T7: Tech-Taxonomy with a Text To Text Transfer Transformer

You Zuo (Almanach, Inria), Kim Gerdes (Lisn, Upsay, ISS, Inria)

ALMAnaCH project-team Seminar - 22/10/2021

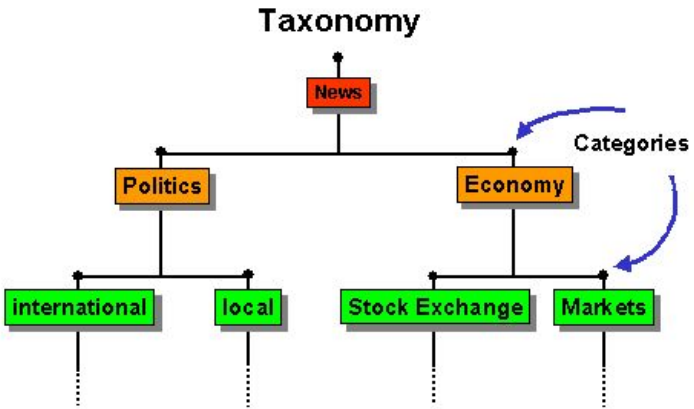
Outline

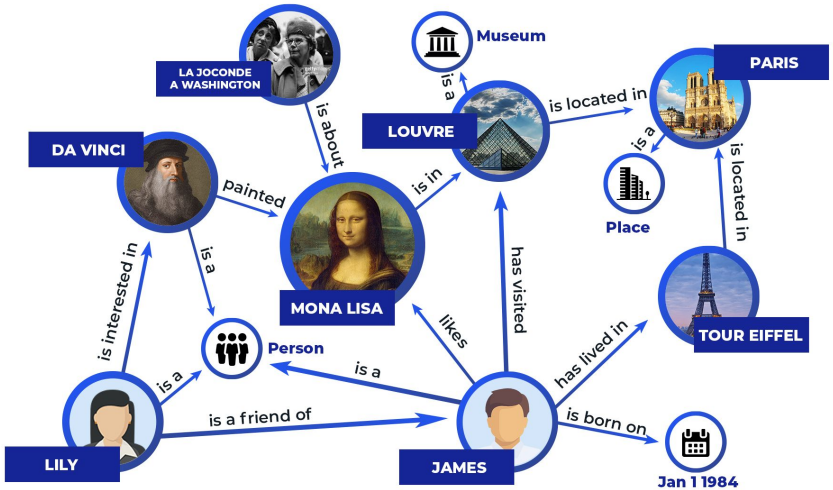
1. INTRODUCTION
   - what is tech-taxonomy?
   - what does qatent.com do?
   - why do we need tech-taxonomy?

2. METHODS
3. EVALUATION
4. CONCLUSION
5. Q & A

Introduction

- Definition of taxonomy
- Taxonomy vs. ontology/ knowledge graph

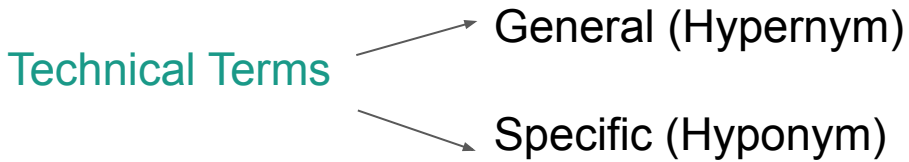




Introduction

- qatent
  - Web application for drafting patents
  - Patent description generator
  - Patent application validator

