Semantic Categorization of Segments of Ancient and Mediaeval Zoological Texts

Catherine Faron Zucker, Irene Pajón Leyra, Konstantina Poulida, Andrea G. B. Tettamanzi

Semantic Categorization of Segments of Ancient and Mediaeval Zoological Texts

Catherine Faron-Zucker[1], Irene Paj´on Leyra[1], Konstantina Poulida[2], Andrea G. B. Tettamanzi[1]

[1] Univ. Nice Sophia Antipolis, France
[2] Inria Sophia Antipolis, France

Abstract. In this paper we present a preliminary work conducted in the framework of the multidisciplinary research network Zoomathia, which aims at studying the transmission of zoological knowledge from Antiquity to the Middle Ages through compilation literature. We propose an approach of knowledge extraction from ancient texts consisting in semantically categorizating text segments based on machine learning methods applied to a representation of segments built by processing their translations in modern languages with Natural Language Processing (NLP) methods and by exploiting a dedicated thesaurus of zoology-related concepts. The final aim is to semantically annotate the ancient texts and reason on these annotations to help epistemologists, historians and philologists in their analysis of these texts.

Keywords: History of Zoology, Knowledge Extraction from Texts, Semantic Categorization

1 Introduction

The Semantic Web has a key role to play to support cultural studies. During the last decade, several works addressed the semantic annotation and search in Cultural Heritage collections and Digital Library systems. They focus on producing Cultural Heritage RDF datasets, aligning these data and their vocabularies on the Linked Data cloud, and exploring and searching among heterogenous semantic data stores. In the framework of the international research network Zoomathia,[3] we address the challenge of adopting such a Linked Data cloud-based approach to support multidisciplinary studies in History of Science. Zoomathia primarily focuses on the transmission of zoological knowledge from Antiquity to the Middle Ages through textual resources, and considers compilation literature such as encyclopaedias.

The automatic annotation of the Zoomathia corpus of selected texts is a first step to enable automatic reasoning on these annotations, supporting the evaluation and interpretation of the development of a zoological knowledge through the ages. The work presented in this paper takes place in the continuation of

[3] `http://www.cepam.cnrs.fr/zoomathia/`

Tounsi et al.'s work presented in [8] on (i) the automatic extraction of zoonyms and zoological topics (ethology, anatomy, medicinal properties, etc.) from the fourth book of the late mediaeval encyclopaedia *Hortus Sanitatis* (15th century), written in Latin and compiling ancient texts on fishes, and (ii) the semantic annotation of the units of this text. The approach for extracting zoonyms was relatively simple, based on a set of patterns (syntactic rules) to recognize the occurence of terms from a taxonomy among the lemmas identified in the Latin texts. The performances of the approach closely depends on the available taxonomic resources. We can now rely on the translation of the TAXREF taxonomic thesaurus of zoological and botanical names in SKOS [2]. As for the extraction of zoological topics, the proposed approach consisted of (i) semi-automatically building a list of semantically related terms for each of the 8 targeted zoological topics, based on the eXtended WordNet Domains[4] (XWND) and BabelNet[5] terminological resources; and (ii) automatically annotating each text segment by a topic when the number of its terms belonging to the set of terms representing a topic was greater than a given threshold. While the overall approach was promising and launched a real dynamic among the participants of the Zoomathia network, the results achieved with the proposed method of knowledge extraction were limited, and the method itself was limited: (i) it required a manual step to build a *representative* set of terms for *each* considered topic; (ii) it required to translate the semantically related terms of each topic in Latin, which had to be done manually by a philologist; (iii) the criterion used to assign a topic to a text segment was too simplistic.

