

SWODCH 2022

Semantic Web and Ontology Design for Cultural Heritage

Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage

Erica Faggiani, Stefano Faralli and Paola Velardi

Sapienza University of Rome, Italy



SAPIENZA
UNIVERSITÀ DI ROMA



Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage



Since December 1st 2021

I'm an Assistant Professor at Computer Science Department Sapienza

Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage

Erica Faggiani, Stefano Faralli and Paola Velardi

Sapienza University of Rome, Italy



SAPIENZA
UNIVERSITÀ DI ROMA



SAPIENZA
UNIVERSITÀ DI ROMA

Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage

SMARTOUR: intelligent platforms for tourism, funded by the Italian Ministry of University and Research.



SMARTOUR



<https://smartour.net>





SAPIENZA
UNIVERSITÀ DI ROMA

Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage



Knowledge Engineering: The Applied Side of Artificial Intelligence

Edward A. Feigenbaum
Computer Science Department
Stanford University
Stanford, CA USA 94305

1.0 Introduction: Symbolic Computation and Inference

This paper will discuss the applied artificial intelligence work that is sometimes called "knowledge engineering". The work is based on computer programs that do symbolic manipulations and symbolic inference, not calculation. The programs I will discuss do essentially no numerical calculation. They discover qualitative lines-of-reasoning leading to solutions to problems stated symbolically.

1.1 Knowledge

Since in this paper I often use the term "knowledge", let me say what I mean by it. The knowledge of an area of expertise—of a field of practice—is generally of two types: a) Facts of the domain—the widely shared knowledge that is written in textbooks, and in journals of a field; that constitutes the kind of material that a professor would lecture about in a class. b) Equally as important to the practice of a field is the heuristic knowledge—knowledge which constitutes the rules of expertise, the rules of good practice, the judgmental rules of the field, the rules of plausible reasoning. These rules collectively constitute what the mathematician, George Polya, has called the "art of good guessing". In contrast to the facts of the field, its rules of expertise, its rules of good guessing, are rarely written down. This knowledge is transmitted in internships, Ph.D. programs, apprenticeships. The programs I will describe require, for expert performance on problems, heuristic knowledge to be combined with the facts of the discipline.

Feigenbaum E. A. Knowledge engineering. The applied side of artificial intelligence. Ann N Y Acad Sci. 1984;426:91-107.
doi: 10.1111/j.1749-6632.1984.tb16513.x.
PMID: 6391332.

SWODCH 2022

Semantic Web and Ontology Design for Cultural Heritage



Adoption of ArCo

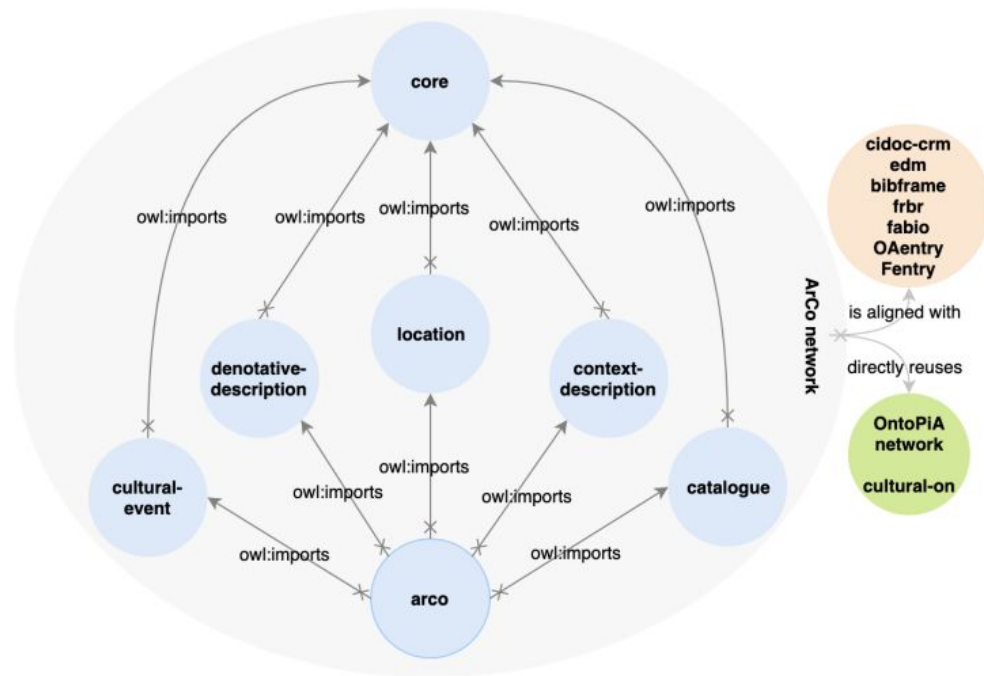
To develop advanced applications in the Italian CH domain and to meet the Smartour project's requirements we decided to adopt ArCo [1].

The project ArCo started in 2017 with the aim of providing an optimal LOD publication for many existing Italian CH catalogs.

[1] Carriero, V.A., Gangemi, A., Mancinelli, M.L., Marinucci, L., Nuzzolese, A.G., Presutti, V., Veninata, C.: Arco: The italian cultural heritage knowledge graph. In: The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II. Lecture Notes in Computer Science, vol. 11779, pp. 36–52. Springer (2019).
https://doi.org/10.1007/978-3-030-30796-7_3



The ArCo Knowledge graph



[1] Carriero, V.A., Gangemi, A., Mancinelli, M.L., Marinucci, L., Nuzzolese, A.G., Presutti, V., Veninata, C.: Arco: The Italian cultural heritage knowledge graph. In: The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II. Lecture Notes in Computer Science, vol. 11779, pp. 36–52. Springer (2019). https://doi.org/10.1007/978-3-030-30796-7_3



ArCo entity example excerpt

property	value
uri	https://w3id.org/arco/resource/HistoricOrArtisticProperty/1200252386
type	"dipinto"
subject	"paesaggio (dipinto) by Salvator Rosa (scuola) (sec. XVII, seconda metà)"
date	"1650-1699"
latitude	41.90745
longitude	12.498603
city	https://w3id.org/arco/resource/City/roma
city label	"ROMA"
author	https://w3id.org/arco/resource/Agent/aea151c6ed45d80e78eb79b4ec150aca
author label	"Salvator Rosa, Scuola"
author date	"1615/ 1673"