- Why do we need tech-taxonomy?
  - patent

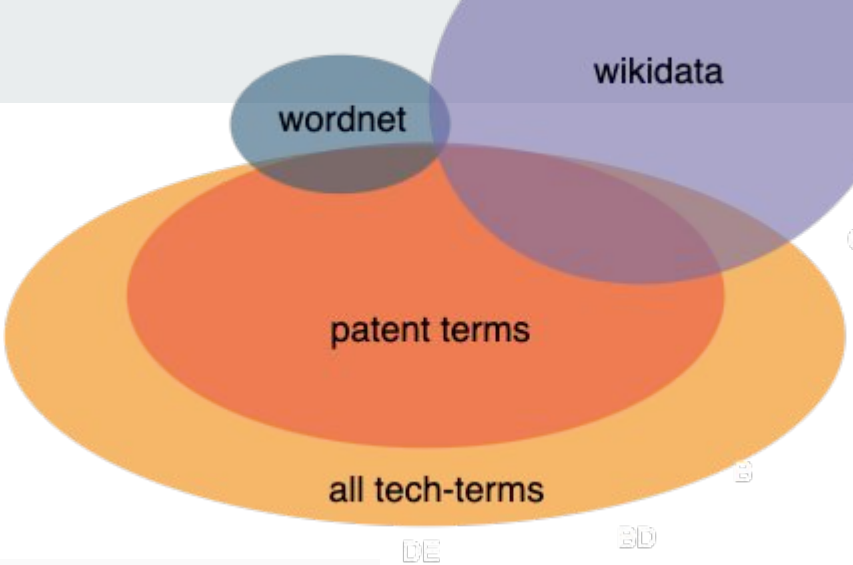


Problems

- Currently existing taxonomies / ontologies

  WordNet ⇒ too general !

  Wikidata ⇒ unrelated items: celebrities, locations, events
  or no information about hypernym



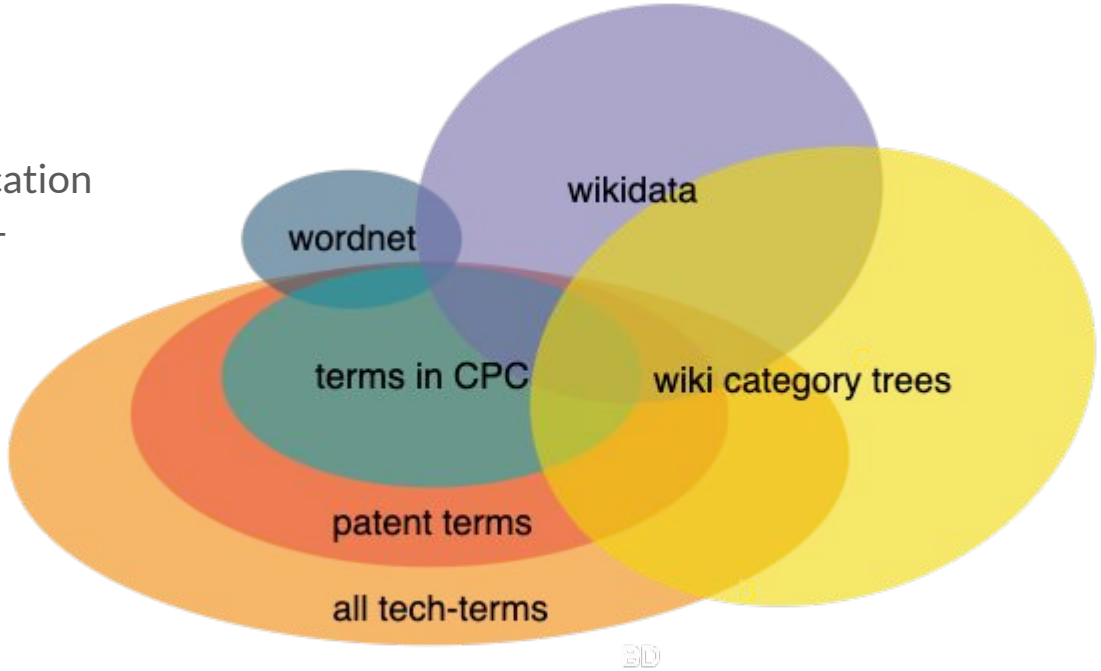
Problem

- Currently existing taxonomies / ontologies
  `WordNet, Wikidata`
- Patent domains

A	HUMAN NECESSITIES
B	PERFORMING OPERATIONS; TRANSPORTING
C	CHEMISTRY; METALLURGY
D	TEXTILES; PAPER
E	FIXED CONSTRUCTIONS
F	MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING
G	PHYSICS
H	ELECTRICITY

Solutions

- Cooperative Patent Classification (CPC) classification system + Wikipedia Category trees



Totally 262K categories in CPC!!

Solutions

CPC classification system

Hierarchy

- Section (one letter A to H and also Y)
  - Class (two digits)
    - Subclass (one letter)
      - Group (one to three digits)
        - subgroup (at least two digits)

Totally 262K categories in CPC!!

