Refining an ontology of NLP research concepts

Karim Arabi

17/11/2022, Thesis kick-off presentation

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
wwwmatthes.in.tum.de

Outline

Introduction

- Problem Statement and Motivation

Methodology

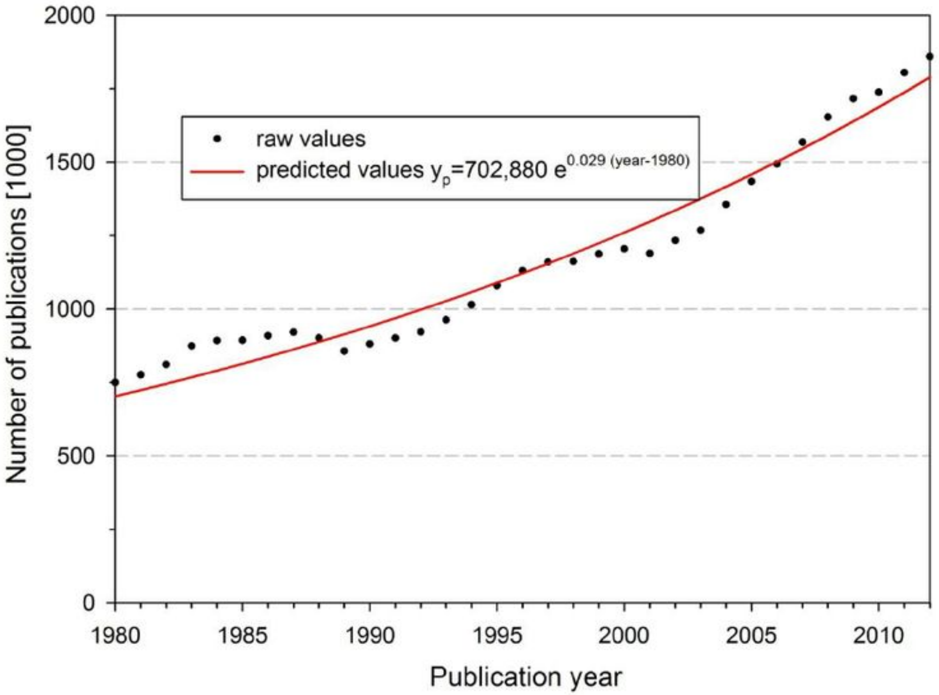
- Research Questions
- Proposed Solutions
- Evaluation Methods and Initial Results

Timeline

Problem Statement and Motivation

With the ever-expanding purview of available research studies and documents becoming available, the discoverability of such papers has become challenging

A domain-specific ontology would satisfy this issue, providing a search through semantic understanding



Bornmann, Lutz & Mutz, Ruediger. (2014). Growth rates of modern science: A bibliometric analysis.

Goal

Construct an automated ontology of NLP concepts and publications that users can browse through and explore

