Living Dictionary: Automatically Generating Wikipedia-like Content

Nishant Balepur, Professor Kevin Chen-Chuan Chang
Siebel Center for Computer Science, Grainger College of Engineering, University of Illinois at Urbana-Champaign

INTRODUCTION

Problems with Current Wikipedia

- Computer science is growing exponentially, so it becomes more challenging to keep track of the field
- Without up-to-date information, entering computer science becomes more difficult

Our Approach

- Recent advancements in text generation employ structured outlines for a more informative output
- We adopt a similar approach with LivDic by creating a content tree rooted in the specific keyword

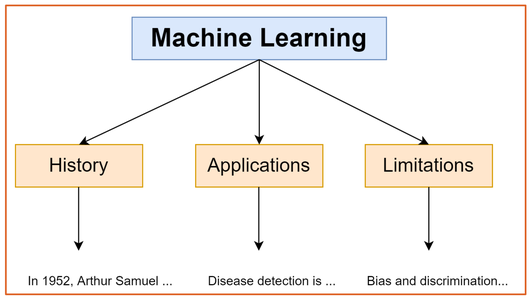

Figure 1: An example content tree for machine learning

PURPOSE

What Previous Work Exists?

- Past work focuses on generating the Wikipedia abstract [1] or needs references as input [2]
- With LivDic, we solve these issues with a novel approach to text summarization

Why does LivDic Matter?

- We enable information in computer science to become more accessible to beginners
- Improving upon information retrieval supports the idea of knowledge freedom as opposed to commoditization
- We can avoid historical problems of outdated or biased information from a single source

METHOD

Our LivDic framework has two subproblems:
1. Identifying sections for our content tree
2. Generating text under each section

Section Identification

- Related fields to machine learning, like databases and NLP, should have similar section titles
- We can select the top 5 most frequent sections under the keyword's category to structure our content tree

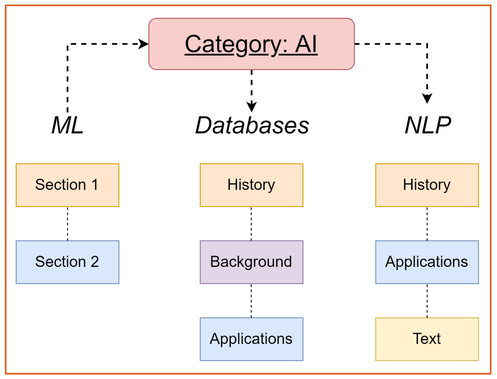

Figure 2: Mining visualization for machine learning

Instance Classifier

- Hypothesis: Wikipedia sections can be split into a list of "instances"
- An instance can broadly be defined as a topic or idea within a certain section (dates, people, etc.)
- If we teach LivDic what instances of a section should look like, we can use it to identify instances on websites and use that as our content

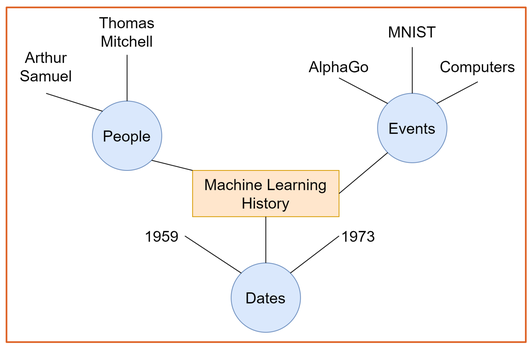

Figure 2: Instance knowledge graph for machine learning history

RESULTS

Section Identification

- Framework is simple but effective: finds highly plausible candidates

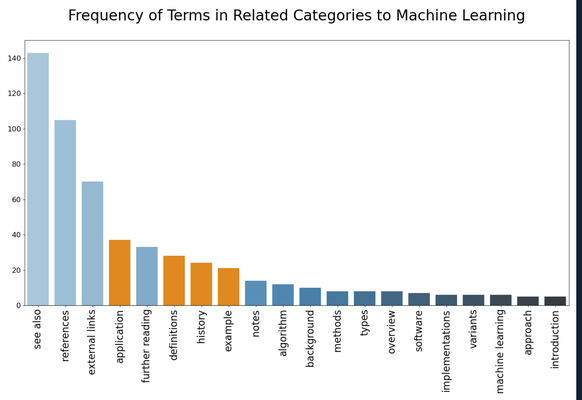

Figure 4: Mined sections for machine learning

Instance Classifier

- LivDic is evaluated using recall, the ability to identify true positives
- Training, validation, and testing sets boast impressive recall values of **1.0**, **0.8047**, and **1.0**, respectively

Sample Article Generation

- LivDic's summaries contain more in-depth information than current techniques [3]
- Our content is thorough, but not too long, preserving around **40%** of the original article

History of Machine Learning

Businesses increasingly rely on tools that use ML algorithms to provide them with accurate data for improving and growing their organization.

While machine learning may seem like a very recent concept, you may be surprised to know that the history of machine learning dates back to the 1940s. However, it wasn't until the 1950s that we saw how ML works for the first time.

. . .

Figure 5: Generated section for 'history of machine learning'

CONCLUSIONS

Contributions

- LivDic is highly proficient at completing its two subproblems
- It can identify probable sections for our content tree and can extract summaries for these sections

Future Work

- Fixing current issues: source article selection and eliminating extraneous web scraped text
- Generating a large dataset for computer science keywords
- Evaluating on traditional ROUGE metrics for more empirical results
- Using insights from human perspectives on readability, relevancy and redundancy

REFERENCES

[1] Liu, Peter J., et al. "Generating Wikipedia by Summarizing Long Sequences." ArXiv:1801.10198 [Cs], Jan. 2018. arXiv.org, http://arxiv.org/abs/1801.10198.

[2] Zhu, Fangwei, et al. "TWAG: A Topic-Guided Wikipedia Abstract Generator." ArXiv:2106.15135 [Cs], June 2021. arXiv.org, http://arxiv.org/abs/2106.15135.

[3] Gambhir, Mahak, and Vishal Gupta. "Recent Automatic Text Summarization Techniques: A Survey." Artificial Intelligence Review, vol. 47, no. 1, Jan. 2017, pp. 1–66. Springer Link, https://doi.org/10.1007/s10462-016-9475-9.

ACKNOWLEDGEMENTS

I would like to thank *Professor Kevin Chen-Chuan Chang* for his continued support throughout this project. Also, a special thanks to *Dr. Natasha Mamaril* and the *ISUR/C3SR* teams for making this experience possible and opening my eyes up to the world of research.

ILLINOIS