Solutions

CPC classification system

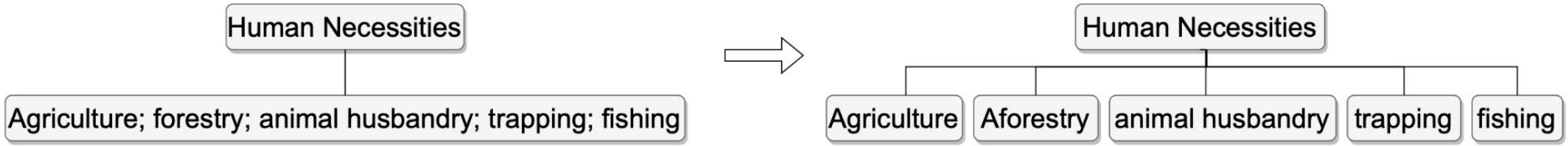
example "A01B33/02"

- **Section A** => Human Necessities
  - **Class 01** => Agriculture; forestry; animal husbandry; trapping; fishing
    - **Subclass B** => Soil working in agriculture or forestry, parts, details, or accessories of agricultural machines or implements, in general (making or covering furrows or holes for sowing, planting, or manuring A01C5/00; soil working for engineering purposes E01, E02, E21; measuring areas for agricultural purposes G01B)
      - **Group 33** => tilling implements with rotary driven tools (e.g. in combination with fertiliser distributors or seeders, with grubbing chains, with sloping axles, with driven discs)
        - **Subgroup 02** => ... with rigid tools

Solutions

example "A01B33/02"

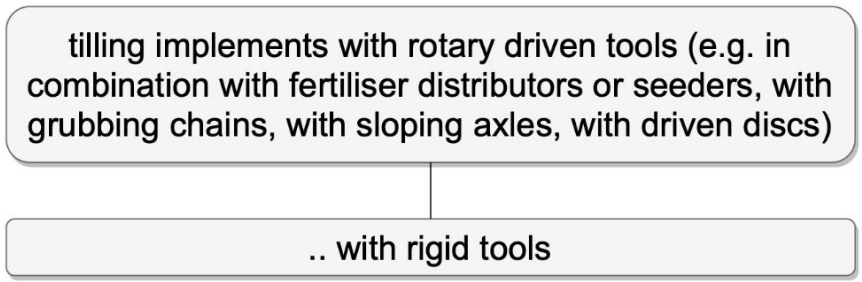
- **Section A** => Human Necessities
  - **Class 01** => Agriculture; forestry; animal husbandry; trapping; fishing

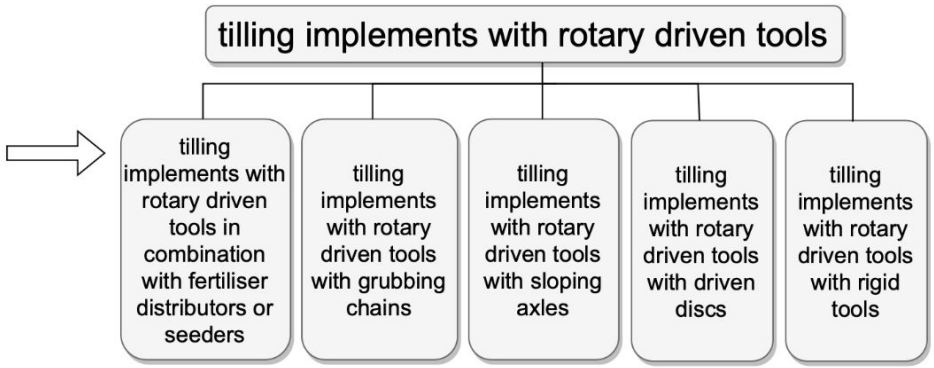


Solutions

example "A01B33/02"

- **Group 33** => tilling implements with rotary driven tools (e.g. in combination with fertiliser distributors or seeders, with grubbing chains, with sloping axles, with driven discs)
  - **Subgroup 02** => ... with rigid tools





308K nodes retrieved finally !

