Word Embeddings Pointing the Way for Late Antiquity

Johannes Bjerva
University of Groningen
The Netherlands
j.bjerva@rug.nl

Raf Praet
University of Groningen
The Netherlands
r.g.l.praet@rug.nl

Abstract

Continuous space representations of words are currently at the core of many state-of-the-art approaches to problems in natural language processing. In spite of several advantages of using such methods, they have seen little usage within digital humanities. In this paper, we show a case study of how such models can be used to find interesting relationships within the field of late antiquity. We use a word2vec model trained on over one billion words of Latin to investigate the relationships between persons and concepts of interest from works of the 6^{th}-century scholar Cassiodorus. The results show that the method has high potential to aid the humanities scholar, but that caution must be taken as the analysis requires the assessment by the traditional historian.

1 Introduction

Continuous space representations of words are currently the backbone of several state-of-the-art approaches to problems in natural language processing. The distributional hypothesis, summarised as: 'You shall know a word by the company it keeps' (Firth, 1957) is the basis of many approaches for obtaining such representations. Word embeddings are an example of such a model (e.g. Collobert and Weston (2008)), and have been found to encapsulate interesting semantic properties; in a model presented by Mikolov et al. (2013b), the result when calculating $\overrightarrow{\text{KING}} - \overrightarrow{\text{MAN}} + \overrightarrow{\text{WOMAN}}$, is close to $\overrightarrow{\text{QUEEN}}$.

In this work in progress within the Cassiodigitalis project, we investigate how such representations can be adapted to aid humanities researchers, using the case of late antiquity as an example. Using such a model has several advantages, such as speed and cost-effectiveness. An automated method such as presented here can save time by, e.g., finding potentially interesting interrelations between historical figures and concepts, or quantitatively corroborate results of an otherwise qualitative study. It is, of course, not expected that this can replace the manual perusal of a historian. The goal is, indeed, to use these models to point the way for late antiquity. We expect that the method outlined in this paper can also be used for other disciplines within the humanities.

The digital approach is easily applicable to historical research of periods which are highly documented, i.e., from the beginning of printing up to today. Yet in this paper we want to ascertain whether a digital approach could be relevant to periods which are less documented. As for classical studies, the field of late antique studies is relatively recent, following the seminal work of Brown (1971). This crucial period of transition from antiquity to the middle ages could prove a fertile ground for a digital approach; the late antique world abounds in dense networks of scholars and politicians who publish their letters in order to further their ambitions. We chose to focus on the person of Cassiodorus for several reasons; this 6^{th}-century scholar was a pivotal figure in the transition of classical literature and knowledge through his Vivarium monastery (O'Donnell, 1979). Yet many aspects of his biography remain enigmatic. A digital analysis of his vast oeuvre could show us new ways to answer questions as why Cassiodorus abandoned his political ambitions to found his monastery.

In this paper, we demonstrate how word embeddings can be used to aid humanities scholars by showing that relations between concepts and historical characters can be found and corroborated. The embeddings used in this paper are released along with scripts to reproduce our plots.[1]

[1] github.com/bjerva/cassiodigitalis

Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 53–57,
Beijing, China, July 30, 2015.

The rest of the paper is organised as follows. Related work is briefly covered in Section 2. The methodology is detailed in Section 3. Section 4 contains the experiment overview and results. We discuss the results in Section 5 and conclude in Section 6.

2 Related Work

Word embeddings have seen much recent use within computational linguistics, however usage within digital humanities appears to be limited. Recent work by Koopman et al. (2015) employs vector representations to calculate similarities between entities such as authors and journals in an article database. Usage of word embeddings in the humanities is further discussed by Tahmasebi et al. (2015), who suggest that they could be useful for comparing word vectors trained on different epochs of time, thus revealing changes in usage of words across time. The usage of digital methods within the late antiquities in particular largely focusses on approaches such as the use of geopositioning data to aid classicists and archaeologists, or linking data from, e.g., funerary monuments in order to facilitate research (cf. Bodard and Mahony (2012)). This field should therefore provide fertile grounds for this relatively new approach. Our contribution to previous work thus constitutes a first study showing concrete usages of word embeddings for the late antiquities in particular, and digital humanities in general.

3 Method

The core of the method used in this paper is based on the freely available *word2vec* tool, which can be used to quickly create high quality word embeddings based on a large corpus of text (Mikolov et al., 2013a).[2] We train word2vec using parameters similar to those used for the best performing English vectors in Baroni et al. (2014). We use the continuous bag-of-words model, a window size of 5, a vector dimensionality of 400, 10 negative samples and set subsampling to $1e^{-5}$. We further allow the model to train on the corpus over the course of 100 epochs.