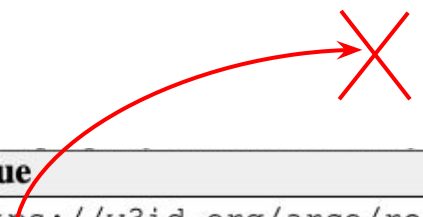


ArCo entity example excerpt

property	value
uri	https://w3id.org/arco/resource/HistoricOrArtisticProperty/1200252386
type	"dipinto"
subject	"paesaggio (dipinto) by Salvator Rosa (scuola) (sec. XVII, seconda metà)"
date	"1650-1699"
latitude	41.90745
longitude	12.498603
city	https://w3id.org/arco/resource/City/roma
city label	"ROMA"
author	https://w3id.org/arco/resource/Agent/aea151c6ed45d80e78eb79b4ec150aca
author label	"Salvator Rosa, Scuola"
author date	"1615/ 1673"



Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage



property	value
uri	https://w3id.org/arco/resource/HistoricOrArtisticProperty/1200252386
type	"dipinto"
subject	"paesaggio (dipinto) by Salvator Rosa (scuola) (sec. XVII, seconda metà)"
date	"1650-1699"
latitude	41.90745
longitude	12.498603
city	https://w3id.org/arco/resource/City/roma
city label	"ROMA"
author	https://w3id.org/arco/resource/Agent/aea151c6ed45d80e78eb79b4ec150aca
author label	"Salvator Rosa, Scuola"
author date	"1615/ 1673"



ArCo entity example excerpt

property	value
uri	https://w3id.org/arco/resource/HistoricOrArtisticProperty/1200252386
type	"dipinto"
subject	"paesaggio (dipinto) by Salvator Rosa (scuola) (sec. XVII, seconda metà)"
date	"1650-1699"
latitude	41.90745
longitude	12.498603
city	https://w3id.org/arco/resource/City/roma
city label	"ROMA"
author	https://w3id.org/arco/resource/Agent/aea151c6ed45d80e78eb79b4ec150aca
author label	"Salvator Rosa, Scuola"
author date	"1615/ 1673"



we identified the needs for:

- normalizing human curated (hence error-prone) literal properties to let both humans and machines better understand CHs' attributes;
- enriching the instances' semantics given by the ArCo ontology, by means of performing entity linking on textual descriptions by targeting semantic-rich knowledge graphs, i.e., GVP (AAT, TGN and ULAN) [2] and DBpedia [3].

[2] Harpring, P.: Development of the getty vocabularies: Aat, tgn, ulan, and cona. Art Documentation: Journal of the Art Libraries Society of North America 29 (1), 67–72 (2010), <http://www.jstor.org/stable/27949541>

[3] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: The Semantic Web. pp. 722–735. Springer Berlin Heidelberg, Berlin, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_52



We started the ArCo, GVP and DBpedia Linking Initiative (AGDLI).

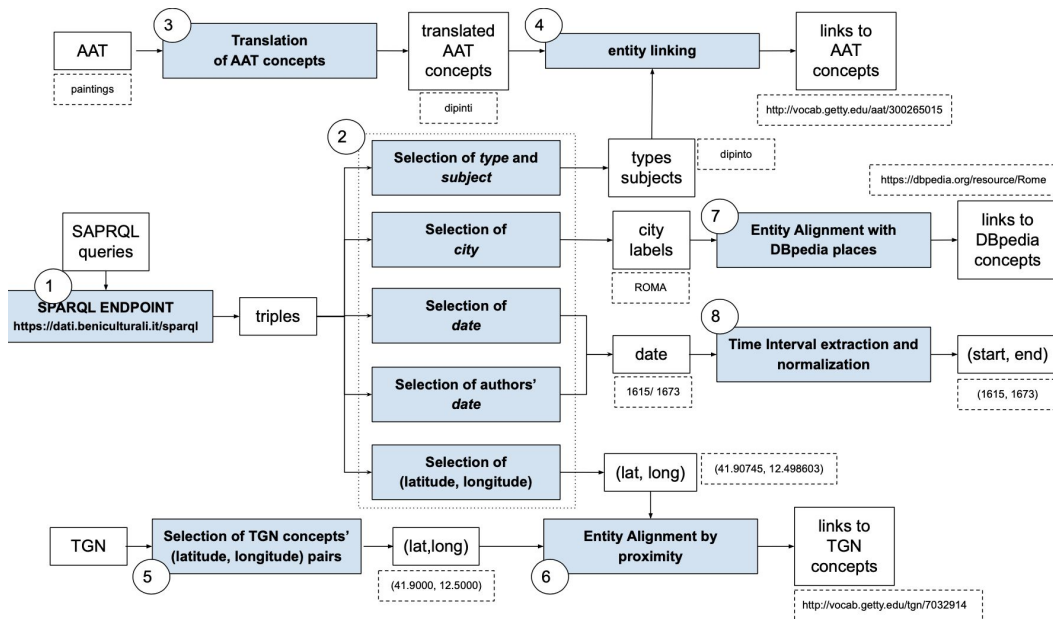


source code, datasets and updates are available at this link:

<https://sites.google.com/uniroma1.it/agdli/>



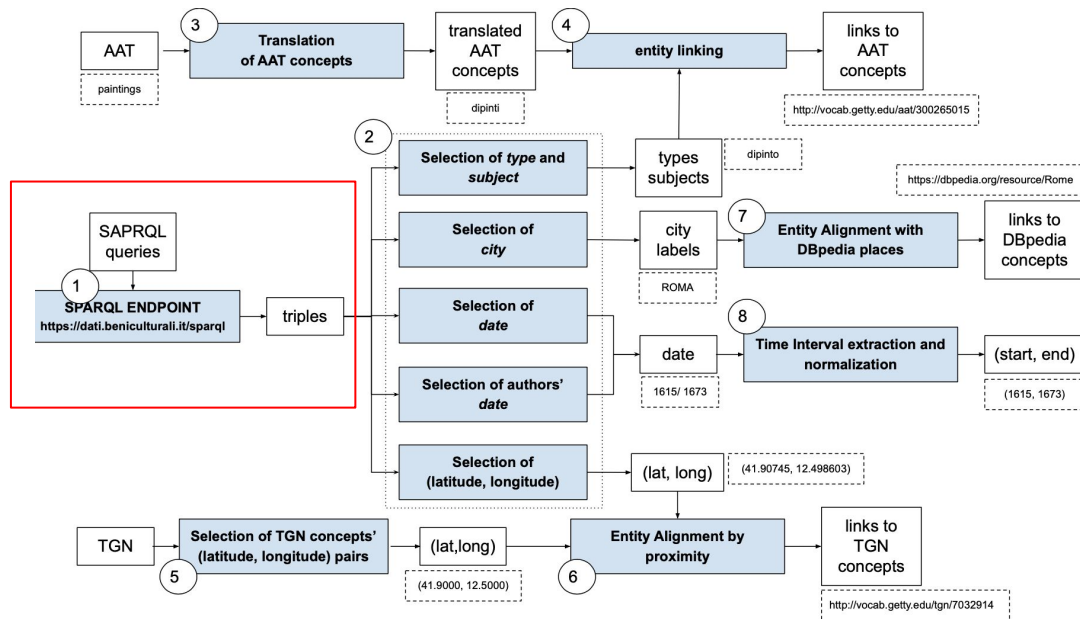
Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage