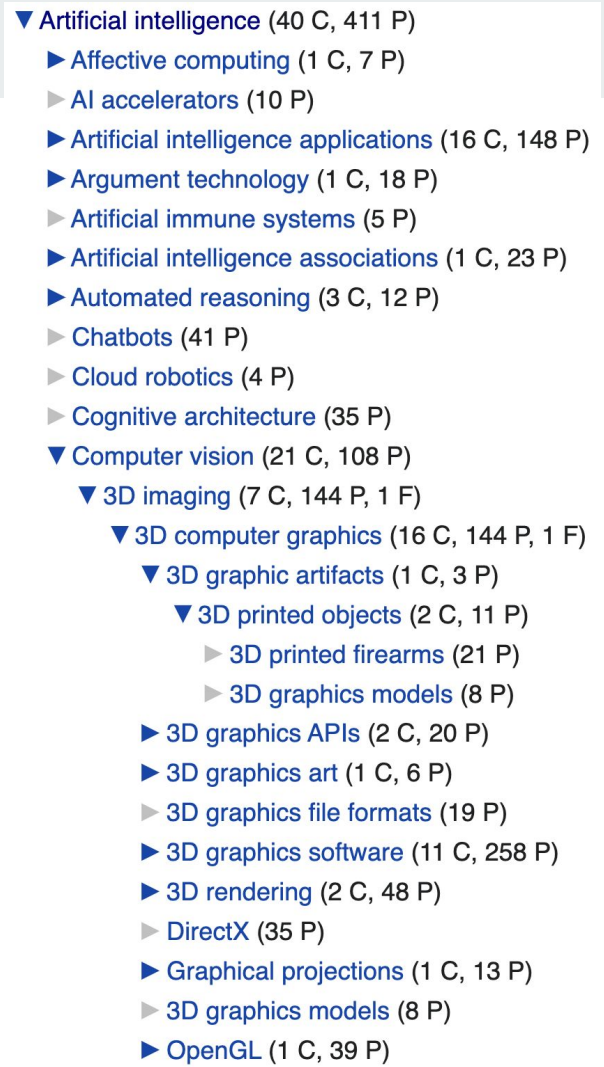
Parse analysis by Alma Parias García

Solutions

- Wikipedia Category trees

Extracted Wikipedia Category Trees under eight domains:

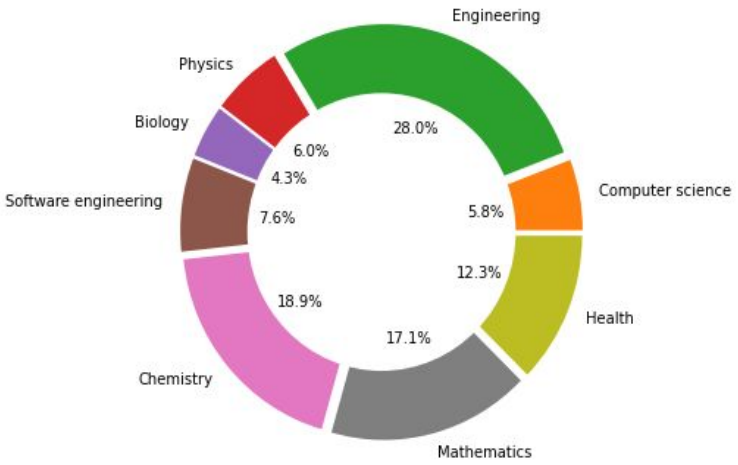
- Physics
- Biology
- Mathematics
- Engineering
- Computer Science
- Health
- Chemistry
- Software Engineering



Solutions

Category	Number of terms	Number of cleaned terms
Physics	57752	1032
Health	51137	2121
Software Engineering	47262	1307
Computer Science	21098	991
Chemistry	33854	3246
Mathematics	65019	2935
Biology	76470	737
Engineering	46932	4816
Total	399521	17185

Number of terms for each Wikipedia Category Tree



Distribution by topic of Wikipedia Category Trees

Solutions

- Cooperative Patent Classification (CPC) classification system
  + Wikipedia Category trees
  ⇒ **339k** nodes in our tech-taxonomy

Evaluation

- Human evaluation

- models for hypernym prediction
  - CRIM *(Bernier-Colborne & Barrière, 2018)*
  - TransE *(Bordes et al., 2013)*
  - hypo2hyper *(Cho et al., 2020)*
  - T5 *(Raffel et al., 2020)*

Models for hypernym prediction

CRIM *(Bernier-Colborne & Barrière, 2018)*

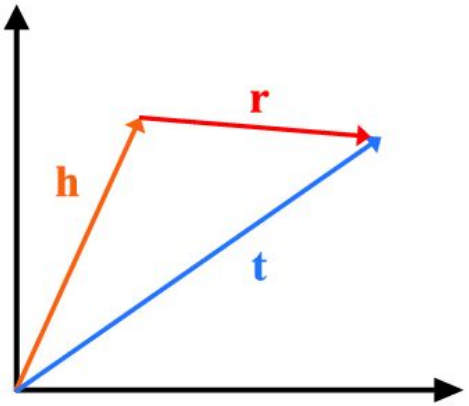
Ranked 1st in SemEval 2018 Task 9

- Learning Projection matrices from query embedding ⇒ target hypernym embeddings
- fastText as term representation (OOV, rare/ unfamiliar words)

