Arretium or *Arezzo*? A Neural Approach to the Identification of Place Names in Historical Texts

Rachele Sprugnoli
Fondazione Bruno Kessler, Via Sommarive 18, Povo (TN)
sprugnoli@fbk.eu

Abstract

English. This paper presents the application of a neural architecture to the identification of place names in English historical texts. We test the impact of different word embeddings and we compare the results to the ones obtained with the Stanford NER module of CoreNLP before and after the retraining using a novel corpus of manually annotated historical travel writings.

Italiano. *Questo articolo presenta l'applicazione di un'architettura neurale all'identificazione dei nomi propri di luogo all'interno di testi storici in lingua inglese. Abbiamo valutato l'impatto di vari word embedding e confrontato i risultati con quelli ottenuti usando il modulo NER di Stanford CoreNLP prima e dopo averlo riaddestrato usando un nuovo corpus di lettaratura di viaggio storica manualmente annotato.*

1 Introduction

Named Entity Recognition (NER), that is the automatic identification and classification of proper names in texts, is one of the main tasks of Natural Language Processing (NLP), having a long tradition started in 1996 with the first major event dedicated to it, i.e. the Sixth Message Understanding Conference (MUC-6) (Grishman and Sundheim, 1996). In the field of Digital Humanities (DH), NER is considered as one of the important challenges to tackle for the processing of large cultural datasets (Kaplan, 2015). The language variety of historical texts is however greatly different from the one of the contemporary texts NER systems are usually developed to annotate, thus an adaptation of current systems is needed.

In this paper, we focus on the identification of place names, a specific sub-task that in DH is envisaged as the first step towards the complete geoparsing of historical texts, which final aim is to discover and analyse spatial patterns in various fields, from environmental history to literary studies, from historical demography to archaeology (Gregory et al., 2015). More specifically, we propose a neural approach applied to a new manually annotated corpus of historical travel writings. In our experiments we test the performance of different pre-trained word embeddings, including a set of word vectors we created starting from historical texts. Resources employed in the experiments are publicly released together with the model that achieved the best results in our task[1].

2 Related Work

Different domains - such as Chemistry, Biomedicine and Public Administration (Eltyeb and Salim, 2014; Habibi et al., 2017; Passaro et al., 2017) - have dealt with the NER task by developing domain-specific guidelines and automatic systems based on both machine learning and deep learning algorithms (Nadeau and Sekine, 2007; Ma and Hovy, 2016). In the field of Digital Humanities, applications have been proposed for the domains of Literature, History and Cultural Heritage (Borin et al., 2007; Van Hooland et al., 2013; Sprugnoli et al., 2016). In particular, the computational treatment of historical newspapers has received much attention being, at the moment, the most investigated text genre (Jones and Crane, 2006; Neudecker et al., 2014; Mac Kim and Cassidy, 2015; Neudecker, 2016; Rochat et al., 2016).

`Person`, `Organization` and `Location` are the three basic types adopted by general-purpose NER systems, even if different entity types can be detected as well, depending on

[1] https://dh.fbk.eu/technologies/place-names-historical-travel-writings

the guidelines followed for the manual annotation of the training data (Tjong Kim Sang and De Meulder, 2003; Doddington et al., 2004). For example, political, geographical and functional locations can be merged in a unique type or identified by different types: in any case, their detection has assumed a particular importance in the context of the spatial humanities framework, that puts the geographical analysis at the center of humanities research (Bodenhamer, 2012). However, in this domain, the lack of pre-processing tools, linguistic resources, knowledge-bases and gazetteers is considered as a major limitation to the development of NER systems with a good accuracy (Ehrmann et al., 2016).

Compared to previous works, our study focuses on a text genre not much investigated in NLP but of great importance from the historical and cultural point of view: travel writings are indeed a source of information for many research areas and are also the most representative type of intercultural narrative (Burke, 1997; Beaven, 2007). In addition, we face the problem of poor resource coverage by releasing new historical word vectors and testing an architecture that does not require any manual feature selection, and thus neither text pre-processing nor gazetteers.