<u>Deliverable:</u> Ontology of NLP research concepts

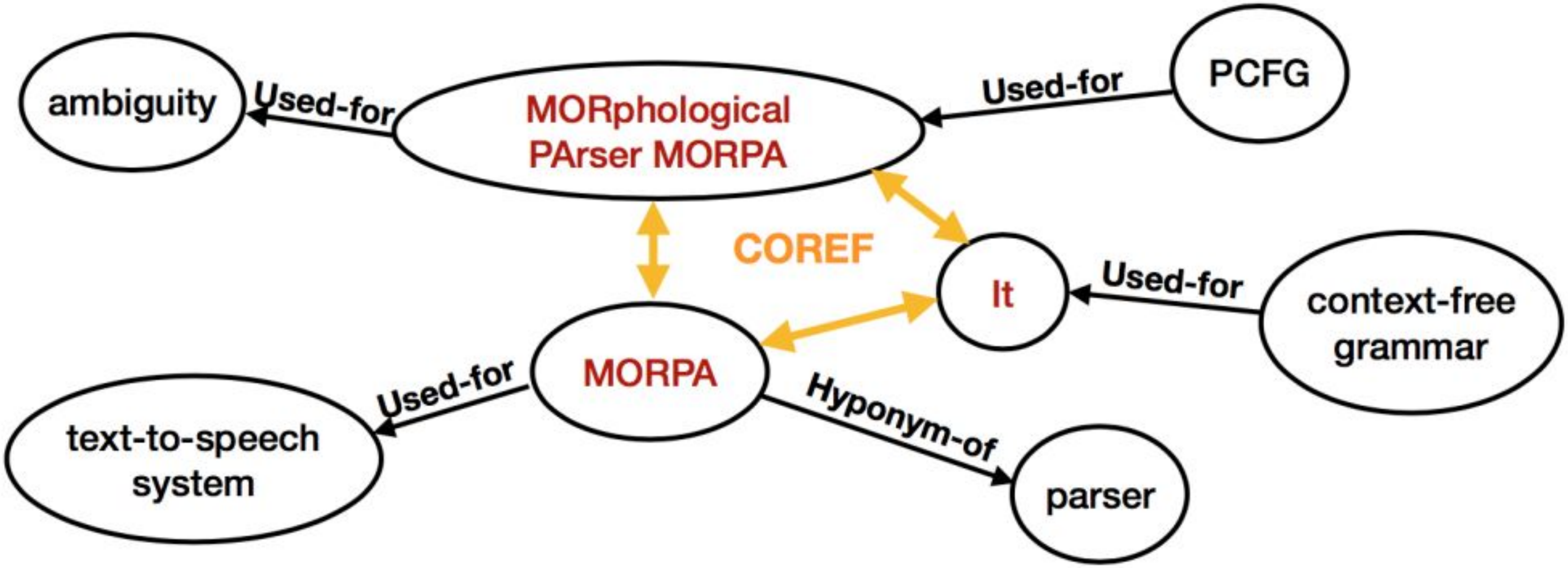


Figure 1: example of an NLP domain ontology

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. (2018). Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction.

Previous Work Completed

- **Learning Hierarchical Relations between Research Concepts from Abstracts and Titles of NLP Publications - Simon Klimek**

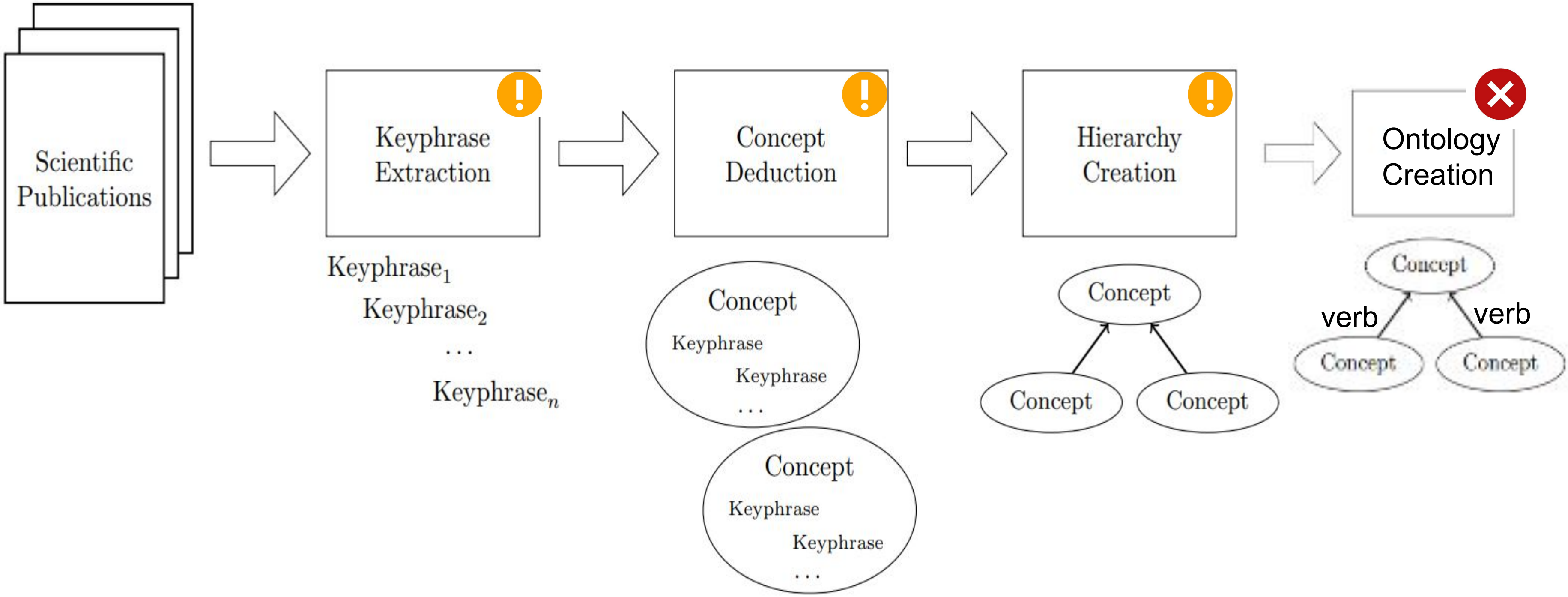


Figure 2: Pipeline of taxonomy creation steps in Simon Klimek's thesis.