- Stefano Faralli, Andrea Lenzi, Paola Velardi: AGDLI: ArCo, GVP and DBpedia Linking Initiative. ISWC 2021.
- Stefano Faralli, Andrea Lenzi, Paola Velardi: A Large Interlinked Knowledge Graph of the Italian Cultural Heritage. LREC 2022

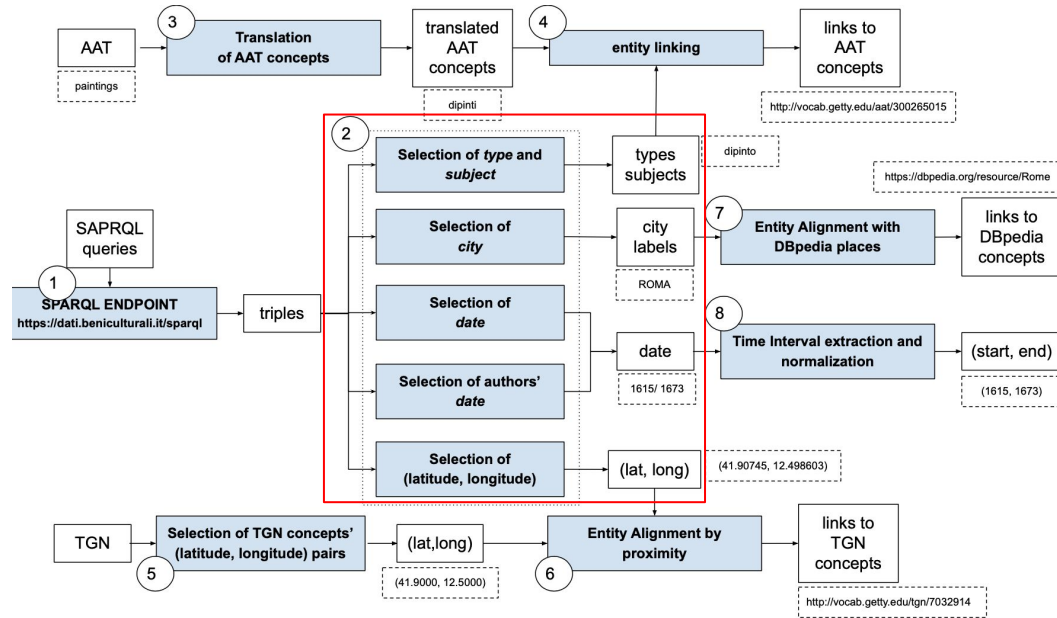


Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage



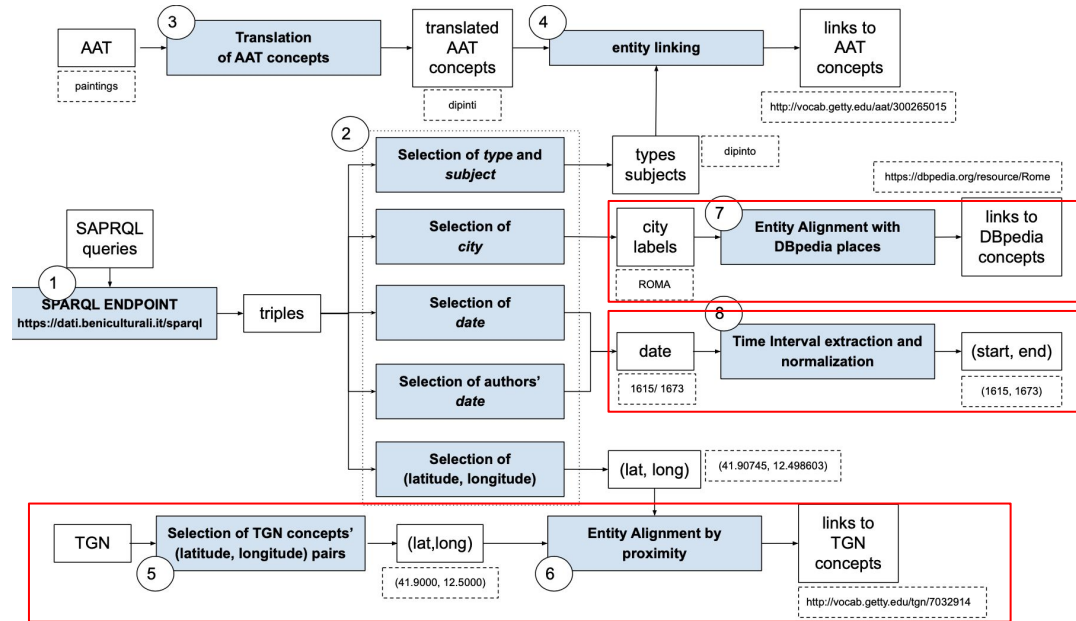
- Stefano Faralli, Andrea Lenzi, Paola Velardi: AGDLI: ArCo, GVP and DBpedia Linking Initiative. ISWC 2021.
- Stefano Faralli, Andrea Lenzi, Paola Velardi: A Large Interlinked Knowledge Graph of the Italian Cultural Heritage. LREC 2022

Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage



- Stefano Faralli, Andrea Lenzi, Paola Velardi: AGDLI: ArCo, GVP and DBpedia Linking Initiative. ISWC 2021.
- Stefano Faralli, Andrea Lenzi, Paola Velardi: A Large Interlinked Knowledge Graph of the Italian Cultural Heritage. LREC 2022

