Refining an Ontology of NLP Research Concepts

Karim Arabi

17/07/2023, Thesis final 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
- Implemented Solutions

Evaluation

- Evaluation Methods
- Results

Conclusion and Future Work

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

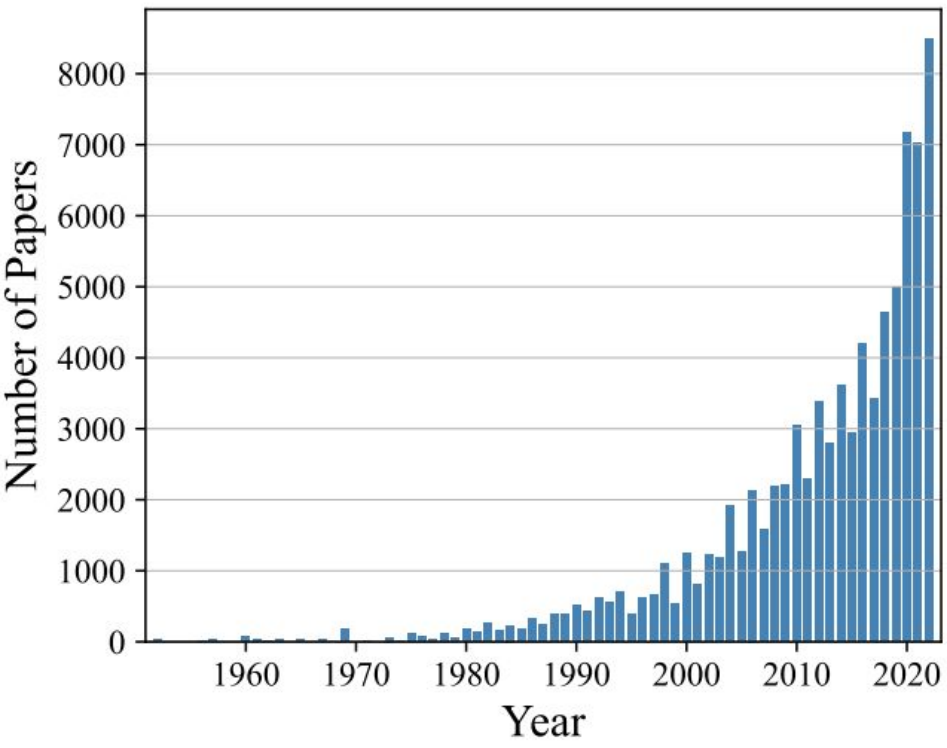


Figure 1: # of publications added to ACL Anthology over years.