Klimek, S. (2022). Learning Hierarchical Relations between Research Concepts from Abstracts and Titles of NLP Publications

Previous Work Completed

Keyphrase Extraction

- Ranking of keyphrase candidates by cosine-similarity of keyphrase and document embeddings (by best ‘document representation’).
- K-means algorithm to manually remove off-topic keyphrases.
- Extracted keyphrases are unsanitized

Concept Deduction

- Bert-based lexical substitution to generate list of substitutes for every keyphrase + merging if overlap of substitutes is > 5%.
- Underperforms with multi-word keyphrase substitution and merging.

Hierarchy Creation

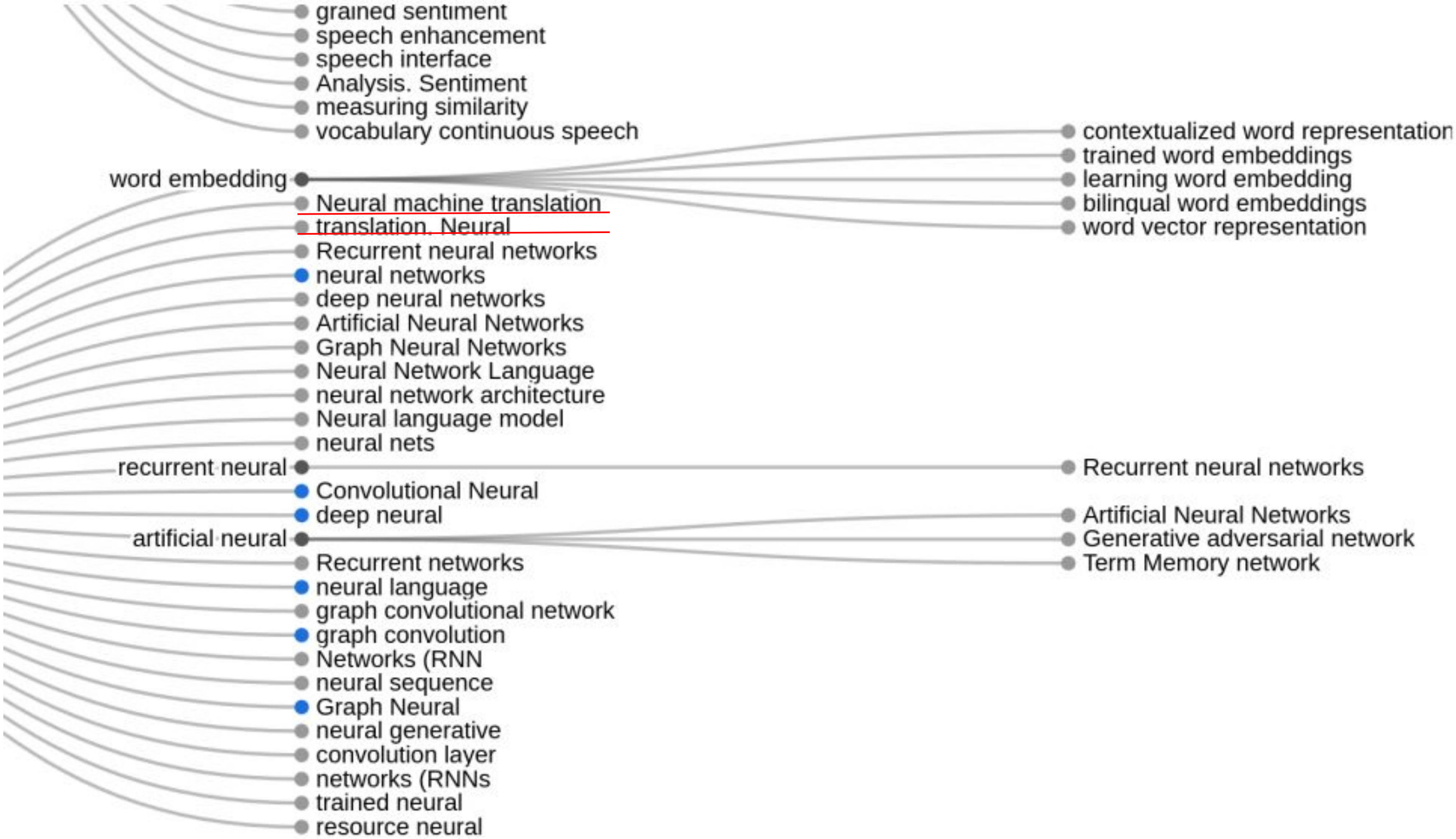
- Subsumption Method for edge creation.
- Simple solution due to time constraints.

