Automatic Semantic Classification of Ancient Zoological Texts

Molka Tounsi Dhouib, Quentin Merilleau, Carla Guerrero, Marco Corneli, Catherine Faron, Arnaud Zucker

1st Symposium in Artificial Intelligence and Applied Mathematics for History and Archaeology
« IAMAHA » - Nice, 27-28 November 2023

Automatic Semantic Classification of Ancient Zoological Texts

Molka Dhouib[1], Quentin Merilleau[2], Carla Guerrero[2],
Marco Corneli[3,4], Catherine Faron[1], Arnaud Zucker[3]

[1]Université Côte d'Azur, CNRS, Laboratoire I3S, Sohpia-Antipolis, France
[2]Université Côte d'Azur, Polytech Nice Sophia
[3]Université Côte d'Azur, CNRS, Labotatoire CEPAM, Nice, France
[4]Université Côte d'Azur, CNRS, INRIA, Laboratoire LJAD, Nice, France
dhouib@i3s.unice.fr

Abstract. We present an approach to semantically annotate the paragraphs of the ancient zoological text *Naturalis Historia* (Pliny the Elder) according to the concepts in the domain thesaurus TheZoo.

Key words. Ancient Zoology, Semantic Classification, Thesaurus.

1 Introduction

This work is conducted in the context of the Zoomathia international research network studying the constitution and transmission of zoological knowledge from Antiquity to the Middle Ages. We combine methods from natural language processing, knowledge representation and machine learning to classify and automate the semantic annotation of ancient texts using the TheZoo thesaurus [Leyra et al., 2015]. We address the following research questions:

(i) What is the best vector representation of concepts and paragraphs that we can use as input for a classifier? (ii) What is the impact of taking into account the semantics captured by a thesaurus in the vector representations of concepts?

2 Proposed approach

We consider two approaches to automatically classify paragraphs of Pliny's *Naturalis Historia* on ancient zoology into one or more macro collections of concepts (i.e. "Places", "Anthroponym", etc.) from the TheZoo thesaurus: (i) **The baseline method** consists into training a Support Vector Machine for each collection separately. In more detail, for a given collection, a binary classifier is trained to label a paragraph with a 1 if one concepts from the collection appears in, 0 otherwise. Each paragraph is represented as a vector of 512 dimensions generated by the Universal Sentence Encoder (USE) [Cer et al., 2018]. (ii) **The knowledge-based method** extends the baseline by using the hierarchical information extracted from the thesaurus. First, we compute the embedded vectors of each concept under the top concept of each collection. Second, we obtain the centroid of each hierarchical group based on the embedded vectors [Tounsi Dhouib et al., 2021]. Finally, we compute the cosine similarity between the vectors representing each centroid and each paragraph.

As a result, each paragraph is represented by a vector of its cosine similarities with respect to the centroids of each group of concepts. We use these similarities to train a classifier.

3 Experiments and results

We evaluated the performance of our approach on the books 8 to 11 of Pliny's *Naturalis Historia*. These four books are divided into paragraphs, which are manually annotated by linguists with concepts from TheZoo. We count 765 paragraphs annotated with 10 collections from TheZoo. Since our dataset is not balanced for all the collections, we used oversampling methods to balance it, and we applied the train-test split procedure with 80% for train and 20% for validation/test. In order to evaluate the performance of the proposed settings, we reported the precision (P), recall (R) and F1 score.

Collection	Baseline method (SVM)			knowledge-based (level 1)			knowledge-based (level 2)		
	P	R	F1	P	R	F1	P	R	F1
Anatomy	0.630	0.615	**0.621**	0.625	0.604	0.613	0.626	0.615	0.619
Anthroponym	0.651	0.599	**0.619**	0.636	0.572	0.597	0.651	0.580	0.609
Environment	0.575	0.546	0.557	0.569	0.539	0.551	0.578	0.546	**0.560**
Ethology	0.550	0.543	0.545	0.557	0.559	**0.556**	0.547	0.548	0.546
Gen. descr.	0.562	0.529	0.543	0.566	0,525	0.543	0.559	0.532	**0.544**
Place	0.665	0.620	0.639	0.673	0.624	**0.645**	0.658	0.603	0.627
Rel. btw man and animal	0.450	0.383	**0.412**	0.446	0.383	0.410	0.439	0.381	0.405
Topic	0.559	0.375	**0.444**	0.645	0.550	0.356	0.532	0.361	0.425
Zoo. info.	0.524	0.383	0.439	0.651	0.532	0.405	0.522	0.402	**0.451**
Zoonyms	0.625	0.600	0.611	0.630	0.606	0.617	0.633	0.610	**0.620**

Tab. 1: Classification methods performance.

Based on the F1 score, we see that our domain Knowledge-based approach outperforms the baseline for some collections such as *Place*, *Zoological information*, *Zoonym*.

References

D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al. Universal sentence encoder for english. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, pages 169–174, 2018.

I. P. Leyra, A. Zucker, and C. F. Zucker. Thezoo: un thesaurus de zoologie ancienne et médiévale pour l'annotation de sources de données hétérogènes. *Archivum Latinitatis Medii Aevi*, 73:321–342, 2015.

M. Tounsi Dhouib, C. Faron, and A. G. Tettamanzi. Measuring clusters of labels in an embedding space to refine relations in ontology alignment. *Journal on Data Semantics*, 10(3-4):399–408, 2021.