Info on ACL: https://www.aclweb.org/portal/

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

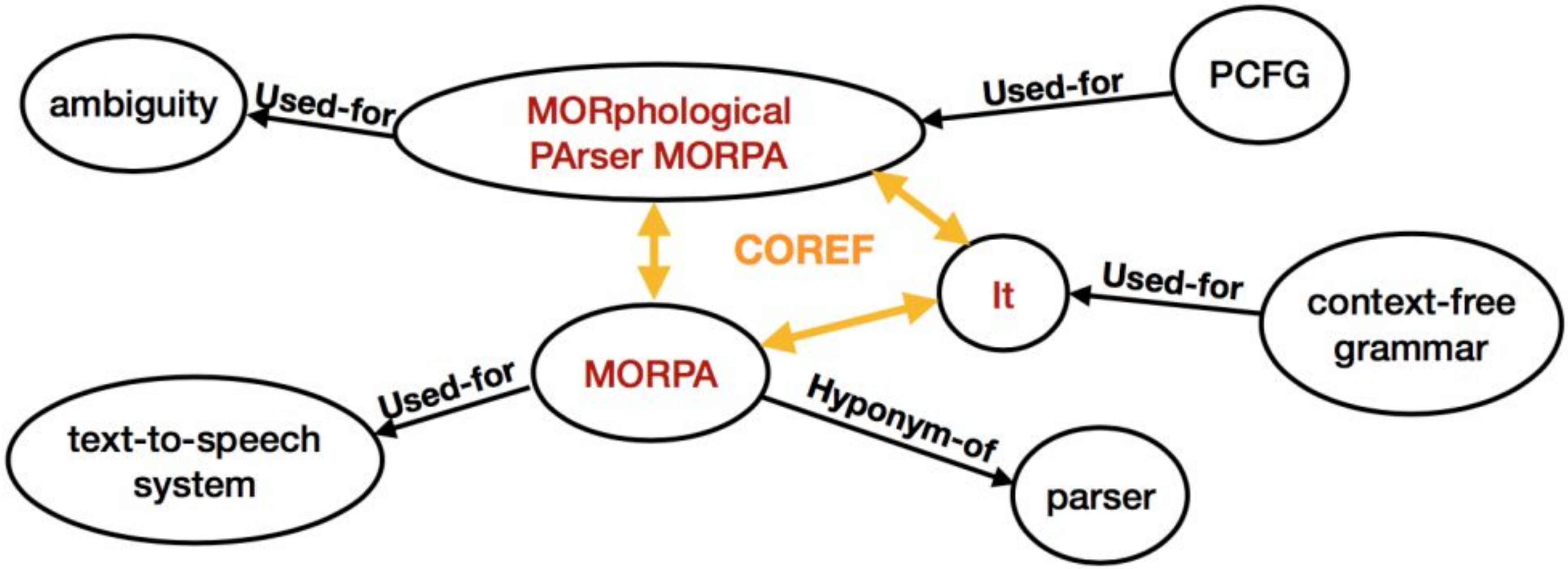


Figure 2: example of an NLP domain ontology

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**

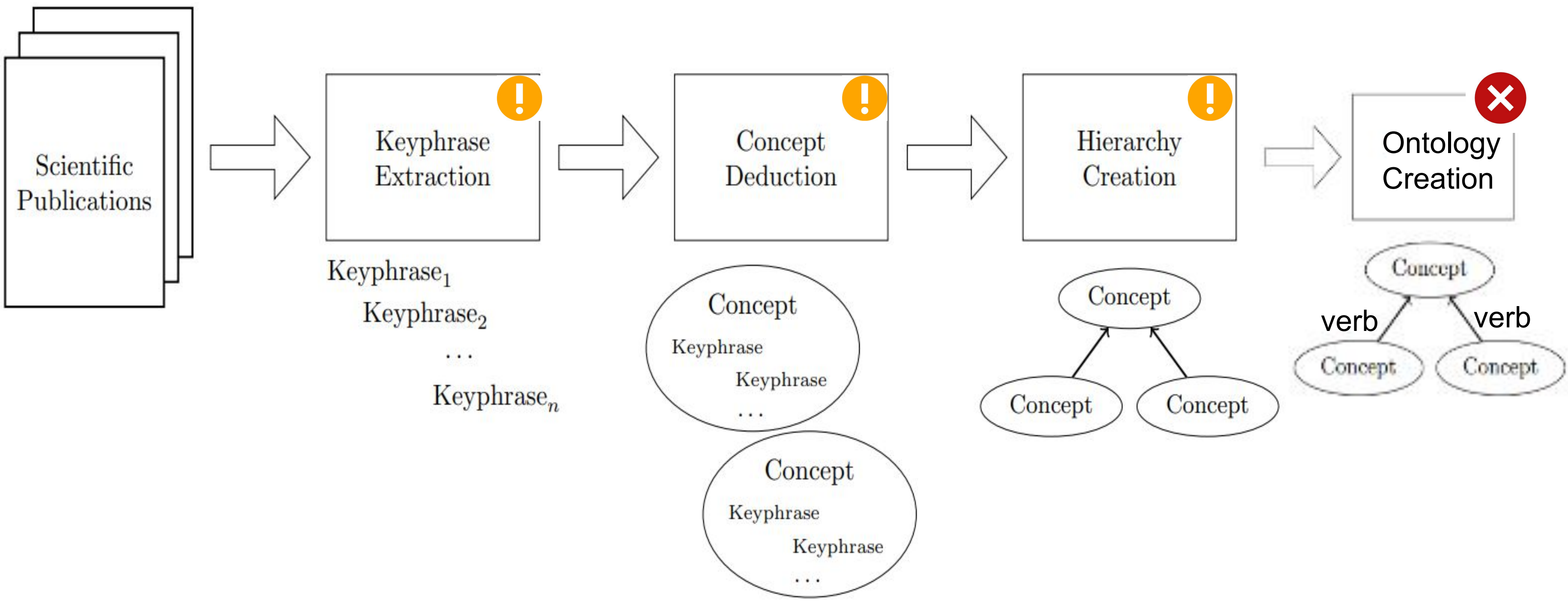


Figure 3: Pipeline of taxonomy creation steps in Simon Klimek's thesis.