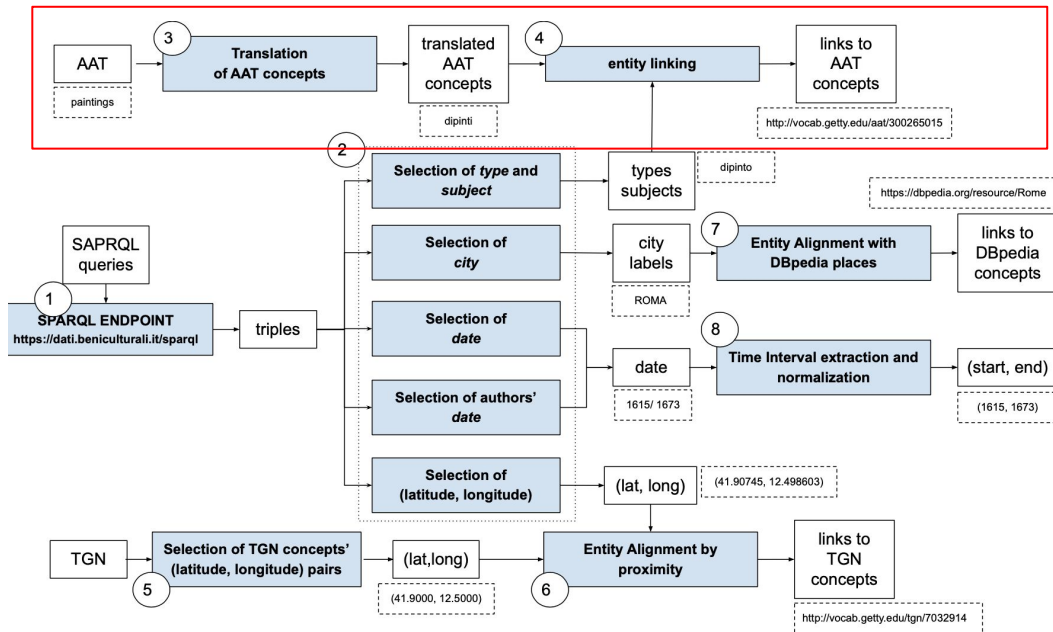
Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage



- Stefano Faralli, Andrea Lenzi, Paola Velardi: AGDLI: ArCo, GVP and DBpedia Linking Initiative. ISWC 2021.
- Stefano Faralli, Andrea Lenzi, Paola Velardi: A Large Interlinked Knowledge Graph of the Italian Cultural Heritage. LREC 2022



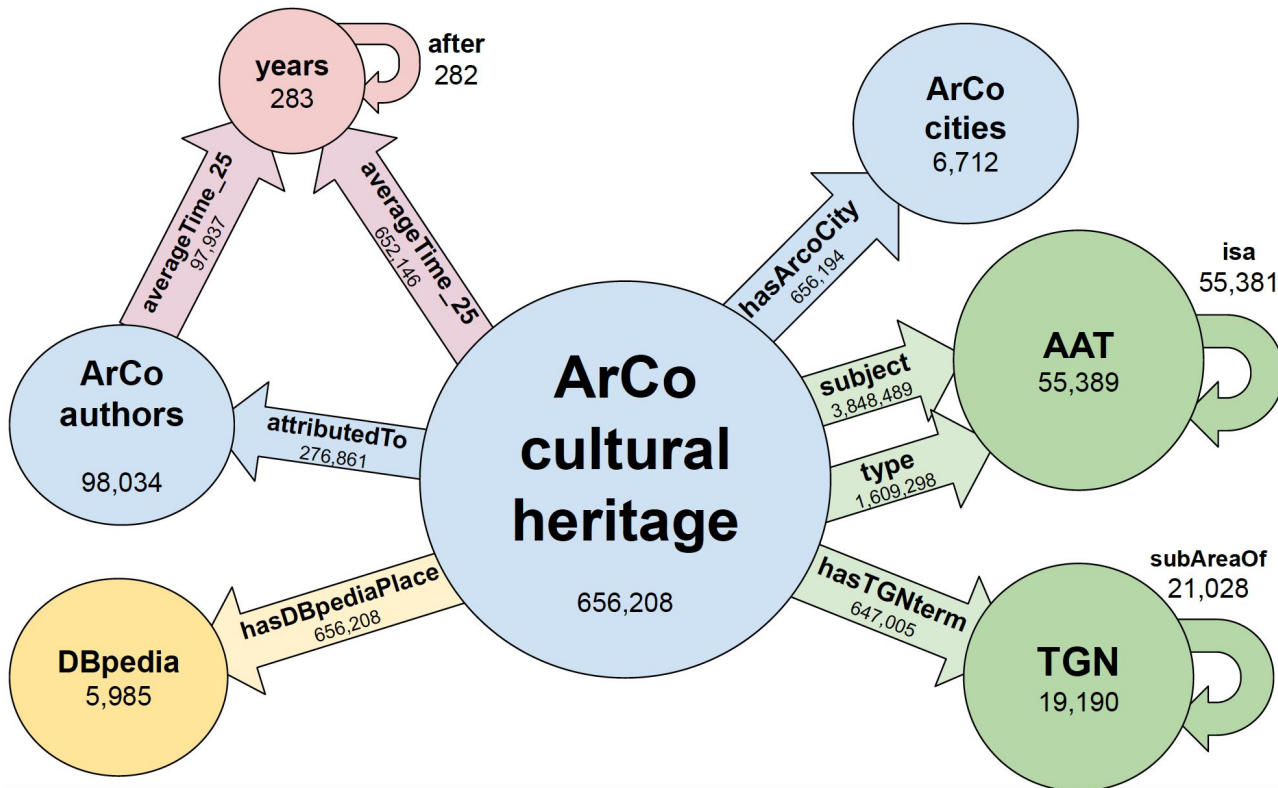
Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage



- Stefano Faralli, Andrea Lenzi, Paola Velardi: AGDLI: ArCo, GVP and DBpedia Linking Initiative. ISWC 2021.
- Stefano Faralli, Andrea Lenzi, Paola Velardi: A Large Interlinked Knowledge Graph of the Italian Cultural Heritage. LREC 2022