Schopf, T.; Klimek, S. and Matthes, F. (2022). ***PatternRank: Leveraging Pretrained Language Models and Part of Speech for Unsupervised Keyphrase Extraction****. In Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR, ISBN 978-989-758-614-9; ISSN 2184-3228, pages 243-248.*

Klimek, S. (2022). Learning Hierarchical Relations between Research Concepts from Abstracts and Titles of NLP Publications

Previous Work Completed

Improvements to be made



Klimek, S. (2022). Learning Hierarchical Relations between Research Concepts from Abstracts and Titles of NLP Publications

Research Questions

- **RQ1: How to use manual refinement to improve top-level navigation for users?**

- **RQ2: How to enhance the existing concepts and relations through automated refinement approaches?**

- **RQ3: How to transition from a taxonomy to an ontology with more complex relations?**

Project Plan: Step 1

- **Manually define first layers of NLP taxonomy for higher-quality navigation**

Why: The microsoft academic graph (an outdated but similar concept) found clearly defined top level-navigation is important for users.

How: Inspired by ACL conferences, NLP surveys, and CSO ontology.

Evaluation: 2-part process

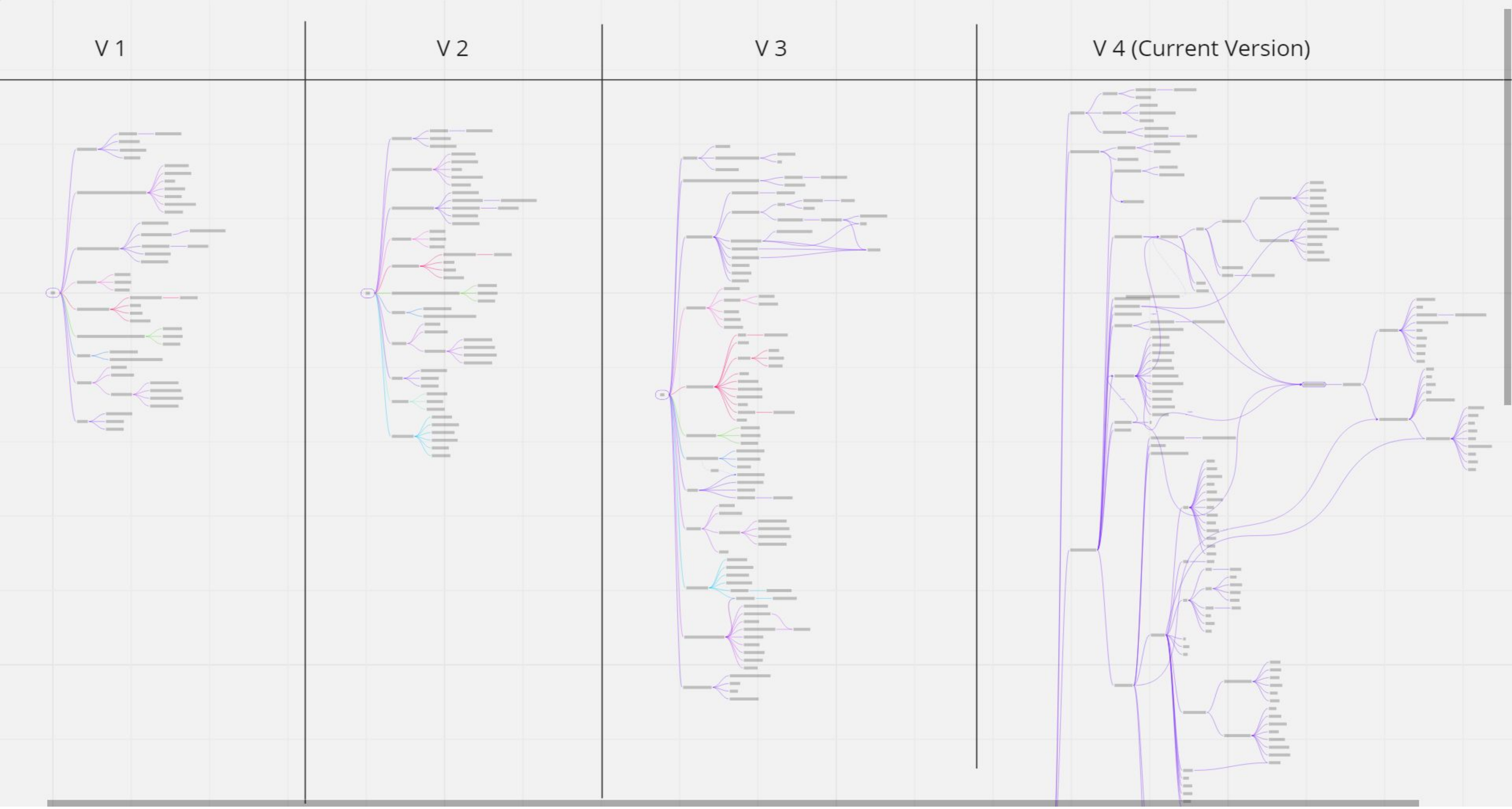
- Loosely-structured interviews with domain experts (researchers in the field). (completed)
- Quantitative user tests with domain experts to measure the clarity and ease-of-use of our manual ontology.
  (incomplete)

Figure 4: example of manually defined top 3 layers of NLP taxonomy.