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

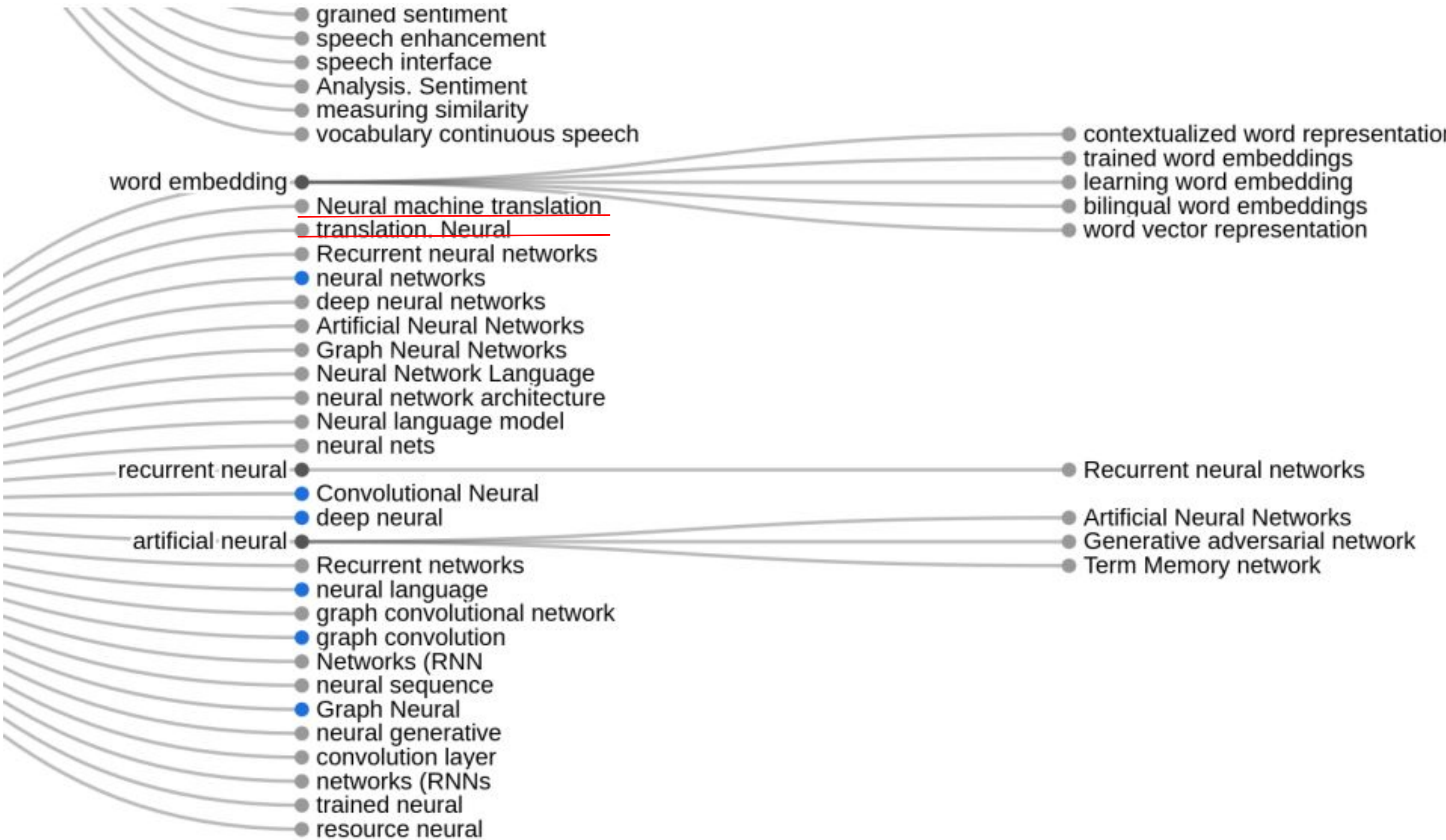


Figure 4: snippet of Klimke's generated NLP taxonomy

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?**

RQ1

- **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:

- The Association for Computational Linguistics conferences (ACL)
- NLP surveys and papers.
- The Computer Science Ontology (CSO)

Based on semi-structured qualitative interviews with domain researchers, we reach a final prototype of the taxonomy that satisfies the largest common denominator of the researchers' expectations.

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/

RQ1: Design Process of the manual ontology

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

Figure 5: evolution of final manual ontology layers.

Figure 6: snippet of final manual ontology layers.