To overcome these limitations, we conceived a possibly more promising method to automatically annotate segments of ancient texts with zoological concepts. First, we take advantage of the terminological work conducted in the meantime in Zoomathia which led to the publication of the THEZOO thesaurus in SKOS, gathering all the zoology-related concepts encountered in Pliny the Elder's *Naturalis Historia* (1st century),[6] considered as representative of the zoological knowledge in the Zoomathia corpus of texts [6]. Second, we reuse state-of-the-art Natural Language Processing (NLP) methods and supervised learning algorithms and libraries for the categorization of text segments. A text segment may be classified into several categories: our classifier is a set of binary classifiers deciding for each considered category whether a segment belongs to it or not. Categories can be any concepts of the THEZOO thesaurus and the semantics of the subsumption relations among concepts are taken into account in our classifier. Third, to take advantage of the amount of available terminological resources developed in the community for modern languages (much more rare for ancient languages), we consider modern translations of ancient texts; and to compensate the possible lost of precision in processing a translation rather than

[4] http://adimen.si.ehu.es/web/XWND

[5] http://babelnet.org/

[6] For the moment, only books VIII–XI are concerned, respectively dealing with: VIII—terrestrial animals; IX—aquatic animals; X—birds; XI—insects and other terrestrial invertebrates.

the original text, we consider several modern translations for each ancient text and we combine the results of their processing. Finally, we use the identified categories to annotate the original ancient text.

Our research question is thus: *How can we effectively categorize ancient text segments by relying on their translation in modern languages and taking advantage of the terminological resources and NLP APIs available for modern languages?*

This paper is organized as follows: Section 2 presents our approach to automatic classification of ancient texts. Section 3 presents the experiments of our approach to the classification of text segments of Book 9 of Pliny's *Naturalis Historia* on aquatic animals and discusses the obtained results. Section 4 concludes.

2 A Semantic Approach to Segment Classification

The problem we tackle is essentially a particular case of text categorization, which may be defined as the classification of documents into a fixed number of predefined categories, where each document may belong in one, more than one, or no category at all [10]. The solution we propose falls within statistical text categorization, in that we rely on machine-learning methods to learn automatic classification rules based on human-labeled training "documents" (in our case, text segments). In addition, to take advantage of linked-data resources and structured domain knowledge, we follow a variant of text segment vector representation whereby the features correspond to *senses* (i.e., meanings) of words or phrases occurring in the text, rather than directly to the words or phrases themselves. In this sense, our approach may be called *semantic*.

By the way, the semantic approach is also a fundamental aspect in the philological work. Precisely, the general idea of THEZOO is to overcome the lexical and grammatical levels of texts and to work at the level of meaning.

One specificity of our problem is that the texts we are interested in categorizing are written in ancient languages (primarily Latin and ancient Greek), for which computational linguistic resources like structured machine-readable lexica and parsers are hard to find, somewhat incomplete, or not interoperable with Semantic Web technologies. We propose, as a workaround, to use one or more translations into modern languages (for which such resources are available) as proxies of the original text. As a matter of fact, translation into modern languages exist for most ancient and medieval texts; furthermore, such translations are of a particularly high quality, being the work of well-trained philologists who strive to convey, as accurately as they can, the full meaning of the ancient text.

2.1 Dataset Construction

Our approach is in two steps. The first one consists in a semantic-based approach for extracting from texts a representation of text segments which will be processed in a second step to categorize them.

We first process the corpus of texts studied to extract from WordNet the list of synsets occuring at least once in the corpus. Then each text segment is represented by a binary vector of the size of this list, indicating the presence or absence of terms belonging to a synset in the segment. The vectors are then weighted by using the term frequency-inverse document frequency (TF-IDF) statistic to reflect how important each synset is to a text segment. This processing step mainly relies on tools available in the Natural Language Toolkit (NLTK).[7]

Second, for each concept of interest in the thesaurus, a binary classifier is constructed, with a training set built by considering the manual annotation of a subset of the text segments in the corpus with terms from the THEZOO thesaurus. This manual annotation activity was conducted by a philologist. At this step, the semantics of the thesaurus is taken into account by considering all the concepts specializing the concept targeted by each classifier.

Finally, with the same training sets, we tested several implementations of classifiers available in the Weka machine learning suite.[8]

2.2 Combining Several Modern Translations

Upon undergoing the treatment described in the previous section, even the most accurate modern translation of an ancient text is likely to introduce noise in the process.