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, Anshul Kanakia; Microsoft Academic Graph: When experts are not enough. Quantitative Science Studies 2020; 1 (1): 396–413.
Info on CSO: https://cso.kmi.open.ac.uk/home
Info on ACL: https://www.aclweb.org/portal/

Initial Results of Step 1

- **6 loosely-structured interviews with NLP researchers**
- **Iterated Ontology design process**



Project Plan: Step 2

- **Enhance concept and hierarchy inference**

Why: Weaknesses in current implementation can be improved.

How:

- Sanitize extracted keyphrases (such as acronym conglomeration)
  (completed)
- Improve substitute generation by investigating alternative solutions (such as BART-LS)
  (incomplete)
- Concept merging solutions (such as SciConceptMiner)
  (completed)
- Alternative taxonomy relation construction such as a weighted ensemble method (of Subsumption method and Lexical Syntactic method).
  (not started)

Evaluation: Use of user studies to evaluate concept coherence and hierarchical relations (as per the thesis that this topic builds upon).

A. Cattan, A. Eirew, G. Stanovsky, M. Joshi, and I. Dagan. (2020). Streamlining Cross-Document Coreference Resolution: Evaluation and Modeling

Lexical Syntactic Method

```
1. such KEYPHRASE as (KEYPHRASE,)* (and|or) (KEYPHRASE ,)+

2. (KEYPHRASE,?)+ (and|or) other KEYPHRASE

3. KEYPHRASE, (especially|including) (KEYPHRASE,)+ (and|or) KEYPHRASE
```

Subsumption Method

$$\exists k \in C_1, \exists k' \in C_2 : P(k|k') \geq \alpha \wedge P(k'|k) < 1 \Rightarrow (C_2, C_1) \in E.$$

$$P(x|y) = \frac{\#\text{sentences contain } x \text{ and } y}{\#\text{sentences contain } y}.$$

Klimek, S. (2022). Learning Hierarchical Relations between Research Concepts from Abstracts and Titles of NLP Publications