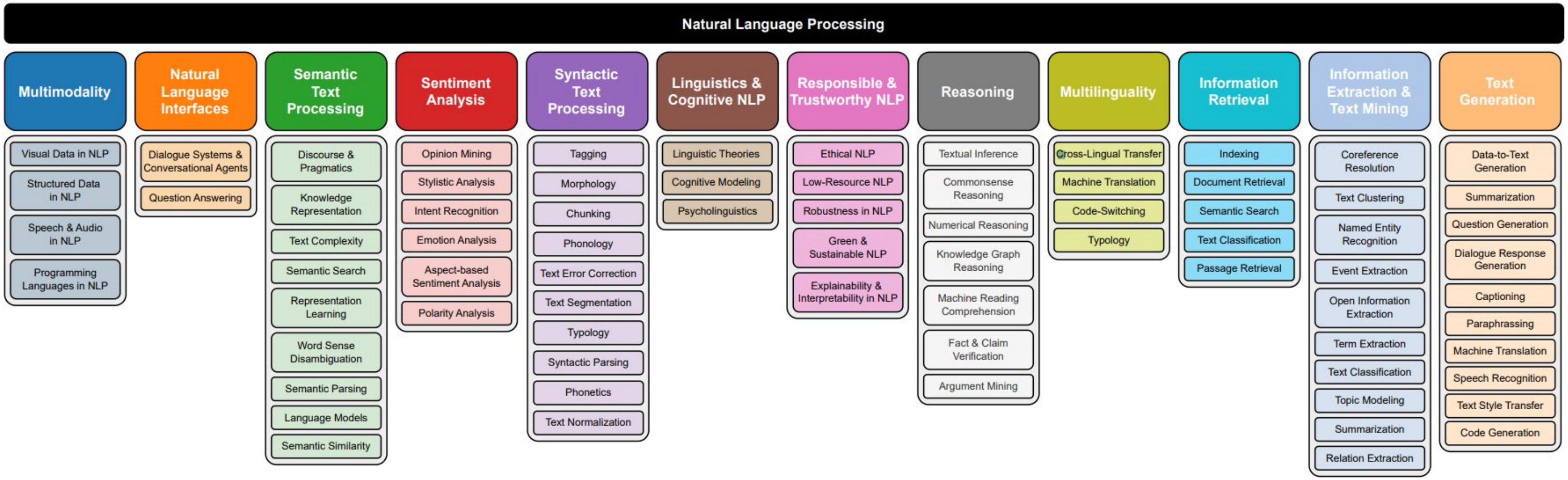


Figure 7: Top layer manually chosen concepts

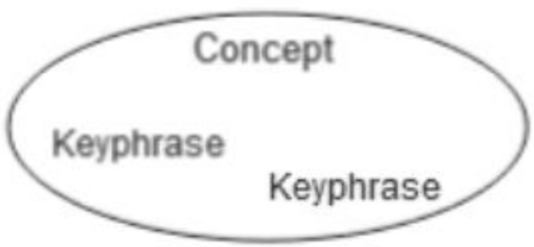
RQ2

- **Enhance concept and hierarchy inference**

Why: Weaknesses in current implementation can be improved.

How:

<u>Step 1</u>: Improve Concept Coherence



<u>Step 2</u>: Improve Taxonomic Relation Inference

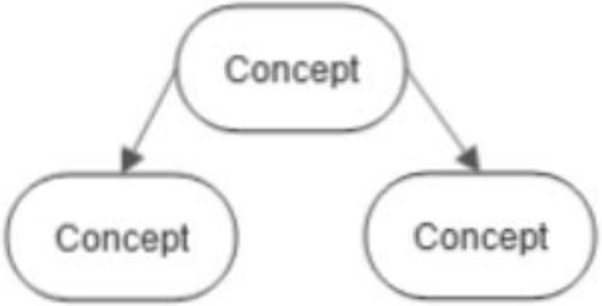


Figure 8: Concept Coherence and Hierarchy relation schematics

RQ2, Step 1: Improve Concept Coherence

Pre-processing:

- **Sanitize Extracted Keyphrases**

Existing Solution:

- **BERT-based Lexical Substitution:** Promising but flawed
  Improved with **BART-based Lexical Substitution**

New Solutions:

- **SciConceptMiner**
- **Sentence Transformers**

RQ2, Step 1: Improve Concept Coherence

Pre-processing:

- **Sanitize Extracted Keyphrases**

Existing Solution:

- **BERT-based Lexical Substitution:** Promising but flawed
  Improved with **BART-based Lexical Substitution**

New Solutions:

- **SciConceptMiner**
- **Sentence Transformers**

RQ2, Step 1: Pre-Sanitize extracted keyphrases

Pre-processing of keyphrases before merging methods:

- **First**: Trim keyphrases with **acronyms** and **punctuation marks**.
  e.g: 'machine learning (ml)?' >> 'machine learning'

- **Second:** Discard keyphrases with **punctuation marks** or that start with a **number** or contain only **one character**.
  e.g: 'language? Text' '3 step process' 'a'

- **Third**: Extend incomplete hyphenated keyphrases.
  e.g: 'automatically-' >> 'automatically-obtained'

- **Fourth**: Discard keyphrases with low **Information Content (IC)** score.
  e.g: 'science' 'languages'

D. Dessı, F. Osborne, D. R. Recupero, D. Buscaldi, and E. Motta. "SCICERO: A deep learning and NLP approach for generating scientific knowledge graphs in the computer science domain". In: Knowledge-Based Systems 258 (2022), p. 109945

RQ2, Step 1: Improve Concept Coherence

Pre-processing:

- **Sanitize Extracted Keyphrases**

Existing Solution:

- **BERT-based Lexical Substitution:** Promising but flawed
  Improved with **BART-based Lexical Substitution**

New Solutions:

- **SciConceptMiner**
- **Sentence Transformers**

RQ2, Step 1: BART-LS approach

Idea: 2 keyphrases have enough synonyms in common >> merged

BERT-LS shortcomings: only generates synonyms with same number of tokens!
'*token token token*' >> '*synonym synonym synonym*'

Alternative: BART-LS

*We explore different **Machine Translation** approaches.*
*[Machine Translation] We explore different **<mask>** approaches.*

BART is trained on noising all the input, it can predict and change parts of the sentence that go beyond just the masked portion. Therefore:

- Limit of up to five newly generated tokens.
- Discard newly generated keyphrases that fail the sanitation check.
- Discard generated outputs that made any changes to the input beyond just the token.