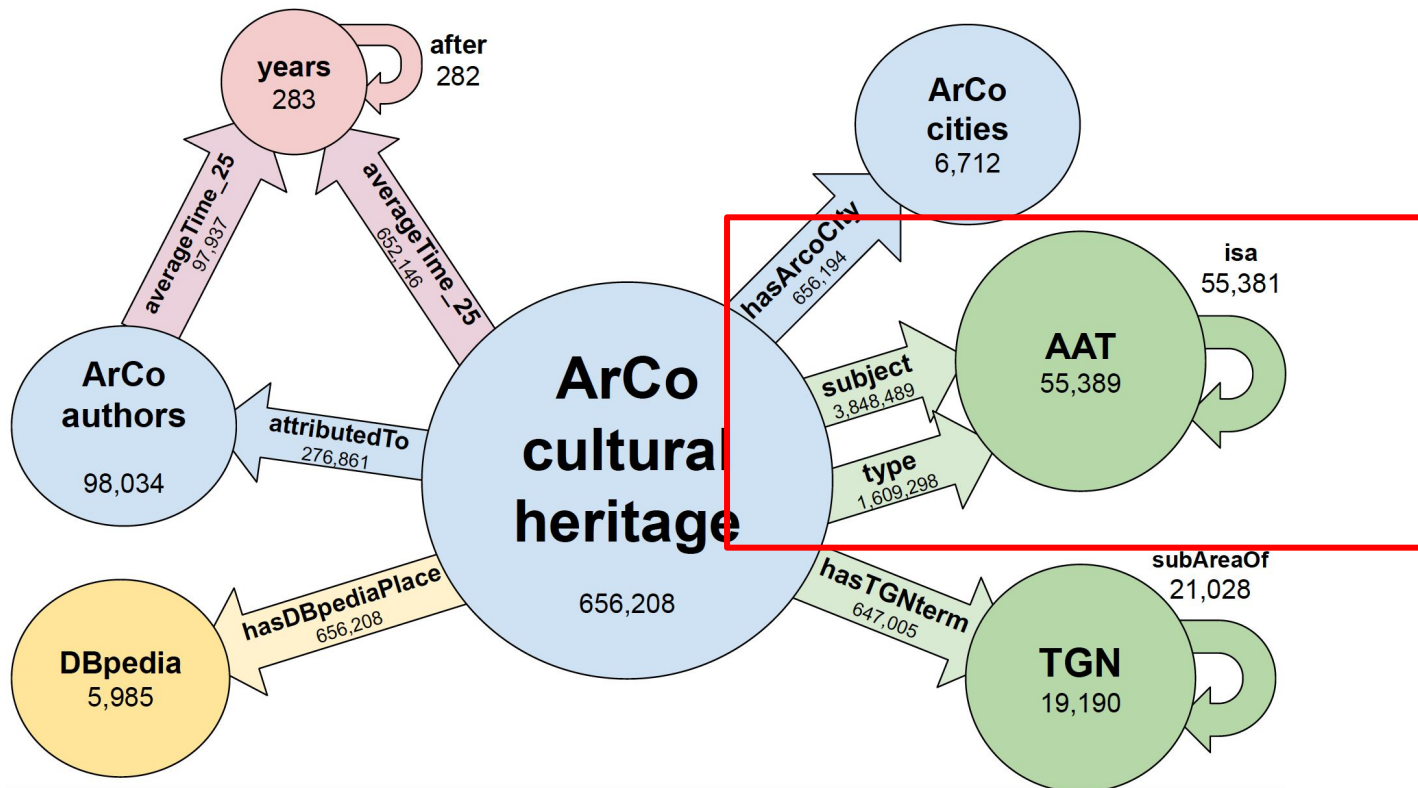


Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage



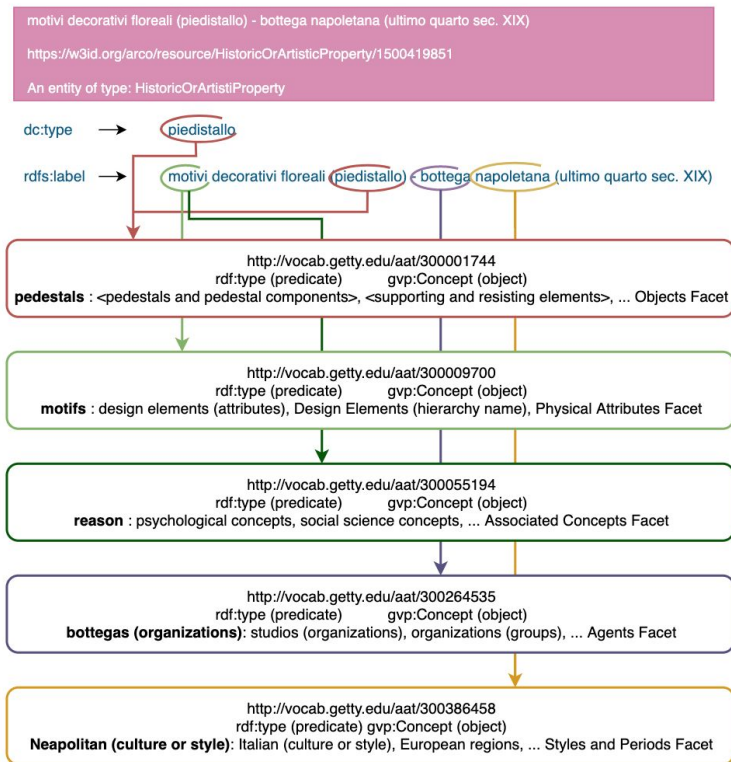


Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage





Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage





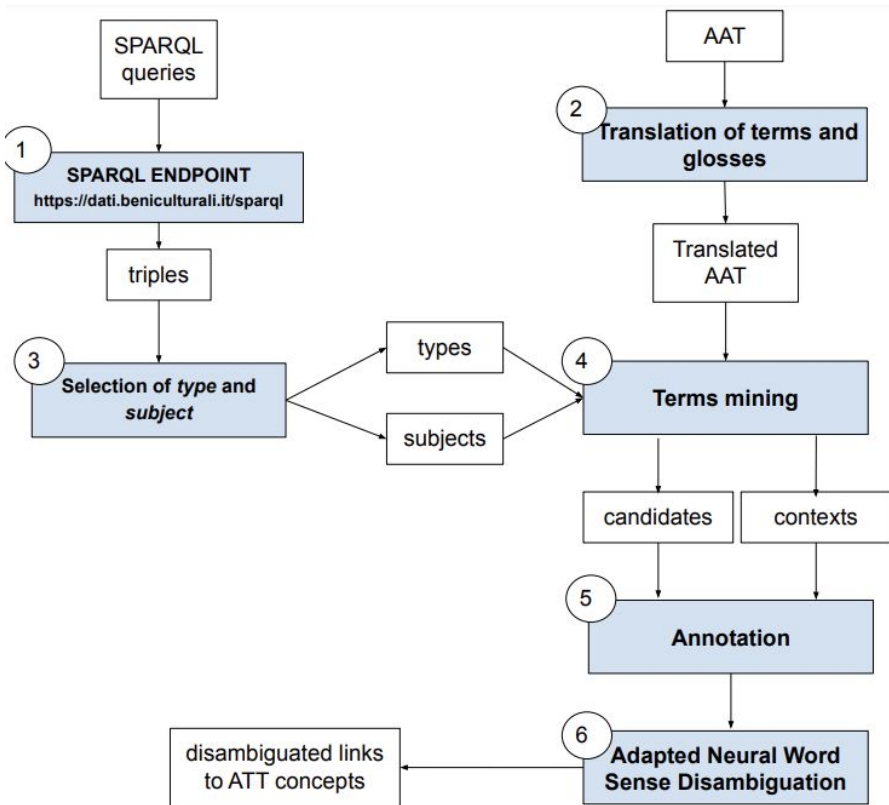
How to remove noisy interlinks?

