SHADE: SEMANTIC HYPERNYM ANNOTATOR FOR DOMAIN-SPECIFIC ENTITIES - DUNGEONS AND DRAGONS DOMAIN USE CASE

Akila Peiris

We will cover…

Introduction

Related Work

Rationale

SHADE

Discussion

Conclusion

INTRODUCTION

- Data annotation
  - Labelling data
  - Teams of annotators
  - Experts/ non-experts
- Annotation consistency
  - Inter Annotator Consistency (IAA)
- Annotation app SHADE
  - Forgotten Realms Wiki (FRW) dataset[1]
  - Entity classification

[1] A. Peiris and N. de Silva, “Synthesis and Evaluation of a Domain-specific Large Data Set for Dungeons & Dragons,” in PACLIC, Oct. 2022, pp. 415–424.

INTRODUCTION

FRW DATASET

- Dungeons and Dragons domain
- Extracted from Forgotten Realms wiki (https://forgottenrealms.fandom.com/)
- Automatically extracted entities using Wikia *First Link*
  - First link - first link found in a Wikipedia page*
  - Wikipedia guidelines
  - Typically points to the higher, abstract entity (super class)
  - Forms a "Classification Chain"
- But there are ambiguities

[1] A. Peiris and N. de Silva, "Synthesis and Evaluation of a Domain-specific Large Data Set for Dungeons & Dragons," in PACLIC, Oct. 2022, pp. 415–424.

* One that does not refer a pronunciation guide or other miscellaneous page and points to another Wikipedia article. https://en.wikipedia.org/wiki/Wikipedia:Manual of Style/Lead section

RELATED WORK

- Crowdsourcing annotations [2][3]
  - Efficiency and scalability
  - 4 non-expert annotations per item ≈ expert level annotation
- Active learning techniques to reduce the amount of annotation required[4]
  - Human input for complex cases
- Inter-annotator agreement (IAA) [5][6]
  - Consistency of annotations produced by different annotators
  - Annotator expertise
  - Annotation guidelines
  - Annotation complexity

[2] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng, "Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks," in Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii: Association for Computational Linguistics, Oct. 2008, pp. 254–263. [Online].
[3] A. Dumitrache, L. Aroyo, and C. Welty, "Achieving expert-level annotation quality with crowdtruth," in Proc. of BDM2I Workshop, ISWC, 2015.
[4] F. Olsson, "A literature survey of active machine learning in the context of natural language processing," 2009.
[5] E. Ouyang, Y. Li, L. Jin, Z. Li, and X. Zhang, "Exploring n-gram character presentation in bidirectional rnn-crf for chinese clinical named entity recognition," in CEUR workshop proceedings, vol. 1976, 2017, pp. 37–42.
[6] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," Computational linguistics, vol. 34, no. 4, pp. 555–596, 2008

RATIONALE

- First links

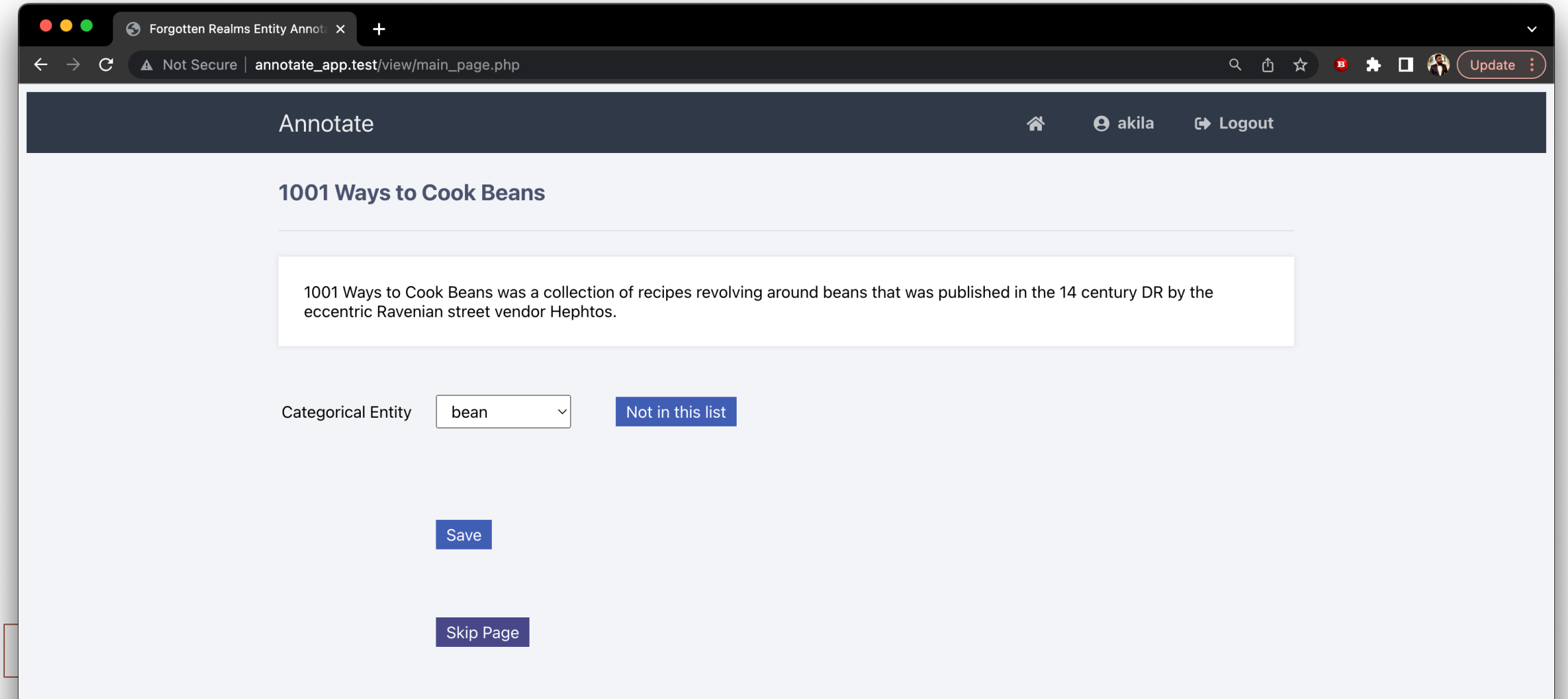
  “ Tiamat was the [lawful evil] [dragon] goddess of greed, queen of [evil dragons] and, for a time, reluctant servant of the [greater gods] [Bane] and later [Asmodeus]. Before entering the [Faerûnian pantheon], she was a member of the [Draconic pantheon], and for some time she was also a member of the [Untheric pantheon]. “