M. Pogoda, K. Gawron, N. Ropiak, M. Swędrowski, and J. Kocoń. "Deep Neural Sequence to Sequence Lexical Substitution for the Polish Language". In: Computational Science–ICCS 2022: 22nd International Conference, London, UK, June 21–23, 2022, Proceedings, Part I. Springer. 2022, pp. 692–705.

RQ2, Step 1: Improve Concept Coherence

Pre-processing:

- **Sanitize Extracted Keyphrases**

Existing Solution:

- **BERT-based Lexical Substitution:** Promising but flawed
  Improved with **BART-based Lexical Substitution**

New Solutions:

- **SciConceptMiner**
- **Sentence Transformers**

RQ2, Step 1: SciConceptMiner Approach

- **Idea: 2 keyphrases have enough common URLs >> merged**

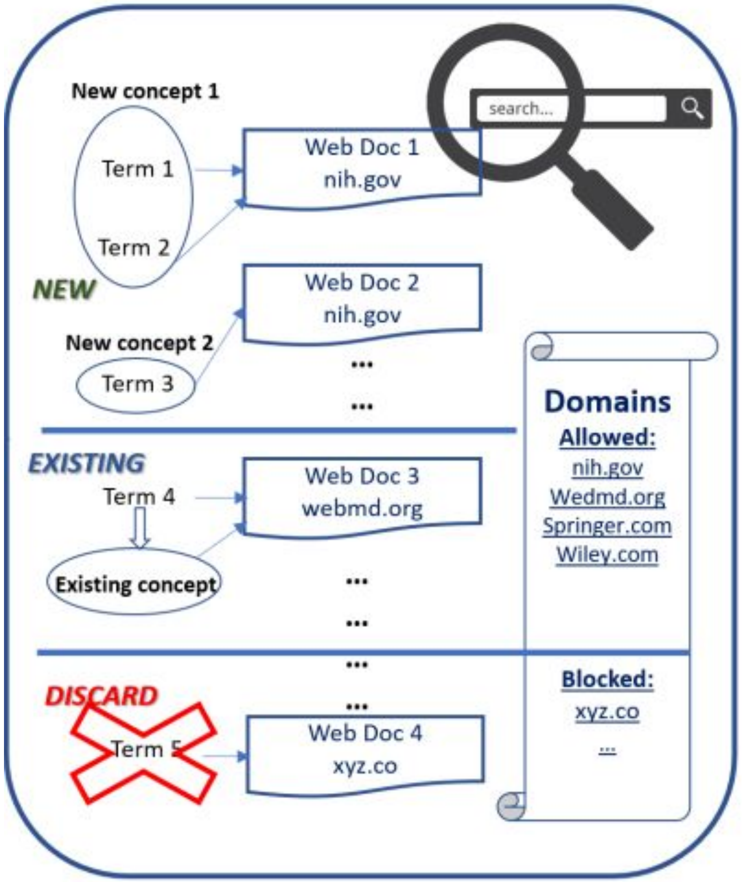


Figure 9: SciConcptMiner Approach

```
('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

RQ2, Step 1: Improve Concept Coherence

Pre-processing:

- **Sanitize Extracted Keyphrases**

Existing Solution:

- **BERT-based Lexical Substitution:** Promising but flawed
  Improved with **BART-based Lexical Substitution**

New Solutions:

- **SciConceptMiner**
- **Sentence Transformers**

RQ2, Step 1: Sentence Transformers Approach

Idea: All corpus papers are related to NLP. 2 keyphrases have 1 token in common and cosine similarity > 0.9 >> merged

E.g: [Emotion]: *Emotion Detection* and *Emotion Recognition*
Cosine Similarity > 0.9

[Detection]: *Emotion Detection* and *Sentiment Detection*
Cosine Similarity > 0.9

Merged: [*Emotion Detection, Emotion Recognition, Sentiment Detection*]

D. Dessı, F. Osborne, D. R. Recupero, D. Buscaldi, and E. Motta. "SCICERO: A deep learning and NLP approach for generating scientific knowledge graphs in the computer science domain". In: Knowledge-Based Systems 258 (2022), p. 109945

RQ2, Step 2: Improve Taxonomic Relation Inference

Existing Solutions:

- **Lexical Syntactic Method:** Underperforms, can be improved.
- **Subsumption Method**: Performs decently.

New Solutions:

- **String Inclusion**
- **Weighted Ensemble**

RQ2, Step 2: Improve Taxonomic Relation Inference

Existing Solutions:

- **Lexical Syntactic Method:** Underperforms, can be improved.
- **Subsumption Method**: Performs decently.

New Solutions:

- **String Inclusion**
- **Weighted Ensemble**

RQ2, Step 2: Lexical Syntactic Method

Idea: If the sentence structure containing Concepts 1 and 2 follow certain patterns, then there is a relation between them.

Existing Rules

1. such KEYPHRASE as (KEYPHRASE,)* (and|or) (KEYPHRASE ,)+
2. (KEYPHRASE,?)+ (and|or) other KEYPHRASE
3. KEYPHRASE, (especially|including) (KEYPHRASE,)+ (and|or) KEYPHRASE

...

Newly Added Rules

1. KEYPHRASE is (a|an) KEYPHRASE
2. KEYPHRASE is a (kind|type) of KEYPHRASE

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

RQ2, Step 2: Improve Taxonomic Relation Inference

Existing Solutions:

- **Lexical Syntactic Method:** Underperforms, can be improved.
- **Subsumption Method**: Performs decently.

New Solutions:

- **String Inclusion**
- **Weighted Ensemble**

RQ2, Step 2: Subsumption Method

Idea: If Concept 1 occurs very frequently in the same context as Concept 2, then it is a hyponym of Concept 2.

Subsumption Method (unchanged)

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