Our ongoing experimental research is currently focusing on cleaning the resulting knowledge graph with:

- Knowledge Graph Embedding-based models of Triple classification and Link Prediction in noisy settings;
- Neural Word Sense Disambiguation.

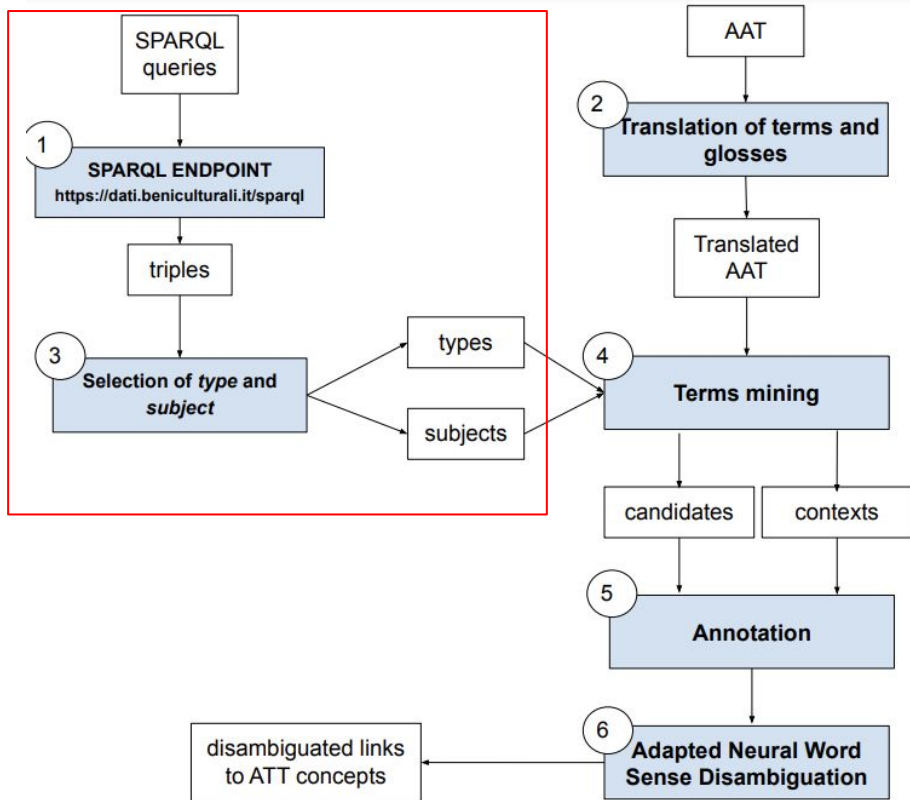


Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage





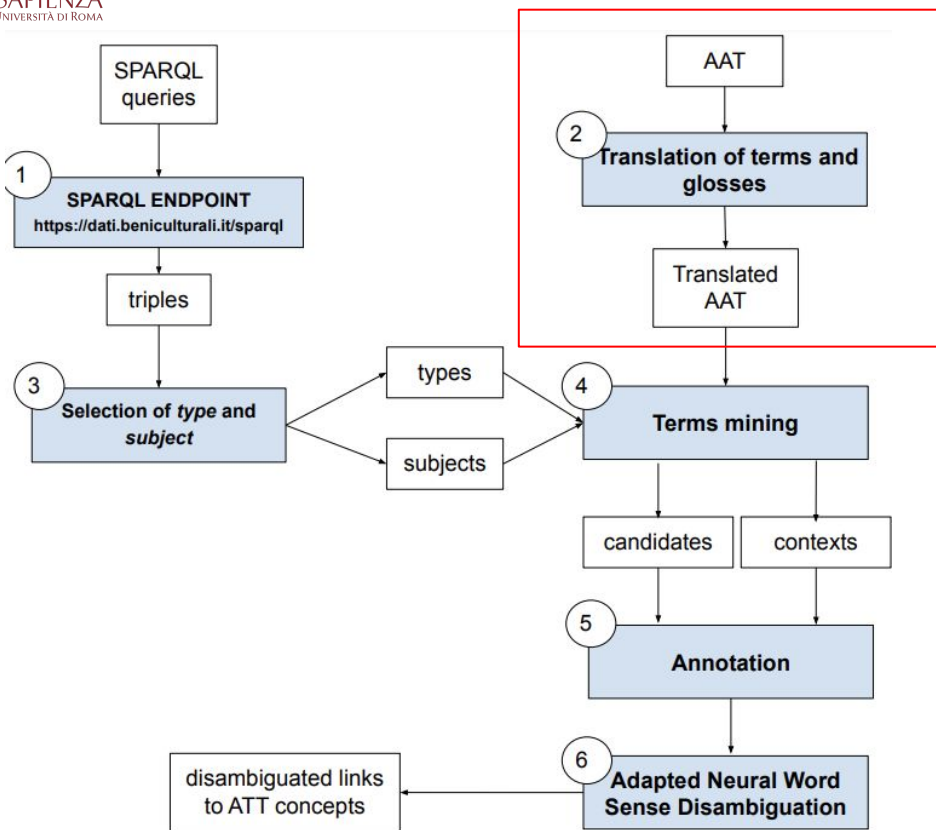
Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage



collection of the subject and type
literal properties of the entities



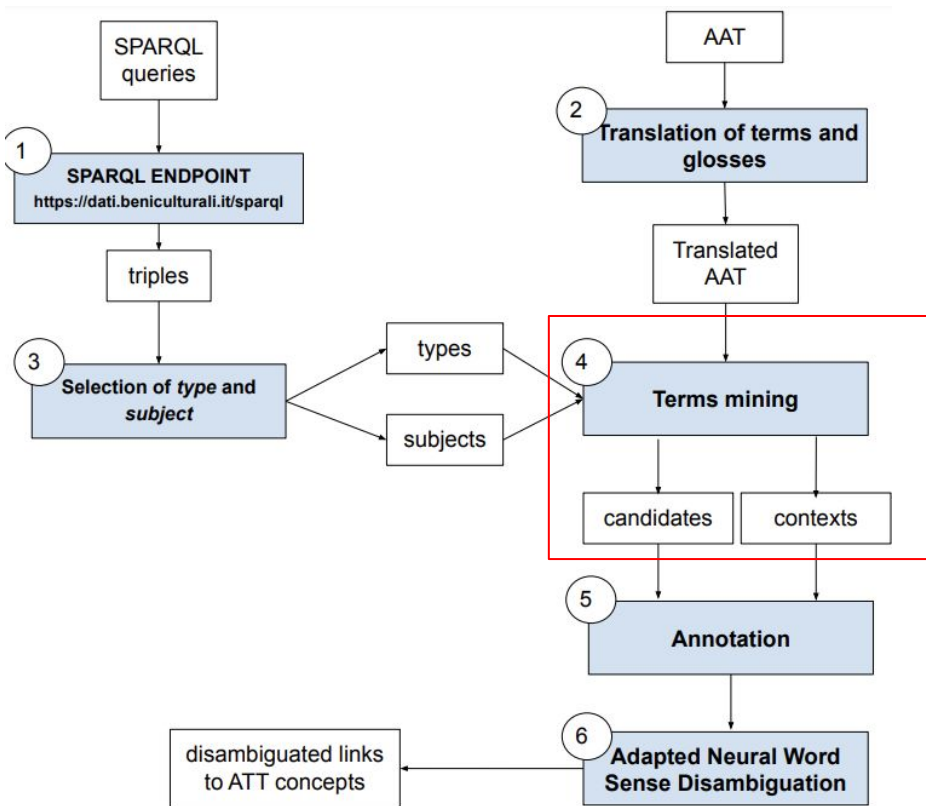
Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage



automatic translation of AAT
terms and glosses



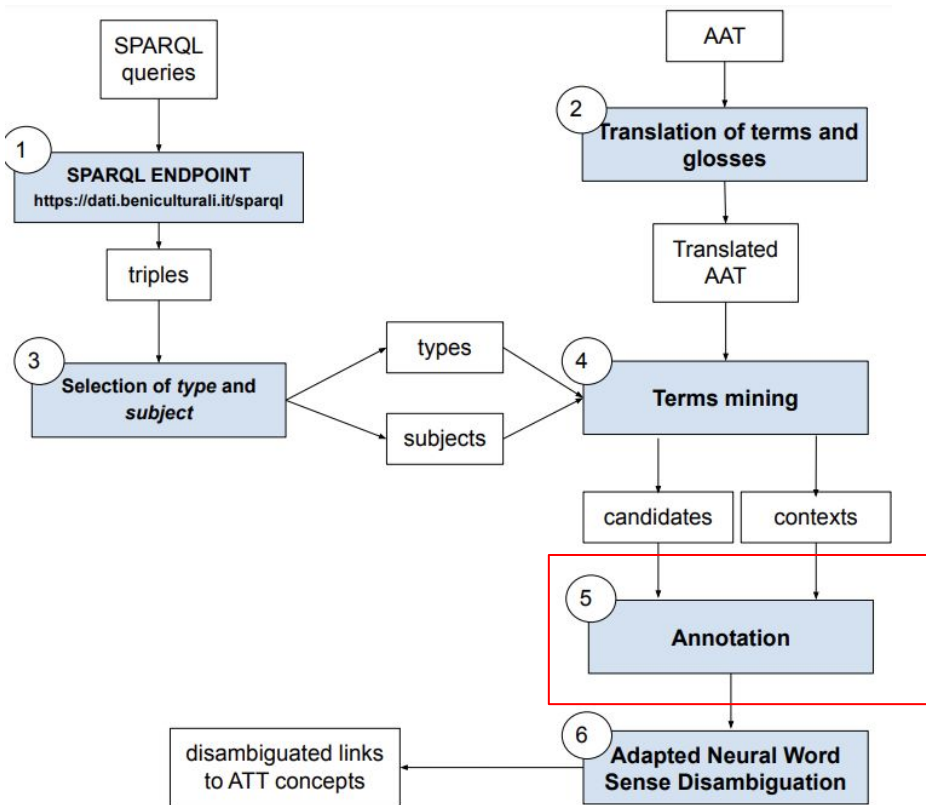
Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage



extraction of terms occurrences
and harvesting of word meaning
candidates in contexts



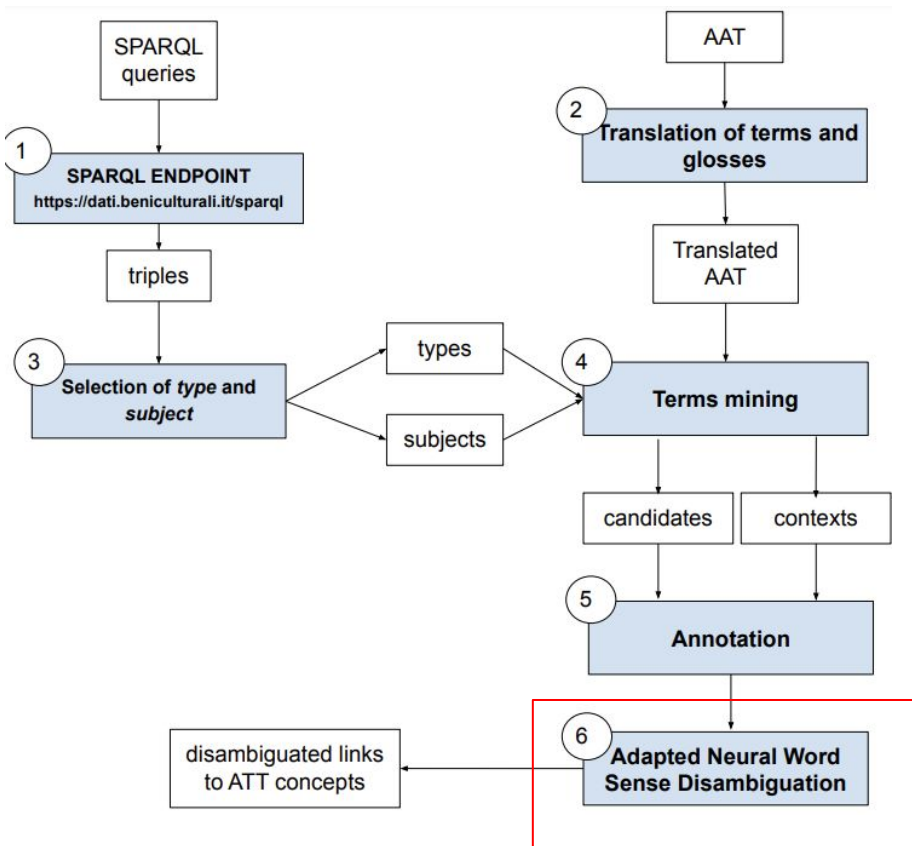
Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage



creation of a manual curated dataset for training and testing purposes



Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage

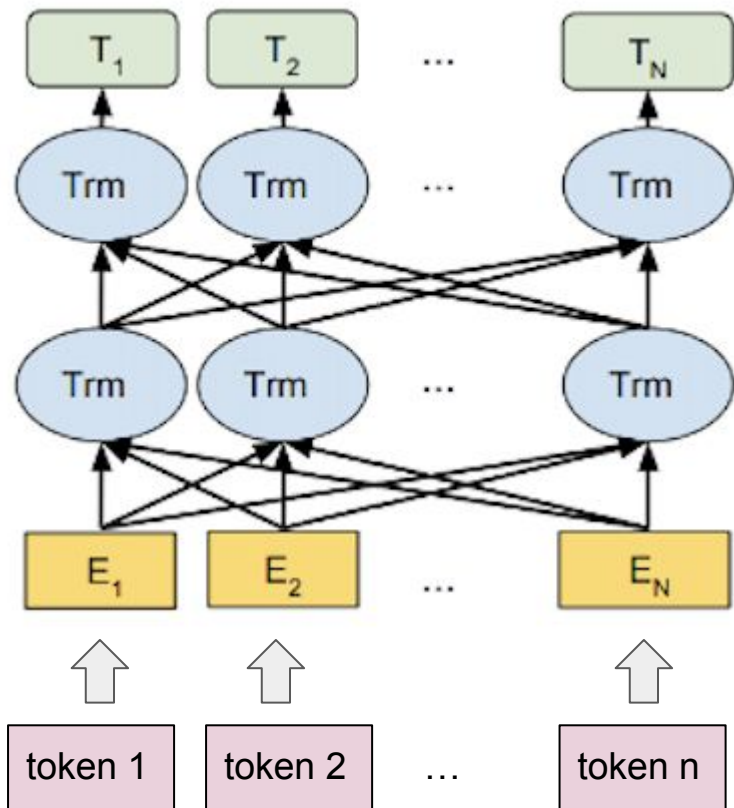


Adaptation of a state-of-the-art
NWSD algorithm

Boon Peng Yap, Andrew Koh, and Eng Siong Chng. 2020. [Adapting BERT for Word Sense Disambiguation with Gloss Selection Objective and Example Sentences](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 41–46, Online. Association for Computational Linguistics.



Pre-Training BERT



BERT

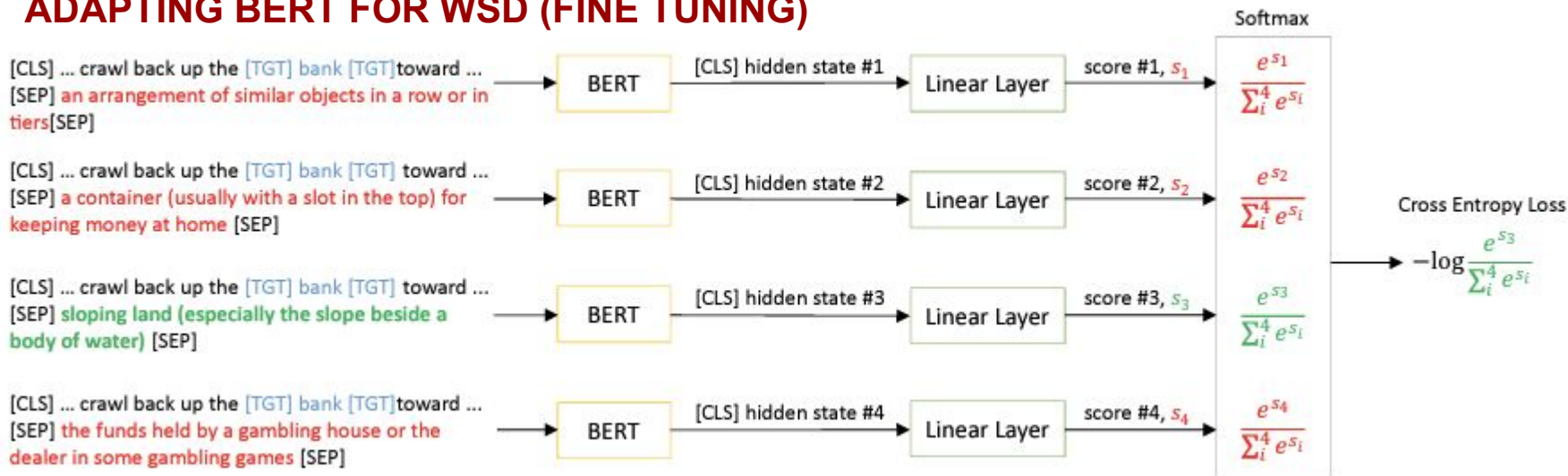
Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) , pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics

ATTENTION NETWORK

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017



ADAPTING BERT FOR WSD (FINE TUNING)



Boon Peng Yap, Andrew Koh, and Eng Siong Chng. 2020. [Adapting BERT for Word Sense Disambiguation with Gloss Selection Objective and Example Sentences](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 41–46, Online. Association for Computational Linguistics.

6-folds cross evaluation results, on our manually curated dataset with adaptation of ItalianBERT for Neural Gloss-based WSD targeting the translated AAT word senses inventory

System	F1	Standard Deviation
Our NWSD system	0.65	0.15
Random Baseline	0.21	0.03

Italian BERT (dbmdz/bert-base-italian-cased)

<https://huggingface.co/dbmdz/bert-base-italian-cased>



ERROR ANALYSIS

- erroneous translations of AAT terms and scope notes.
- missing scope notes for many AAT concepts;
- erroneous text preprocessing, more specifically, errors caused by part of speech tagging of Italian sentences.



CONCLUSIONS

- **Limitations of this work:**

small dataset, errors from automatic translation and NLP preprocessing pipeline

- **Future work:**

- we need a larger dataset of annotated word meanings in contexts;
- we aim to compare the performances of NWSD against those of KGE-based triple classification, link prediction and link deletion systems.



SAPIENZA
UNIVERSITÀ DI ROMA

Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage

SMARTOUR: intelligent platforms for tourism, funded by the Italian Ministry of University and Research.



SMARTOUR



<https://smartour.net>





SAPIENZA
UNIVERSITÀ DI ROMA

Neural Word Sense Disambiguation to prune a large Knowledge Graph of the Italian Cultural Heritage

Thank you for your attention

stefano.faralli@uniroma1.it
faralli@di.uniroma1.it



intelligent information mining

<http://iim.di.uniroma1.it/>