- Which is it?
  - Dragon, diety
- First link?
  - Lawful evil

RATIONALE

- Why SHADE?
  - Why not a spreadsheet?
- Consistency of annotations
  - Different annotators, different patterns
- Control the annotation options
  - Non-paged entity categories
- Reduce human error
  - Faerûnian vs. Faerunian
  - Outright wrong spellings leading to different label
- All of the above can be enforced via a task specific app

SHADE APP


Forgotten Realms Entity Annot
Not Secure | annotate_app.test/view/main_page.php
Update
Annotate
akila
Logout
1001 Ways to Cook Beans
1001 Ways to Cook Beans was a collection of recipes revolving around beans that was published in the 14 century DR by the eccentric Ravenian street vendor Hephtos.
Categorical Entity
bean
Not in this list
Save
Skip Page

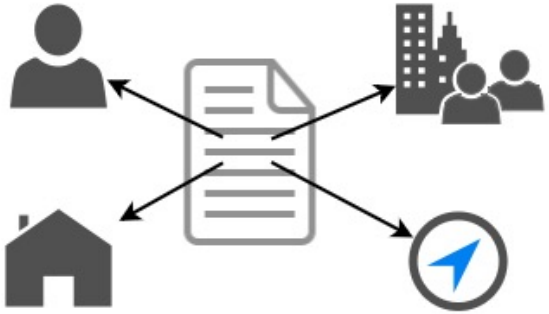
SHADE APP

PROCESS FLOW

Select from a list of links

No suitable match in list

Select from a list of named entities



No suitable match in list

Tye in the category

Click “Not in this list” button

Click “Not in this list either” button

SHADE APP

LINK LIST

- Forgotten Realms Wiki*
- XML and Markdown text
- Anatomy of a link: [[Linked_page_name|Display_text]]
- Cannot use parsers
  - Parsers remove links
- First link
- Remove infobox
  - When count(“[[“) == count(“]]”)
  - And next few characters do not contain “[[“