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

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

RQ2, Step 2: Improve Taxonomic Relation Inference

Existing Solutions:

- **Lexical Syntactic Method:** Underperforms, can be improved.
- **Subsumption Method**: Performs decently.

New Solutions:

- **String Inclusion**
- **Weighted Ensemble**

RQ2, Step 2: String Inclusion Approach

Idea: Each word from Concept 1 is similar to a word in Concept 2, and at least one word from Concept 1 is a hypernym of a word in Concept 2

String Inclusion

Notation	Meaning
$t_1 \gg t_2$	t_1 is a hypernym of t_2
$t_1 \approx t_2$	t_1 semantically equals or is similar to t_2
$t_1 \gg_{WN} t_2$	t_1 is a direct or inherited hypernym of t_2 according to WordNet
$t_1 \approx_{WN} t_2$	t_1 and t_2 belong to the same synset of WordNet

E.g: '*Suicide Attack*' and '*1983 self-destruction bombing*'.
"attack" ≫Wn "bombing" and "suicide" ≈Wn "self-destruction"

Therefore: '*Suicide Attack*' is the hypernym of '*1983 self-destruction bombing*'.

A. T. Luu, J.-j. Kim, and S. K. Ng. "Taxonomy construction using syntactic contextual evidence". In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014, pp. 810–819

RQ2, Step 2: Improve Taxonomic Relation Inference

Existing Solutions:

- **Lexical Syntactic Method:** Underperforms, can be improved.
- **Subsumption Method**: Performs decently.

New Solutions:

- **String Inclusion**
- **Weighted Ensemble**

RQ2, Step 2: Weighted Ensemble Approach

Idea: Place equal weights on the 3 previous approaches. If at least 2 out of 3 indicate a taxonomic relation, it is valid. Otherwise, it is discarded.

Lexical Syntactic Method

Subsumption Method

String Inclusion Method

RQ3

- **Add more complex non-taxonomic relations**

Why: Allows for deeper semantic topic exploration than parent-child (hypernym-hyponym relations)
How: Investigate new relation extraction methods.

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

Figure 10: 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,

RQ3: Infer Non-Taxonomic Relations.

New Solutions:

- **Dependency Tree Paths Based Approach**
- **PoS Tag-Based Relationship Extractor Approach**

Post-processing:

- **Verbal Relation Mapping**

RQ3: Infer Non-Taxonomic Relations.

New Solutions:

- **Dependency Tree Paths Based Approach**
- **PoS Tag-Based Relationship Extractor Approach**

Post-processing:

- **Verbal Relation Mapping**

RQ3: Dependency Tree Paths Based Approach

Idea: Leverage 'ideal' dependency trees to extract verbal relations between concept pairs that lie on the same path.

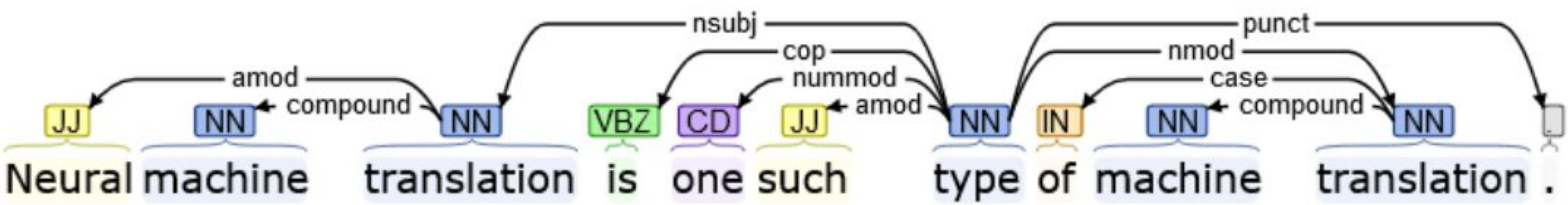


Figure 11: Example of Stanford CoreNLP Dependency Parser.

12 'good' dependency paths generated by Dessi with a correctness rate exceeding 60%.

1. **'nsubj', 'obj'**: The subject of the sentence is connected to the direct object through a verb.
2. **'acl:relcl', 'obj'**: An adjectival clause modifies the object of the main clause.
3. **'nsubj', 'obj', 'conj'**: The subject and the direct object are connected through coordination, indicating multiple subjects or objects in the sentence.

…

D. Dessı, F. Osborne, D. R. Recupero, D. Buscaldi, and E. Motta. "SCICERO: A deep learning and NLP approach for generating scientific knowledge graphs in the computer science domain". In: Knowledge-Based Systems 258 (2022), p. 109945