Link prediction model

TransE *(Bordes et al., 2013)*

- Relationships are represented as translations in the embedding space
- if (h, r, t) holds, then `h+r≅t`



Hypernym prediction as sequence to sequence task

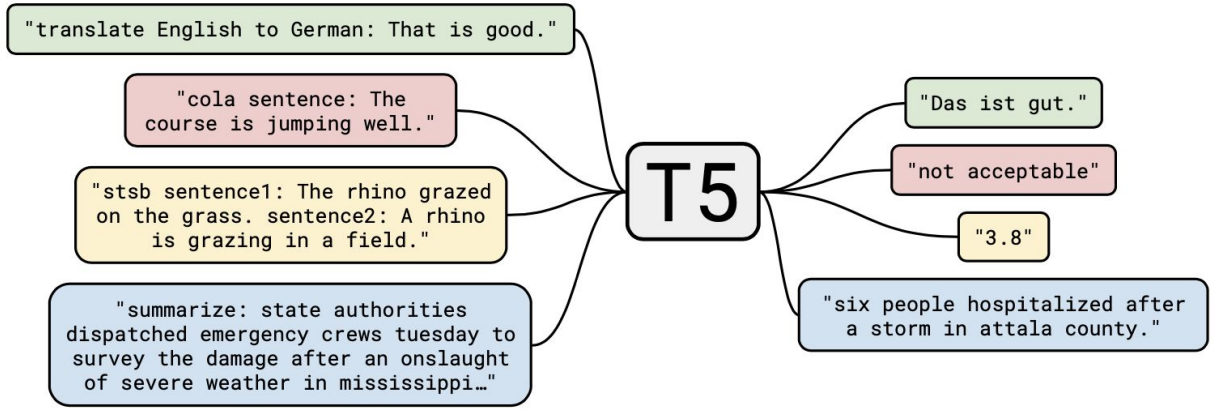
hypo2hyper *(Cho et al., 2020)*

- LSTM-based seq2seq model with Luong attention *(Luong et al., 2015)*

	Hyponym	Generated hypernym path	Gold hypernym
✓	pizza.n.01	dish.n.02 → nutriment.n.01 → food.n.01 → . . . → entity.n.01	dish.n.02
✓	alps.n.01	range.n.04 → geological_formation.n.01 → . . . → entity.n.01	range.n.04
✓	whisper.v.01	talk.v.02 → communicate.v.02 → interact.v.01 → act.v.01	talk.v.02
✗	proletarian.n.01	*worker.n.01 → person.n.01 → causal_agent.n.01 → . . . → entity.n.01	commoner.n.01
✗	austerity.n.01	*punishment.n.01 → social_control.n.01 → . . . → entity.n.01	self-discipline.n.01
✗	compulsive.n.01	*sick_person.n.01 → unfortunate.n.01 → person.n.01 → . . . → entity.n.01	person.n.01

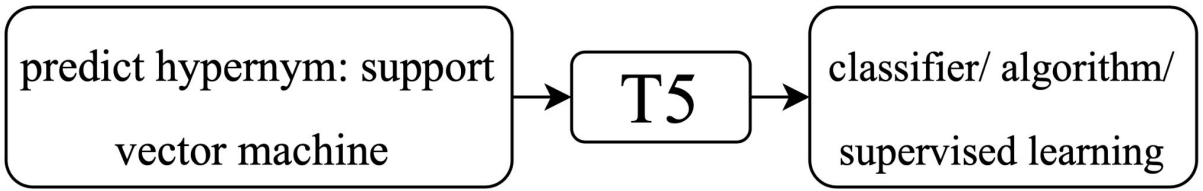
Table 1: We frame hypernym prediction as a sequence generation problem. Given a query hyponym (e.g., *pizza.n.01*), the ***hypo2path rev*** model generates its taxonomy path, from its direct hypernym (*dish.n.02*) to the root node (*entity.n.01*). ✓ and ✗ indicate a correct and an incorrect prediction, respectively. In each example, an underlined synset corresponds to what the model predicted as a direct hypernym.