* https://forgottenrealms.fandom.com/

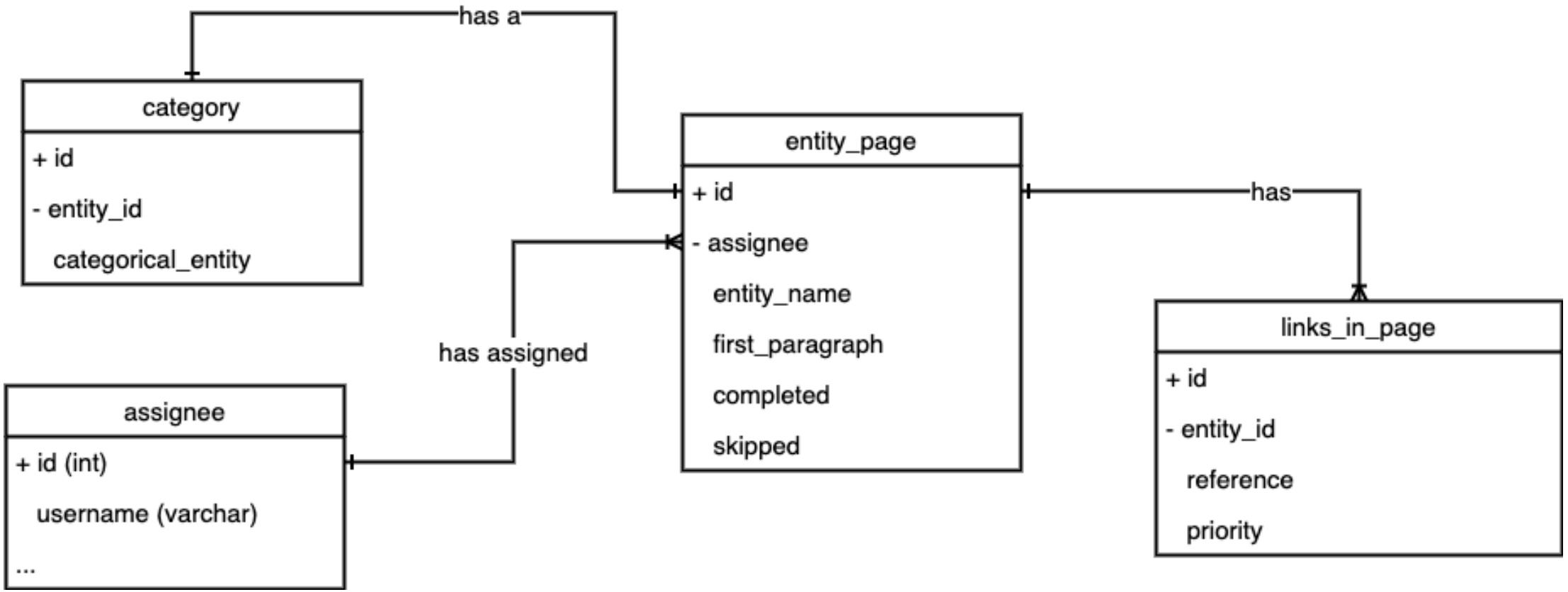
SHADE APP

NOUN-PHRASE LIST AND TEXT INPUT

- When “Not in this list”
  - First paragraph
  - Clean text
  - Extract nouns and phrases
- When “Not in this list either”
  - Enter manually

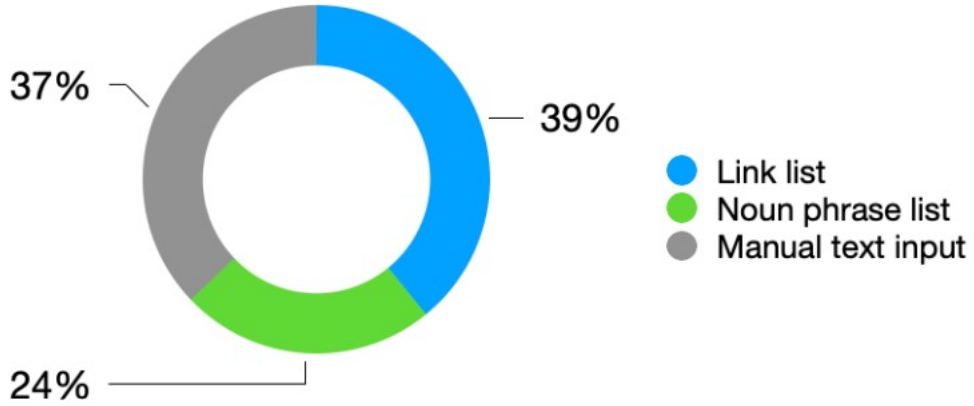
SHADE APP

ER DIAGRAM



DISCUSSION

- At the time of writing the paper,
  - Total annotations: 3984
  - Entities other than “years”: 1020
  - From links list: 399
  - From noun phrases list: 242
  - Typed in: 379



- 2/3 of the annotations come from lists
  - Proves that SHADE can reduce manual text inputs during annotation significantly

CONCLUSION AND FUTURE WORK

- How to structure an application used for data annotations
- Easily be ported to other data sources using MediaWiki stack
- Can be used even if no corpora, if annotation option space can be procured
- Improvements
  - Text suggestion feature (from existing labels)
  - Multiple annotations per same entity

REFERENCES

[1] A. Peiris and N. de Silva, “Synthesis and Evaluation of a Domain-specific Large Data Set for Dungeons & Dragons,” in PACLIC, Oct. 2022, pp. 415–424.

[2] R. Snow, B. O’Connor, D. Jurafsky, and A. Ng, “Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks,” in Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii: Association for Computational Linguistics, Oct. 2008, pp. 254–263. [Online].

[3] A. Dumitrache, L. Aroyo, and C. Welty, “Achieving expert-level annotation quality with crowdtruth,” in Proc. of BDM2I Workshop, ISWC, 2015.

[4] F. Olsson, “A literature survey of active machine learning in the context of natural language processing,” 2009.

[5] E. Ouyang, Y. Li, L. Jin, Z. Li, and X. Zhang, “Exploring n-gram character presentation in bidirectional rnn-crf for chinese clinical named entity recognition,” in CEUR workshop proceedings, vol. 1976, 2017, pp. 37–42.

[6] R. Artstein and M. Poesio, “Inter-coder agreement for computational linguistics,” Computational linguistics, vol. 34, no. 4, pp. 555–596, 2008

THANK YOU