3.1 Data

3.1.1 Large Corpus of Latin

Our word2vec model is trained on a large corpus of Latin texts, containing about 1.38 billion tokens, collected from 11,261 texts spanning two millennia of use of Latin (Bamman and Crane, 2011; Bamman and Smith, 2012). This corpus is freely available.[3] The texts have been manually confirmed as containing Latin text. Seeing as the texts have been OCR-scanned, the quality varies widely. Prior to training the word2vec model, we pre-process the corpus in order to reduce noise. We convert all text to lower case, remove all punctuation and non-alphanumeric characters.

3.1.2 Cassiodorus' Variae

Flavius Magnus Aurelius Cassiodorus Senator (c. 485 – c. 585) served under the Ostrogothic king Theodoric and his successors until the collapse of the kingdom under the Byzantine armies (535 – c. 540). After his stay (or detention) in Constantinople (c. 540 – 554), he fully concentrated on his own Christian didactical project within the confines of his Vivarium monastery in the south of Italy (O'Donnell, 1979). The main testimony to his political career were the Variae, a collection of state letters in twelve books (Fridh and Halporn, 1973; Zecchini et al., 2014). Cassiodorus wrote them on behalf of king Theodoric, his successors, or on his own account as praetorian prefect. The date of the compilation and composition of the Variae is posited between 540 and the mid-540's. In this paper, the Variae are used as a source of historical figures and concepts.

4 Mapping Person – Concept Relations

Our experiment deals with investigating relations between historical figures and central concepts in the period in question. We compile a list of six concepts with their related Latin words which were deemed relevant for the investigation in question. The selected concepts are shown in Table 1. We further compile a list of 14 persons of interest within Cassiodorus' Variae. These historical figures were selected based on their proximity to Cassiodorus and their significance in the presentation of Cassiodorus and his Ostrogothic masters in the Variae. We selected several historical characters who were Cassiodorus' peers and competitors in cultural networks (Boethius, Symmachus), political networks (Liberius) and ecclesiastical circles (Agapetus). Furthermore we added representatives of the political forces with whom the Ostrogothic kingdom in Italy in-

[2] code.google.com/p/word2vec/

[3] cs.cmu.edu/~dbamman/latin.html

teracted and competed: apart from the Ostrogothic kings themselves (Theodoric, Athalaric, Theodahad), we have their barbarian predecessors (Alaric and Odoacer), and the Roman emperors from the Byzantine east (Anastasius, Iustinianus, Theodora). This selection of persons, along with relevant details, is shown in Table 2.

Table 1: Relevant concepts used in the study, with related Latin words.

Concept	Words		
Modernity	Modernus	Novus	Novitas
Romanness	Romuleus	Quirites	Latialis
Greekness	Graecus	Graeculus	Atticus
Gothness	Gothus	Hamalus	Gothicus
Antiquity	Vetus	Antiquitas	Senex
Liberty	Libertas	Libertatus	Liber

Table 2: Persons of interest used in the study, along with personal details.

Name	Status	Lifetime
Cassiodorus	scholar, Ostrogothic official	c. 485 – c. 585
Theoderic	Ostrogothic king of Italy	454 – 526
Alaricus	Visigothic king	c. 370 – 410
Odovacer	barbarian general, king of Italy	433 – 494
Athalaricus	Ostrogothic king of Italy	516 – 534
Theodahadus	Ostrogothic king of Italy	c. 480 – 536
Anastasius	Byzantine emperor	c. 431 – 518
Iustinianus	Byzantine emperor	c. 482 – 565
Theodora	Byzantine empress	c. 500 – 548
Boethius	scholar, Ostrogothic official	c. 480 – 524
Symmachus	mecenas, Ostrogothic official	? – 526
Liberius	Ostrogothic/Roman official	c. 465 – c. 554
Agapetus	pope	? – 536

We calculate the relatedness between each person and concept as follows. For each concept, x, we amass a set of vectors $\mathbb{X}$ based on the related Latin words. For each person, y, we use the vector representation in our model based on the nominative form of the person's name. We then find the smallest cosine distance between each vector representation of a concept, $\vec{x_i} \in \mathbb{X}$, and each person's vector representation, $\vec{y}$. We take this distance to be a measure of the relationship between a person of interest and the concept in question.