To begin with, contemporary translation studies [4] have made it clear that, when applied to texts of cultural and literary relevance, translation is not just a means of recovering a source text, but also a process of interpretation and production of literary meaning and value. The translator faces multiple choices when having to render the sense of a word or phrase in the target language and some of these choices imply an interpretation of the meaning of the original text which might be subject to debate. Whereas all possible choices are implicitly contained in the original text—in potentiality, once the translator commits to a particular interpretation and choice, there is necessarily a loss of meaning.

At the same time, and besides the possible loss of meaning, there is also the risk of introducing novel meaning, which was not necessarily implied by the original text, and this because the terms employed by the translator to convey the intended meaning of the original text may be ambiguous or polysemous.

One way to obviate both the problem of sense loss and the problem of ambiguity/polysemy is to consider multiple translations, in the same or different modern languages. We concentrate on the case of combining translations in different modern languages, because it is most general and, by solving all the challenges it poses, an approach providing for it is then suitable for dealing with multiple translations in the same language as well. Besides, different languages are not always equally capable of expressing the nuances of the original text. Therefore, using versions in different languages helps recovering a more complete perspective of the original meaning.

[7] `http://www.nltk.org/`
[8] `http://weka.wikispaces.com/`

An essential requirement for combining multiple translations is that the original text and their translations must be aligned. In the case of classic and medieval texts, a conventional segmentation of the text into books, chapters, and paragraphs is generally agreed upon by philologists. Therefore, if the granularity of the segments we are interested in categorizing is, as we assume here, the same as the smallest unit of such traditional segmentation, this step does not pose particular problems, all the more so because, in general, translations into modern languages preserve it.

At the level of a given segment, the combination of multiple translations works as follows:

1. the multiset S_i of synsets giving the senses of all terms (after eliminating stopwords) occurring in each translation T_i of the segment is computed;
2. each multiset S_i is converted into a multiset S'_i by mapping every synset id $s_{ij} \in S_i$ into the corresponding synset id s'_{ij} in the Princeton WordNet; if no corresponding synset id can be determined based on the available index files, s_{ij} is simply dropped;
3. the intersection of the converted multisets, $S = \bigcap_i S'_i$, is computed and it is used as a basis for constructing the feature vector representation of the segment, using the TF-IDF as described above.

The main rationale for taking the intersection of the multisets computed from the various translations is that, by keeping only the senses which are shared among them, we hope to reduce the noise due to polysemous terms occurring in the translations and, in an indirect way, to disambiguate the original text. One possible drawback of taking the intersection is that, if two of the translations considered were based on radically different interpretations of the original text, the synsets corresponding to some important term in the original text might disappear altogether. However, this is very unlikely to happen in reality, for even if two different sense of the same word are construed by two translators, chances are that the terms employed to render them are not too distant semantically, so that the intersection of their respective synsets is not empty. A quantitative investigation of this claim, however, is left for future work.

3 Experiments

To test our approach, we focused our attention on Book 9 of Pliny the Elder's *Naturalis Historia* on aquatic animals, which consists of 186 paragraphs. In this case, paragraphs are the segments of text which are categorized; on average they are 56 word long.

We have used translations which are now in the public domain, namely [1, 7] for English, [5] for French, and [9] for German. As for linguistic resources, we have used Princeton WordNet[9] for English, WOLF (Wordnet Libre du Français)[10] for French, and GermaNet[11] for German.

[9] `https://wordnet.princeton.edu/`
[10] `http://alpage.inria.fr/~sagot/wolf-en.html`
[11] `http://www.sfs.uni-tuebingen.de/GermaNet/index.shtml`

Seven pairs of training and test datasets have been constructed for the following translation languages or combinations of languages:

1. English;
2. French;
3. German;
4. English and French;
5. English and German;
6. French and German;
7. English, French, and German.

Each paragraph has been transformed into a vector of features, where each feature is the TF-IDF in the paragraph of a synset whose lexicalization occurs in the translations of Book 9 in the modern languages considered. When the translations in two or three modern languages are considered, the synsets of languages other than English are translated into the corresponding Princeton Wordnet synset and the intersection of the synsets from each modern translation is taken for computing the feature vector for that paragraph.

We manually assigned paragraphs (and, by extension, their associated feature vectors) to the categories corresponding to their topic. A paragraph may belong to more than one category.