Evaluation



T5 *(Raffel et al., 2020)*

- Text to Text Transfer Transformer



Evaluation

Metrics

- Hits@k score
  percentage of correct true labels that appeared in the top k (k=1, 3, 10) ranked predictions

- Mean reciprocal rank (MRR)
  the position of the first correct results in ranked list of outcomes

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

Results

Table 12. Results on test set of Qatent Taxonomy (incomplete)

MODEL	H@1	H@3	H@10	MRR
CLOSET VECTOR	9.49	17.61	28.26	15.57
CRIM	10.99	20.48	31.57	17.68
TRANSE	37.82	55.25	69.64	48.39
HYPO2HYPER	11.04	21.03	30.18	17.45
T5	**61.00**	**68.35**	**79.10**	**70.43**

Results

- many potential terms that have not been included in our taxonomy

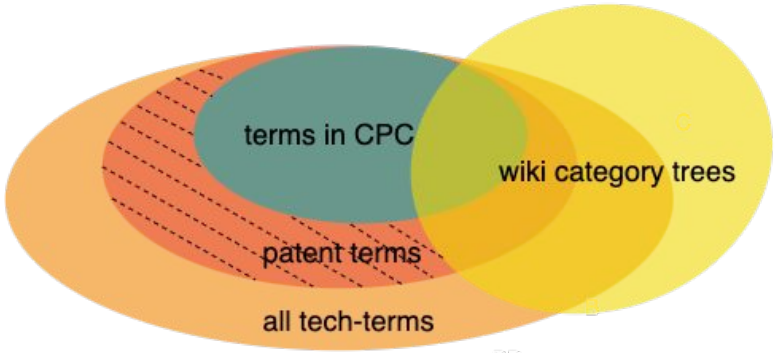
  ⇒ Term recognizer used in qatent:

  8. The method of 1, further comprising: in response to displaying the scrollable refresh trigger, providing first audio feedback; and in response to determining that the scrollable

- 45 % of patent terms are not in our taxonomy

Results

- many potential terms that have not been included in our taxonomy



- 45 % of patent terms are not in our taxonomy
  ⇒ T5

Ongoing and Future Work

1. Semi-supervised method to correct taxonomy extracted from CPC
   Using model trained on wikipedia category trees as a filter ⇒
   eg. keep terms if in the top 10

2. Try different pre-trained transformers: GPT-NeoX
3. Expand patent terms by implementing Hearst patterns in patent and technical wikipedia pages:
   `Y such as X, Y other than X, not all Y are X, Y including X, Y especially X, Y like X, Y for example X, Y which includes X, X are also Y, X are all Y`

Conclusion

- Taxonomies are not dead, still useful for paraphrase generation, intelligent editors
- Non-trivial task to develop domain-specific taxonomies

- Our attempt: Term recognizer (span Categorizer + NER from spaCy) + T5 ⇒ taxonomy completion

Q & A

Thank you for your attention!

Reference

- Bernier-Colborne, G. and Barrière, C. CRIM at SemEval-2018 task 9: A hybrid approach to hypernym discovery. In Proceedings of The 12th International Workshop on Semantic Evaluation, pp. 725–731, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1116.

- Bordes, A., Usunier, N., Garcia-Duran, A., Weston,J., and Yakhnenko, O.Translating embeddings for modeling multi-relational data.In Burges, C.J. C., Bottou, L., Welling, M., Ghahramani, Z.,and Weinberger, K. Q. (eds.),Advances in Neural Information Processing Systems, volume 26. Curran Associates, Inc., 2013.

- Camacho-Collados, J., Delli Bovi, C., Espinosa-Anke, L.,Oramas, S., Pasini, T., Santus, E., Shwartz, V., Navigli, R., and Saggion, H. SemEval-2018 task 9: Hypernym discovery. In Proceedings of The 12th International Workshop on Semantic Evaluation, pp. 712–724, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1115.

- Cho, Y., Rodriguez, J. D., Gao, Y., and Erk, K. Lever-aging WordNet paths for neural hypernym prediction.In Proceedings of the 28th International Conference on Computational Linguistics, pp. 3007–3018, Barcelona,Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.268.

Reference

- Martin Kracker. European patent information and the cpc taxonomy as linkedopen data. InSEMANTICS Posters&Demos, 2018.
- Tung Tran and Ramakanth Kavuluru. Supervised approaches to assign cooperative patent classification (cpc) codes to patents. In MIKE, 2017.
- George A Miller. Wordnet: a lexical database for english.Communications of theACM, 38(11):39–41, 1995.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S.,Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.