Before visualizing the results, we split the persons of interest into two groups. Group 1 consists of the leading figures of the 6^{th}-century political patchwork. Group 2 consists of Cassiodorus' colleagues and competitors. Heat maps of the relationships between each person and concept are shown in Figure 1 and Figure 2. Blue is used to indicate an absent or relatively small relationship, while red is used to indicate a relatively strong relationship.

5 Discussion

5.1 The Blurring of the Barbarian-Roman Boundary

Whereas the Visigoth king Alaric and the barbarian general Odoacer are intensively associated with words which denote the Goths, this association fades away with the rulers of the Ostrogoth kingdom in Italy: Theodoric and his successors Athalaric and Theodahad (see Figure 1). This could reflect the success of Theodoric's cultural and political profiling as being the true heir to the Roman legacy in Italy (Jones, 1962; Heather, 1992). This diminishing association with the Goths does not, however, correlate with an increasing association with the Romans; Theodoric and Theodahad's association with "Romanness" are rather meagre in comparison with Odoacer. This could be explained by Odoacer's exemplary role as the general who officially put an end to the Western Roman empire by deposing its last emperor, Romulus Augustulus (ca. 464 – ca. 507). This negative association with the legacy of Rome apparently endured in the reception of this historical character.

5.2 The Roman Empire is Dead, Long Live the Byzantine Empire!

The transition of the Roman empire into a medieval Byzantine empire was a gradual and evasive process, which cannot be exactly pinpointed in time. However, the rule of the emperor Justinian has been considered to be pivotal in this gradual process (Maas, 2005). The digital approach seems to corroborate Justinian's role; whereas there still is a high association between "Romanness" and Anastasius, the emperor of the Eastern Roman empire before Justinian's dynasty, this association dramatically diminishes in the case of Justinian (see Figure 1). This would mean that in the Latin sources, or, from a western perspective, Justinian is considered emperor of the Greeks instead of Roman emperor. However, caution has to be exhibited when comparing Justinian to Anastasius, as there are several historical characters extant with the name Anastasius and the word representations used only consider the surface forms of the names in question.

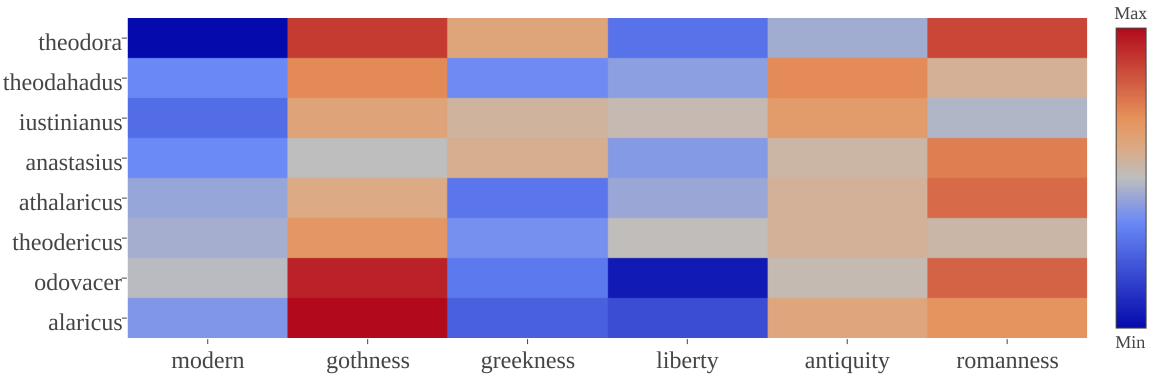

Figure 1: Heatmap showing relationships between persons and concepts in group 1.

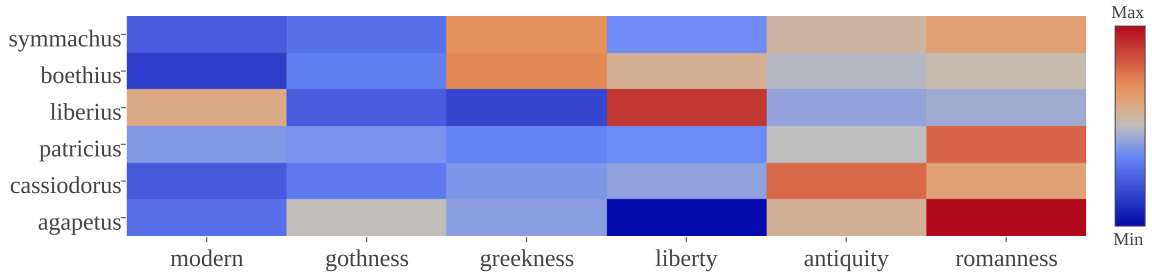

Figure 2: Heatmap showing relationships between persons and concepts in group 2.