Initial Results of Step 2

- **SciConceptMiner concept merging approach**

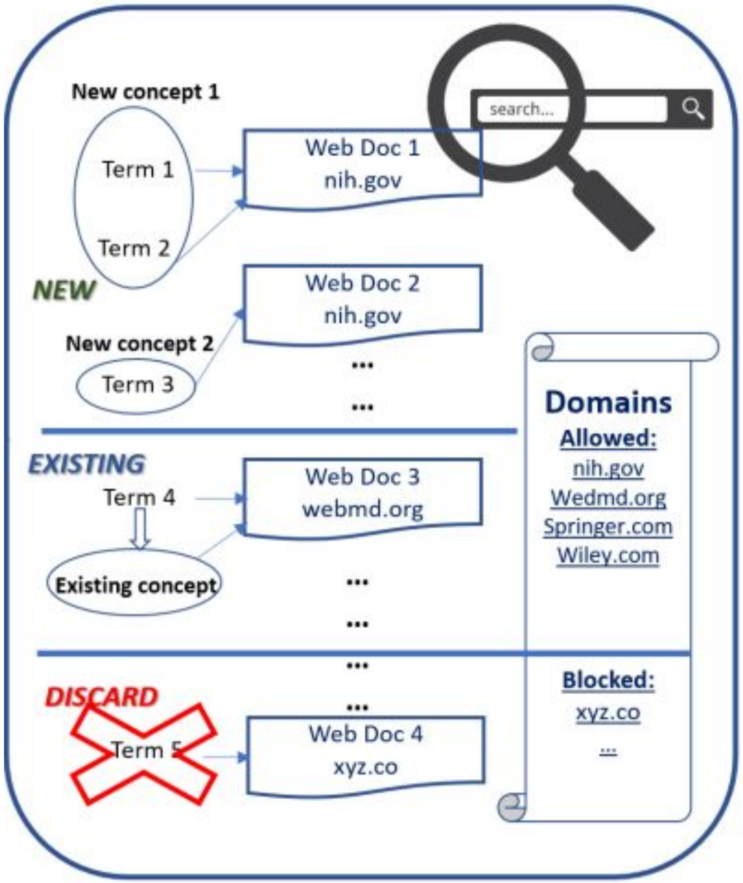
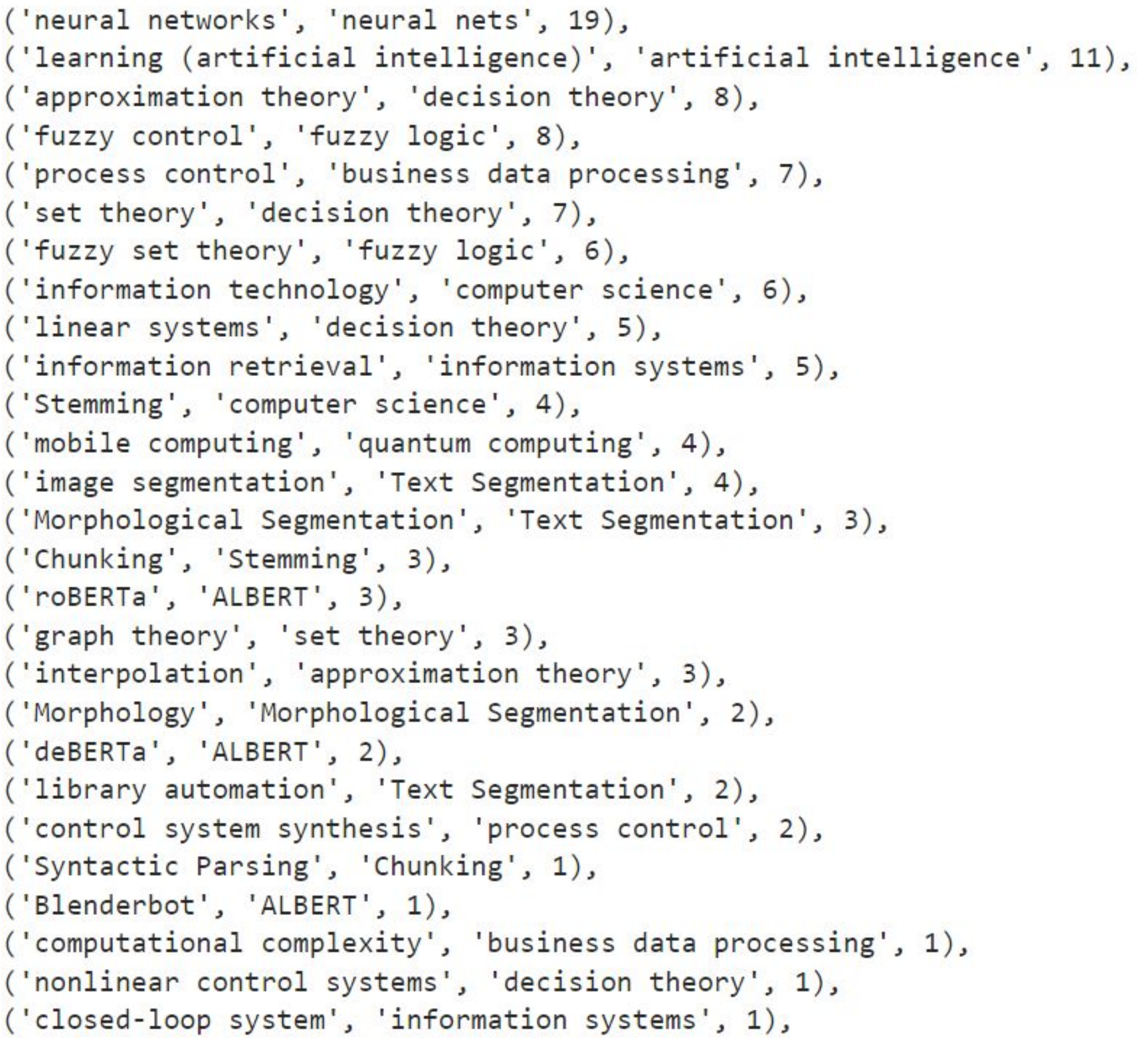