RQ3: Infer Non-Taxonomic Relations.

New Solutions:

- **Dependency Tree Paths Based Approach**
- **PoS Tag-Based Relationship Extractor Approach**

Post-processing:

- **Verbal Relation Mapping**

RQ3: PoS Tag-Based Relationship Extractor Approach

Idea: More generic. Extract all verbs between 2 concepts within at most 10 tokens of each other.

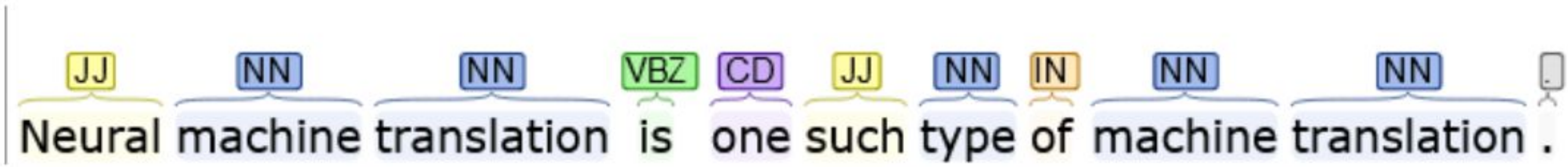


Figure 12: Example of Stanford CoreNLP Part-of-Speech.

D. Dessı, F. Osborne, D. R. Recupero, D. Buscaldi, and E. Motta. “SCICERO: A deep learning and NLP approach for generating scientific knowledge graphs in the computer science domain”. In: Knowledge-Based Systems 258 (2022), p. 109945

RQ3: Infer Non-Taxonomic Relations.

New Solutions:

- **Dependency Tree Paths Based Approach**
- **PoS Tag-Based Relationship Extractor Approach**

Post-processing:

- **Verbal Relation Mapping**

RQ3: Post-processing Verbal Relation Mapping

Idea: Too many triple variations are produced. Use a mapping to condense 464 types of verbal relations to one of 38 representative verbs, and discard the rest.

Mapped final verbs: uses, produces, provides, supports, proposes, base, improves, includes, identify, acquires, adapts, analyzes, links, matches, manages, interacts, queries, guides, automates, lacks, limits, affects, processes, contributes, causes, classifies, annotates, visualizes, predicts, standardizes, learns, executes, outperforms, extracts, highlights, transfers, solves, discusses.

D. Dessı, F. Osborne, D. R. Recupero, D. Buscaldi, and E. Motta. “SCICERO: A deep learning and NLP approach for generating scientific knowledge graphs in the computer science domain”. In: Knowledge-Based Systems 258 (2022), p. 109945

RQ1 Evaluation Method

Interviewee	Topic	# of Steps	Ideal Steps	Correct
#1	semantic parsing	2	2	1
	data-to-text generation	2	2	1
	conversational QA	3	3	1
	dialogue policy learning	5	5	1
	entity linking	12	6	1
	differential privacy	5	4	1
	chatGPT	4	4	1
#2	fact verification	6	2	1
	relation extraction	2	2	1
	sentiment analysis	1	1	1
	generative question answering	2	2	1
	knowledge graph embedding	7	7	1
	DeBERTa	4	4	1
	word edit distance	3	3	1
	green NLP	2	2	1

...

RQ1 Results

Approach	Relation Accuracy
Manual Taxonomy Creation	**0.988**
Subsumption Method, SCOPUS	0.860
TaxoGen, DBLP	0.775

$$\text{MAPE} = \frac{1}{n}\sum \frac{|\text{Total Steps Taken} - \text{Ideal \# Steps}|}{\text{Ideal \# Steps}} = \mathbf{0.478}$$

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

RQ2, Step 1: Evaluation Method

Concept Merging Coherence

Term 0	Term 1	Term 2	Term 3	Term 4	Term 5	Real Intruder Index (0 - 5)	Evaluator Intruder Index (0 - 5)
use convolutional neural	convolutional neural network	convolutional neural	text classification			3	3
processing	embeddings and weights	processing performed	processing works			1	1
nlp literature	word embeddings freely	word embeddings	word embedding	new word embedding		0	0
natural language questions	artificial neural	natural languages	natural language answer	natural language	natural language question	1	1
semantic similarity	automatically obtained synonyms	semantic similarities	semantic feature similarity			1	1
...							
