficient. For example, most person entities have 'human' as the only related entity, which is uninformative. Conversely, considering other relations might also introduce unnecessary noise.

6 Conclusion

We introduce SumMC, a fully unsupervised Entity Linking model that first produces a summary of the document guided by the mention, and then casts the task to a multiple-choice format. Our model achieves new state-of-the-art performance on various benchmarks, including our new WikiHow-Wikidata, the first EL dataset on procedural texts. Notably, our approach of guided summarization may be applied to other tasks that benefit from global contexts. Future work might also extend our methods to supervised settings.

Limitations

Because we focus on fully unsupervised models, we do not consider fine-tuning GPT-3 nor provide a direct comparison with other supervised approaches.

A potential criticism of this work is our use of GPT-3. Although GPT-3 is publicly available to everyone, it is not an open-source model and can be expensive to use at scale.

For direct comparison, we use the candidate generation method from (Le and Titov, 2019) and Arora et al. (2021), which has a low recall on datasets. Although there are better methods (Sil et al., 2012; Charton et al., 2014), we do not consider them in this work.

Acknowledgements

This research is based upon work supported in part by the DARPA KAIROS Program (contract FA8750-19-2-1004), the DARPA LwLL Program (contract FA8750-19-2-0201), the IARPA BETTER Program (contract 2019-19051600004), and the NSF (Award 1928631). Approved for Public Release, Distribution Unlimited. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, IARPA, NSF, or the U.S. Government. We thank the students from the CIS-421/521 course in 2021 at the University of Pennsylvania for annotating WikiHow-Wikidata dataset.

References

Akhil Arora, Alberto Garcia-Duran, and Robert West. 2021. Low-rank subspaces for unsupervised entity linking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8037–8054, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yixin Cao, Lifu Huang, Heng Ji, Xu Chen, and Juanzi Li. 2017. Bridge text and knowledge by learning multi-prototype entity mention embedding. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1623–1633, Vancouver, Canada. Association for Computational Linguistics.

Eric Charton, Marie-Jean Meurs, Ludovic Jean-Louis, and Michel Gagnon. 2014. Improving entity linking using surface form refinement. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4609–4615, Reykjavik, Iceland. European Language Resources Association (ELRA).

Xiao Cheng and Dan Roth. 2013. Relational inference for wikification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1787–1796, Seattle, Washington, USA. Association for Computational Linguistics.

Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. 2018. Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1595–1604, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

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

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Filip Ilievski, Daniel Garijo, Hans Chalupsky, Naren Teja Divvala, Yixiang Yao, Craig Rogers, Ronpeng Li, Jun Liu, Amandeep Singh, Daniel Schwabe, and Pedro Szekely. 2020. KGTK: A toolkit for large knowledge graph manipulation and analysis. In *International Semantic Web Conference*, pages 278–293. Springer.

Phong Le and Ivan Titov. 2019. Distant learning for entity linking with automatic noise detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4081–4090, Florence, Italy. Association for Computational Linguistics.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460, Florence, Italy. Association for Computational Linguistics.

Isaiah Onando Mulang', Kuldeep Singh, Chaitali Prabhu, Abhishek Nadgeri, Johannes Hoffart, and Jens Lehmann. 2020. Evaluating the impact of knowledge graph context on entity disambiguation models. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2157–2160.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Xiaoman Pan, Taylor Cassidy, Ulf Hermjakob, Heng Ji, and Kevin Knight. 2015. Unsupervised entity linking with Abstract Meaning Representation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1130–1139, Denver, Colorado. Association for Computational Linguistics.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710.

Jonathan Raiman and Olivier Raiman. 2018. Deeptype: multilingual entity linking by neural type system evolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Avirup Sil, Ernest Cronin, Penghai Nie, Yinfei Yang, Ana-Maria Popescu, and Alexander Yates. 2012. Linking named entities to any database. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 116–127, Jeju Island, Korea. Association for Computational Linguistics.

Niket Tandon, Keisuke Sakaguchi, Bhavana Dalvi, Dheeraj Rajagopal, Peter Clark, Michal Guerquin, Kyle Richardson, and Eduard Hovy. 2020. A dataset for tracking entities in open domain procedural text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6408–6417, Online. Association for Computational Linguistics.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598, San Diego, California. Association for Computational Linguistics.

Johannes M van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P de Vries. 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.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

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.

Li Zhang. 2022. Reasoning about procedures with natural language processing: A tutorial.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020. Reasoning about goals, steps, and temporal ordering with WikiHow. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4630–4639, Online. Association for Computational Linguistics.

Wenzheng Zhang, Wenyue Hua, and Karl Stratos. 2021. Entqa: Entity linking as question answering. *arXiv preprint arXiv:2110.02369*.