Figure 6: SciConcptMiner Approach



```
('neural networks', 'neural nets', 19),
('learning (artificial intelligence)', 'artificial intelligence', 11),
('approximation theory', 'decision theory', 8),
('fuzzy control', 'fuzzy logic', 8),
('process control', 'business data processing', 7),
('set theory', 'decision theory', 7),
('fuzzy set theory', 'fuzzy logic', 6),
('information technology', 'computer science', 6),
('linear systems', 'decision theory', 5),
('information retrieval', 'information systems', 5),
('Stemming', 'computer science', 4),
('mobile computing', 'quantum computing', 4),
('image segmentation', 'Text Segmentation', 4),
('Morphological Segmentation', 'Text Segmentation', 3),
('Chunking', 'Stemming', 3),
('roBERTa', 'ALBERT', 3),
('graph theory', 'set theory', 3),
('interpolation', 'approximation theory', 3),
('Morphology', 'Morphological Segmentation', 2),
('deBERTa', 'ALBERT', 2),
('library automation', 'Text Segmentation', 2),
('control system synthesis', 'process control', 2),
('Syntactic Parsing', 'Chunking', 1),
('Blenderbot', 'ALBERT', 1),
('computational complexity', 'business data processing', 1),
('nonlinear control systems', 'decision theory', 1),
('closed-loop system', 'information systems', 1),
```

https://aclanthology.org/2021.acl-demo.6.pdf

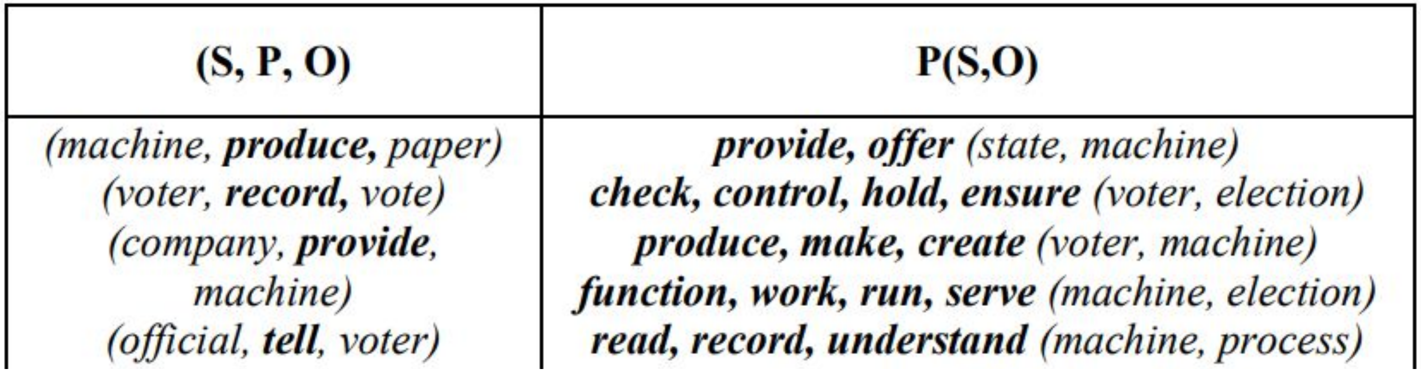
Project Plan: Step 3

- **Add more complex non-taxonomic relations (not started)**

Why: Allows for deeper semantic topic exploration than parent-child (hypernym-hyponym relations)

How: Investigate possible relation extraction methods, such as a ranking of concept-pair verbal dependencies, or SCICERO’s path-based relationship extractor module.

Evaluation: Use of user studies to evaluate non-taxonomic relations.



(S, P, O)	P(S,O)
*(machine, **produce**, paper)* *(voter, **record**, vote)* *(company, **provide**, machine)* *(official, **tell**, voter)*	***provide, offer*** *(state, machine)* ***check, control, hold, ensure*** *(voter, election)* ***produce, make, create*** *(voter, machine)* ***function, work, run, serve*** *(machine, election)* ***read, record, understand*** *(machine, process)*