The training and test datasets for a *category* C (i.e., a topic against which paragraphs are to be classified) are obtained by randomly selecting half of the feature vectors (or records) classified as C and half of the feature vectors classified as $\neg C$ for the training dataset and the remaining half for the test set, so that the training and test datasets contain the same fraction of C and $\neg C$ records.

The datasets thus obtained, however, are imbalanced. For instance, out of the 186 paragraphs in Book 9 of Pliny the Elder's *Naturalis Historia*, 55, or 29.6%, are about "anatomy"; most paragraph are not about anatomy. Such an imbalance, if not properly corrected, may lead many classification methods to take the shortcut of classifying all paragraphs as "not-anatomy", which would be an easy way of obtaining a 70% accuracy.

Random under- and oversampling are two popular techniques to obtain a balanced training set from an imbalanced one [3]. However, undersampling, which works by removing examples from the most represented class, is not suitable for cases, like ours, where training data are scarce and could potentially remove certain important examples; random oversampling, which injects into the least represented class additional copies of its examples, on the other hand, may lead to overfitting if some examples get sampled more than others. To obviate this problem, we adopted a deterministic oversampling strategy which constructs a perfectly balanced dataset of a size n much larger than the size of the original imbalanced dataset by alternatively picking an example from either class and wrapping around when all the examples of a class have been exhausted, as shown in Algorithm 1. As a result, two examples of the same class will always get sampled a number of times which can differ by at most 1. By taking a sufficiently large n, one can make the maximum deviation between the frequency of examples as small as desired.

Algorithm 1 BALANCE(d, n)

Input: an imbalanced dataset d, containing records of two classes, C and $\neg C$; an even integer n, much greater than the size of d;
Output: b: a balanced dataset of size n, containing exactly $n/2$ records of class C and $n/2$ records of class $\neg C$.

```
1: list_C ← [ ]
2: list_¬C ← [ ]
3: for record ∈ d do
4:    if record[class] = C then
5:       list_C.APPEND(record)
6:    else
7:       list_¬C.APPEND(record)
8: b ← [ ]
9: for i = 1, ..., n/2 do
10:   b.APPEND(list_C[i (mod list_C.LENGTH())])
11:   b.APPEND(list_¬C[i (mod list_¬C.LENGTH())])
12: return b
```

Table 1. Performance of the approach on Book 9 of *Naturalis Historia* for the category "anatomy".

Combination			Accuracy	Precision	Recall	F-Measure
en			59.6%	8.3%	3.6%	0.537
	fr		62.8%	11.1%	3.6%	0.555
		de	69.1%	42.9%	10.7%	0.620
en	fr		57.4%	7.1%	3.6%	0.524
en		de	57.4%	12.5%	7.1%	0.534
	fr	de	59.6%	27.3%	21.4%	0.580
en	fr	de	55.3%	18.2%	14.3%	0.536

Specifically, for our experiments, we set $n = 1000$.

A number classification methods implemented in Weka, including complement and multinomial naive Bayes, k-nearest neighbors, and the support vector machines have been applied to the datasets thus obtained. Support vector machines proved to give the best results.

Table 1 summarizes the results obtained by support vector machines when used to classify paragraphs not used for training (test set) with respect to category "anatomy". In this table, accuracy is the percentage of correct classifications; precision is the percentage of paragraphs classified as "anatomy" by the model that were annotated as such by the human expert; recall is the percentage of paragraphs annotated as "anatomy" that are correctly recognized; F-measure is the average of the F-scores for class "anatomy"and for its complement.

In terms of accuracy, these results constitute an improvement over the results obtained in [8].

Although the performance in terms of accuracy looks promising, in reality, when one focuses on the capability of the classification model to recognize and

thus automatically annotate a paragraph about a given topic (category), these results are quite disappointing, with precision and recall figures well below an acceptable level.

A rather surprising fact, which calls for a more in-depth investigation, is that the results obtained by combining translations in three languages (cf. the last row of Table 1) are no better than those obtained by combining translations in two languages, which, in turn, are no better than those obtained by considering a single translation. This preliminary evidence would thus suggest that combining translations in different languages is not a good idea, but we are cautious to jump to such conclusions and we think more evidence based on a larger corpus of texts should be gathered before dismissing this proposal.