5.3 Different Intellectual Profiles

When we compare the results of the contemporaries Cassiodorus, Liberius, Symmachus and Boethius, we can see some differences. Symmachus and Boethius have a distinct association with the Greek cultural sphere, whereas Cassiodorus and Liberius lack this link (see Figure 2). In this case the results shed a light on the social networks and cultural aspirations of both pairs. Symmachus and his son-in-law Boethius had, as members of the senatorial aristocracy, close ties with their counterpart in the Greek Eastern Roman empire, as they still cherished the cultural ideal of a bilingual Roman legacy. Boethius translated Greek philosophical treatises, and Symmachus was involved in the bilingual project of the grammarian Priscian of Caesarea (around 500) (Marenbon, 2003). Liberius and Cassiodorus foreshadow the gradual disintegration of the links between the east and the west. Liberius' long political career was mainly based in Gaul and Italy (O'Donnell, 1981), whereas Cassiodorus was active in the administration of the Ostrogothic realm in Italy. Cassiodorus' association with antiquity points to his success as central figure in the transmission of ancient works of literature and science through his Vivarium monastery (see Figure 2). Furthermore, this association can be traced to the meticulous self-presentation in his letter collection Variae as an ardent intellectual. The high association between Liberius and the concept of liberty should be disregarded because of linguistic reasons; naturally there is a high association between the concept of liberty and a name which can also be a form of the adjective liber, 'free'.

6 Conclusion

In this paper, we have shown an example of how word embeddings can be used to point the way for late antiquity, and in extension, humanities. Such a digital method has high potential to aid the humanities scholar in assessing different historiographical questions. Not only do the results corroborate or nuance the findings of qualitative research. Surprising results also generate new historiographical questions. Nevertheless, the example of Liberius and liberty urges to exhibit caution; the digital approach cannot be used without the guiding assessment of the traditional historian.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful suggestions, and Vivian Bouwer for her input on earlier versions of this manuscript.

References

Bamman, D. and Crane, G. (2011). Measuring historical word sense variation. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, pages 1–10. ACM.

Bamman, D. and Smith, D. (2012). Extracting two thousand years of latin from a million book library. *Journal on Computing and Cultural Heritage (JOCCH)*, 5(1):2.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1.

Bodard, G. and Mahony, M. S. (2012). *Digital research in the study of classical antiquity*. Ashgate Publishing, Ltd.

Brown, P. (1971). *The World of Late Antiquity: from Marcus Aurelius to Muhammad*. London.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Firth, J. R. (1957). A synopsis of linguistic theory. pages 1930–1955. 1952–1959:1–32.

Fridh, A. and Halporn, J. (1973). Magni aurelii cassiodori senatoris opera pars i: Variarum libri xii. *Corpus Christianorum, Series Latina, 96.*

Heather, P. J. (1992). The historical culture of Ostrogothic Italy. In *Teoderico il Grande e i Goti d'Italia*. Atti del XIII Congresso internazionale di studi sull'Alto Medioevo.

Jones, A. H. M. (1962). The constitutional position of Odoacer and Theoderic. *Journal of Roman Studies*, 52(1-2):126–130.

Koopman, R., Wang, S., Scharnhorst, A., and Englebienne, G. (2015). Ariadne's thread: Interactive navigation in a world of networked information. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1833–1838. ACM.

Maas, M. (2005). Roman Questions, Byzantine Answers. *The Cambridge Companion to the Age of Justinian (Cambridge: Cambridge University Press)*, pages 3–27.

Marenbon, J. (2003). *Boethius*. Oxford University Press.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, pages 746–751.

O'Donnell, J. J. (1979). *Cassiodorus*. Los Angeles: University of California Press.

O'Donnell, J. J. (1981). Liberius the Patrician. *Traditio*, pages 31–72.

Tahmasebi, N., Borin, L., Capannini, G., Dubhashi, D., Exner, P., Forsberg, M., Gossen, G., Johansson, F. D., Johansson, R., Kågebäck, M., et al. (2015). Visions and open challenges for a knowledge-based culturomics. *International Journal on Digital Libraries*, 15(2-4):169–187.

Zecchini, G., Giardina, A., Cecconi, G., Tantillo, I., Oppedisano, F., Marcone, A., Lo Cascio, E., LA Rocca, A., La Rocca, C., Neri, V., et al. (2014). *Cassiodoro Varie. Volume 2: Libri III, IV, V*. Erma di Bretschneider.