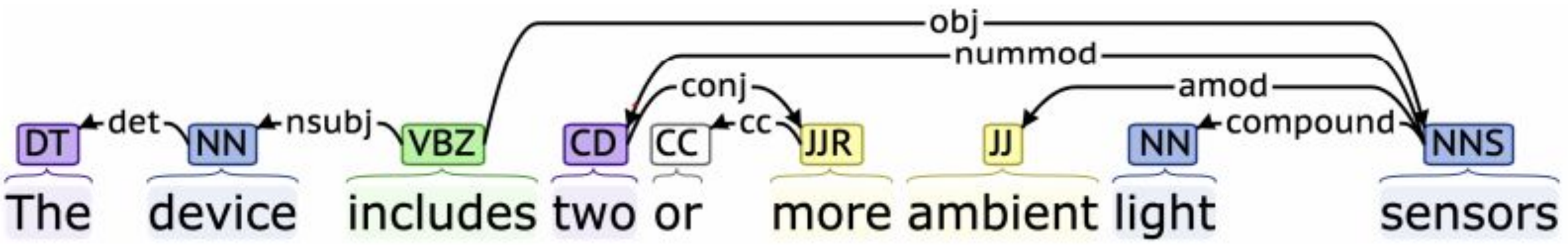
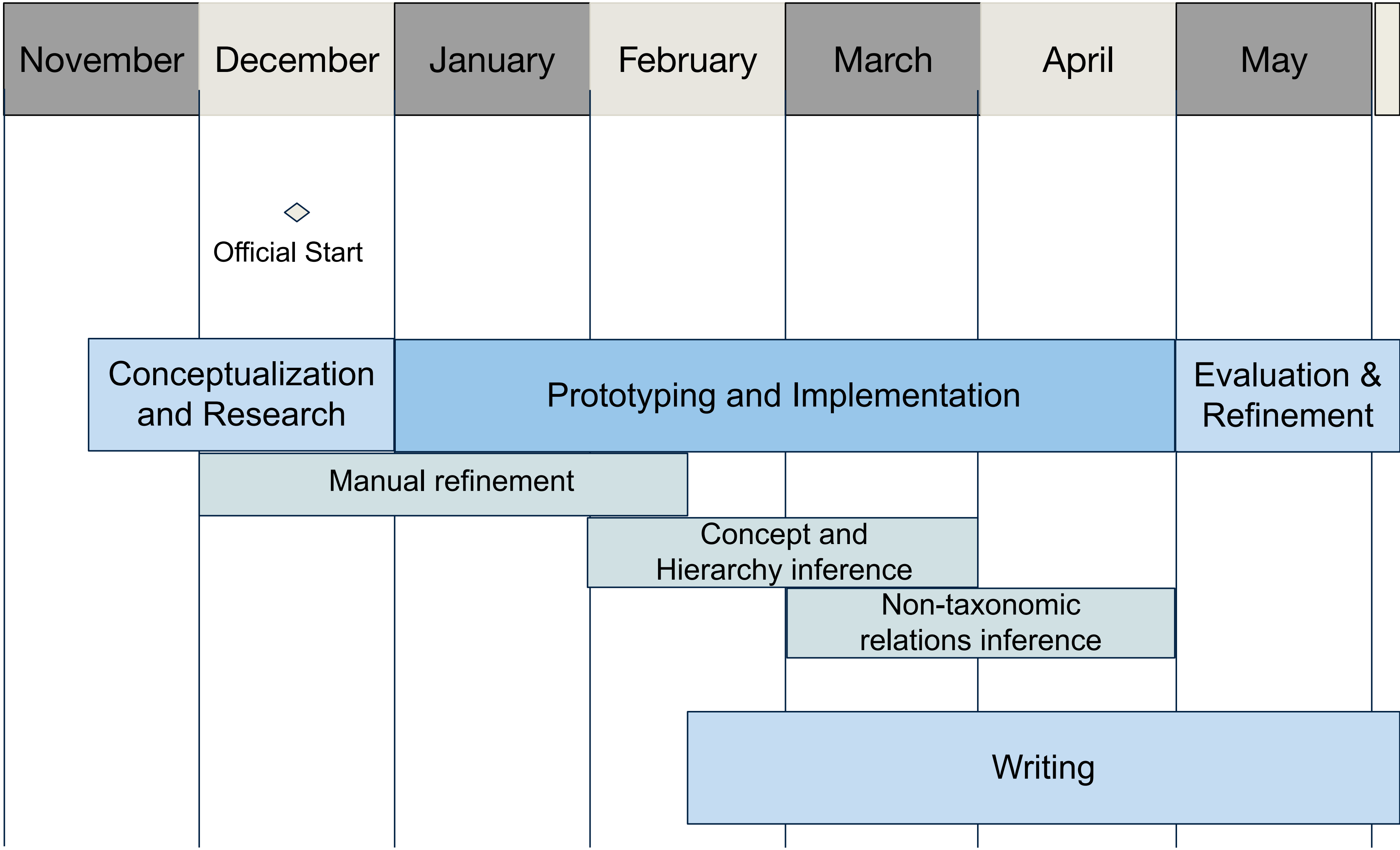


Figure 7 & 8: non-taxonomic relation (verbal) formed between topics.

N. F. Nabila, A. Mamat, M. A. Azmi-Murad and N. Mustapha, "Enriching non-taxonomic relations extracted from domain texts," *2011 International Conference on Semantic Technology and Information Retrieval*, 2011, pp. 99-105,

Timeline



November

December

January

February

March

April

May

Official Start

Conceptualization and Research

Prototyping and Implementation

Evaluation & Refinement

Manual refinement

Concept and Hierarchy inference

Non-taxonomic relations inference

Writing

Karim Arabi

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for
Business Information Systems

Boltzmannstraße 3
85748 Garching bei München

ge75yud@mytum.de