4 Conclusion and Future Work

Despite the disappointing preliminary results, we believe the proposed approach to have the potential to provide a viable solution to the problem of automatic or semi-automatic annotation of ancient texts.

We think that the reason for the observed poor performance of the classification models in our preliminary experiments may be twofold: on the one hand, the number of examples available for training the models is exceedingly small, in the face of a very high-dimensional feature space (ranging from 2,500 to 10,500 features); on the other hand, the features that could prove useful to reach the correct classification are drowned among all the other features. Coming up with a heuristic to select a small number of relevant features given a category would probably alleviate both problems. We plan on concentrating our future efforts in that direction. In addition, we are aware that many tools for semi-automatic analysis are currently under development, for exemple in the Perseus Project. Currently, NLTK does not enable to the exploit the Latin or Classical Greek version of WordNet. For some phases of our work, perhaps a framework like the Classical Language Toolkit,[12] an extension of NLTK, could be useful. Conversely, our research work described here could somehow contribute to these efforts. For example, we plan on aligning the THEZOO thesaurus with WordNet. An implicit assumption of our methodological choice is that the categories in ancient, medieval, 19th-century and contemporary texts are supposed to match perfectly. Of course this is not the case and using a specific thesaurus like THEZOO might contribute to make our approach more anthropologically-aware.

Acknowledgments. Zoomathia is an International Research Group (GDRI) supported by the French National Scientific Research Center (CNRS).

GermaNet[13] is a German lexical-semantic resource developed at the Linguistics Department of the University of Tübingen.

[12] `http://cltk.org`
[13] http://www.sfs.uni-tuebingen.de/GermaNet/

References

1. J. Bostock and H. T. Riley, editors. *Pliny the Elder, The Natural History, Vol. II.* Taylor and Francis, London, 1890.
2. C. Callou, F. Michel, C. Faron-Zucker, C. Martin, and J. Montagnat. Towards a shared reference thesaurus for studies on history of zoology, archaeozoology and conservation biology. In A. Zucker, I. Draelants, C. Faron-Zucker, and A. Monnin, editors, *Proceedings of the First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference, Portorož, Slovenia, June 1st, 2015.*, volume 1364 of *CEUR Workshop Proceedings*, pages 15–22. CEUR-WS.org, 2015.
3. N. V. Chawla. Data mining for imbalanced datasets: An overview. In O. Maimon and L. Rokach, editors, *Data Mining and Knowledge Discovery Handbook, 2nd ed.*, pages 875–886. Springer, 2010.
4. E. Gentzler. *Contemporary Translation Theories: Revised 2nd Edition.* Multilingual Matters, Clevedon, 2001.
5. M. E. Littré, editor. *Histoire Naturelle de Pline, avec la traduction en français.* Firmin-Didot et C^{ie}, Paris, 1877.
6. I. Pajón-Leyra, A. Zucker, and C. Faron-Zucker. Thezoo : un thésaurus de zoologie ancienne et médiévale pour l'annotation de sources de données hétérogènes. *to appear in ALMA (Archivum Latinitatis Medii Aevi)*, 73, 2015.
7. H. Rackham, editor. *Pliny: Natural History volume III (Books VIII–XI).* Cambridge, Massachusetts, 1940.
8. M. Tounsi, C. Faron-Zucker, A. Zucker, S. Villata, and E. Cabrio. Studying the history of pre-modern zoology by extracting linked zoological data from mediaeval texts and reasoning on it. In *The Semantic Web: ESWC 2015 Satellite Events, Portorož, Slovenia, 2015, Revised Selected Papers*, volume 9341 of *LNCS*. Springer, 2015.
9. G. C. Wittstein, editor. *Die Naturgeschichte des Cajus Plinius Secundus, ins Deutsche übersetzt und mit Anmerkungen versehen, zweiter Band (VII–XI Buch).* Gressner & Schramm, Leipzig, 1881.
10. Y. Yang and T. Joachims. Text categorization. *Scholarpedia*, 3(5):4242, 2008.