Document
SOCCER - JAPAN GET LUCKY WIN, CHINA IN SURPRISE DEFEAT. Nadim Ladki AL-AIN, United Arab Emirates 1996-12-06 Japan began the defence of their Asian Cup title with a lucky 2-1 win against Syria in a Group C championship match on Friday. But China saw their luck desert them in the second match of the group, crashing to a surprise 2-0 defeat to newcomers Uzbekistan. China controlled most of the match and saw several chances missed until the 78th minute when Uzbek striker Igor Shkvyrin took advantage of a misdirected defensive header to lob the ball over the advancing Chinese keeper and into an empty net. Oleg Shatskiku made sure of the win in injury time, hitting an unstoppable left foot shot from just outside the area. The former Soviet republic was playing in an Asian Cup finals tie for the first time. Despite winning the Asian Games title two years ago, Uzbekistan are in the finals as outsiders. Two goals from defensive errors in the last six minutes allowed Japan to come from behind and collect all three points from their opening meeting against Syria. Takuya Takagi scored the winner in the 88th minute, rising to head a Hiroshige Yanagimoto cross towards the Syrian goal which goalkeeper Salem Bitar appeared to have covered but then allowed to slip into the net. It was the second costly blunder by Syria in four minutes. Defender Hassan Abbas rose to intercept a long ball into the area in the 84th minute but only managed to divert it into the top corner of Bitar's goal. Nader Jokhadar had given Syria the lead with a well-struck header in the seventh minute. Japan then laid siege to the Syrian penalty area for most of the game but rarely breached the Syrian defence. Bitar pulled off fine saves whenever they did. Japan coach Shu Kamo said: "The Syrian own goal proved lucky for us. The Syrians scored early and then played defensively and adopted long balls which made it hard for us." Japan, co-hosts of the World Cup in 2002 and ranked 20th in the world by FIFA, are favourites to regain their title here. Hosts UAE play Kuwait and South Korea take on Indonesia on Saturday in Group A matches. All four teams are level with one point each from one game.

Mention	Summary
-	Japan began the defence of their Asian Cup title with a lucky 2-1 win against Syria in a Group C championship match on Friday.
Japan	Japan won 2-1 against Syria in the first game of the Asian Cup, while China lost 2-0 to Uzbekistan in the second game of the group.
Syria	Syria lost to Japan 2-1 in the Asian Cup championship, with two late goals coming from defensive errors.
Uzbekistan	Uzbekistan defeated China 2-0 in their first match of the Asian Cup, surprising many observers.

Table 4: Example of guided summarization on '1163testb_soccer' document in AIDA-B dataset.

A Examples of Guided Summarization

Based on the document '1163testb_soccer' in the AIDA-B dataset, we show examples of guided summarization in Table 4. In the first example, the model generates a general document summary since it is not guided with a mention. Thus, information about Uzbekistan is not shown in the summary. The latter three examples are guided with 'Japan', 'Syria', and 'Uzbekistan', and give corresponding summaries specified to the mention.

We also provide example guided summaries of the AIDA-B dataset, which can be found in the uploaded file.

B Creation of WikiHow-Wikidata

Our annotation interface shows example sentences from a Wikihow article and asks the annotator to select the correct sense of one of the three most frequent nouns. Our inventory of senses is a numbered list of possible Wikidata candidate entities, along with a short description of each sense. Participants read the article and select the word sense by picking the closest match from the candidate list or choosing "No Answer" if there is none. Annotators can also input multiple answers if more than one candidate matches the correct sense inferred from example sentences. We do not force participants to input only one answer because it is common in Wikidata that multiple entities describe the same meaning. Our program records the WikiHow article URL, target mention, and the corresponding Wikidata QID students selected. We manually annotated 30 questions for control questions. The program shows a random control question for every ten questions without telling participants. The annotation program is available in the uploaded file.

Eventually, we collect 31,354 responses from 521 participants. We then filtered qualifying participants so that only those with more than 95% accuracy on confident control questions remain. Hence, we end up with a cleaned set of 23,352 responses.

In order to apply to different models examined in our paper, we do further filtering on the cleaned set. We run the candidate generation mentioned in Section 2, and exclude entities that cannot be found in the list of DEEPWALK (Perozzi et al., 2014) graph embedding trained on Wikidata by Arora et al. (2021). Also, we drop mentions with a candidate list that does not have a gold entity or has only one entity in the list. As a result, we get a final set of 11,287 mentions.

C Effect of GPT-3 Engine Size

We also compare the impact of GPT-3 engine size to SumMC model. Guided summarization is very powerful regardless of the engine. Only changing engine size, our model with Ada achieves 0.631 P@1, and Babbage scores 0.633 P@1 on AIDA-B, which tie with 0.636 P@1 by Curie. This gives an alternative option to users with a limited budget but who still want a moderate performance. Compared to Curie, the pricing of Ada is 87% cheaper, but it is still equivalent to the result that Curie achieved. On the other hand, multiple-choice selection requires a large model. Compared with the 0.633 P@1 on AIDA-B with Davinci engine, Curie and Babbage only score 0.204 and 0.196 P@1, respectively, while the Ada engine fails to complete the evaluation.

Using our model's setting, it costs around $0.002 for guided summarization and $0.01 for multiple-choice selection.

D Model Setting Details

Since most of our code is API call of GPT-3, SumMC does not require a strong requirement on computational resources.

In our model, we used default hyperparameter setting for both guided summarization and multiple-choice selection. In detail, we set temperature=0.7, max_tokens=256, top_p=1. frequency_penalty=0, and presence_penalty=0.

Due to the input token limit of GPT-3 engines, we truncated the input document to 512 words surrounding the target mention during guided summarization.

We used the '2021-09-13' dump of Wikidata in our model, and used Knowledge Graph Toolkit (Ilievski et al., 2020) to extract entities and their relations.