Correct guesses (%)							**98.10%**

RQ2, Step 1: Results

Concept Merging Coherence

Approach	Concept Coherence
BART-LS and BERT-LS, SCOPUS	**0.816**
Sentence Transformers, SCOPUS	**0.981**
SciConceptMiner, SCOPUS	**0.988**
BERT-LS, SCOPUS	0.747
TaxoGen, DBLP	0.728

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

RQ2, Step 2: Evaluation Method

Taxonomic Relation Construction

Parent	Child	Interviewee #1	Interviewee #2	Interviewee #3	Interviewee #4	Interviewee #5	Majority
semantic	capturing crucial semantic	1	0	0	0	1	0
Term Memory	term memory network	1	1	1	1	1	1
language processing	natural language processing	1	1	1	1	1	1
neural network	convolutional neural network	1	1	1	1	1	1
semantic	semantic web reasoning	1	1	1	1	1	1
...							
Correctness (%)							**90.00%**

RQ2, Step 2: Results

Taxonomic Relation Construction

Approach	Relation Accuracy
String Inclusion, SCOPUS	**0.667**
Weighted Ensemble, SCOPUS	**0.900**
Lexical Syntactic Method, SCOPUS	0.440
Subsumption, SCOPUS	0.860
TaxoGen, DBLP	0.775

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

RQ3 Evaluation Method

Non-Taxonomic Relation Construction

Subject	Predicate/ Verb	Object	Interviewee #1	Interviewee #2	Interviewee #3	Interviewee #4	Interviewee #5	Majority
indexed journal articles	matches	keywords	0	1	1	1	0	1
GANCoder	produces	programming language codes	1	1	1	1	1	1
deep bidirectional transformers	uses	Improved prediction	0	0	0	1	1	0
lexicon matching	includes	character classifier	0	1	1	1	0	1
base WordNet	affects	replace rare words	0	0	0	0	0	0
...								
Correctness (%)								**53.33%**

RQ3 Results

Non-Taxonomic Relation Construction

Approach	Relation Accuracy
Dependency Tree Paths, SCOPUS	**0.533**
PoS Parsing, SCOPUS	**0.300**
Association Rules Algorithm, CS Corpus	0.728
Probabilistic Algorithm, CS Corpus	0.617

E. Drymonas, K. Zervanou, and E. G. Petrakis. “Unsupervised ontology acquisition from plain texts: the OntoGain system”. In: Natural Language Processing and Information Systems: 15th International Conference on Applications of Natural Language to Information Systems, NLDB 2010, Cardiff, UK, June 23-25, 2010. Proceedings 15. Springer. 2010, pp. 277– 287.

Conclusion

- **Successful manual taxonomy construction**

  Manual relation extraction **98.8%**

  Automated relation extraction **86~90%**

- **Concept coherence achieved better performance**

Sentence Transformers:	**98.1%**	>	**74.7%**
SciConceptMiner:	**98.9%**	>	**74.7%**
BERT-LS + BART-LS:	**81.7%**	>	**74.7%**

- **Hierarchy construction achieved better performance**

Weighted Ensemble:	**90%**	>	**86%**

- **Non-taxonomic relation extraction achieved middling results:**

Dependency Tree Paths:	**53.3%**	<	**73%**
PoS Parsing:	**30%**	<	**73%**

Future Work

- **Alternative datasets (Less domain-focused, more pure NLP).**
- **New RQ1 Evaluation Participants (avoid bias).**
- **Additional non-taxonomic relation extraction methods.**
- **Triple validation step.**

Related Publication

Exploring the Landscape of Natural Language Processing Research

Tim Schopf and **Karim Arabi** and **Florian Matthes**
Technical University of Munich, Department of Computer Science, Germany
{tim.schopf,karim.arabi,matthes}@tum.de

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