3 Manual Annotation

We manually annotated a corpus of 100,000 tokens divided in 38 texts taken from a collection of English travel writings (both travel reports and guidebooks) about Italy published in the second half of the XIX century and the '30s of the XX century (Sprugnoli, 2018). The tag `Location` was used to mark all named entities (including nicknames like *city on the seven hill*) referring to:

- geographical locations: landmasses (*Janiculum Hill*, *Vesuvius*), body of waters (*Tiber*, *Mediterranean Sea*), celestial bodies (*Mars*), natural areas (*Campagna Romana*, *Sorrentine Peninsula*);
- political locations: areas defined by socio-political groups, such as cities (*Venice*, *Palermo*), regions (*Tuscany*, *Lazio*), kingdoms (*Regno delle due Sicilie*), nations (*Italy*, *Vatican*);
- functional locations: areas and places that serve a particular purpose, such as facilities (*Hotel Riposo*, *Church of St. Severo*), monuments and archaeological sites (*Forum Romanum*) and streets (*Via dell'Indipendenza*).

The three aforementioned definitions correspond to three entity types of the ACE guidelines, i.e., `GPE` (geo-political entities), `LOC` (locations) and `FAC` (facilities): we extended this latter type to cover material cultural assets, that is the built cultural inheritance made of buildings, sites, monuments that constitute relevant locations in the travel domain.

The annotation required 3 person/days of work and, at the end, 2,228 proper names of locations were identified in the corpus, among which 657 were multi-token (29.5%). The inter-annotator agreement, calculated on a subset of 3,200 tokens, achieved a Cohen's kappa coefficient of 0.93 (Cohen, 1960), in line with previous results on named entities annotation in historical texts (Ehrmann et al., 2016).

The annotation highlighted the presence of specific phenomena characterising place names in historical travel writings. First of all, the same place can be recorded with variations in spelling across different texts but also in the same text: for example, modern names can appear together with the corresponding ancient names (*Trapani gradually assumes the form that gave it its Greek name of Drepanum*) and places can be addressed by using both the English name and the original one, the latter occurring in particular in code-mixing passages (Sprugnoli et al., 2017) such as in: (*Byron himself hated the recollection of his life in Venice, and I am sure no one else need like it. But he is become a cosa di Venezia, and you cannot pass his palace without having it pointed out to you by the gondoliers.*). Second, some names are written with the original Latin alphabet graphemes, such as *Ætna* and *Tropæa Marii*. Then, there are names having a wrong spelling: e.g., *Cammaiore* instead of *Camaiore* and *Momio* instead of *Mommio*. In addition, there are several long multi-token proper names, especially in case of churches and other historical sites, e.g. *House of the Tragic Poet*, *Church of San Pietro in Vincoli*, but also abbreviated names used to anonymise personal addresses, e.g. *Hotel B.*. Travel writings included in the corpus are about cities and regions of throughout Italy thus there is a high diversity in the mentioned locations, from valleys in the Alps (*Val Buona*) to small villages in Sicily (*Capo S. Vito*). However, even if the main topic of the corpus is the descrip-

tion of travels in Italy, there are also references to places outside the country, typically used to make comparisons (*Piedmont, in Italy, is nothing at all like neighbouring Dauphiné or Savoie*).

4 Experiments

Experiments for the automatic identification of place names were carried out using the annotated corpus described in the previous Section. The corpus, in BIO format, was divided in a training, a test and a development set following a 80/10/10 split. For the classification, we tested two approaches: we retrained the NER module of Stanford CoreNLP with our in-domain annotated corpus and we used a BiLSTM implementation evaluating the impact of different word embeddings, including three new historical pre-trained word vectors.

4.1 Retraining of Stanford NER Module

The NER system integrated in Stanford CoreNLP is an implementation of Conditional Random Field (CRF) sequence models (Finkel et al., 2005) trained on a corpus made by several datasets (CONLL, MUC-6, MUC-7, ACE) for a total of more than one million tokens[2]. The model distributed with the CoreNLP distribution is therefore based on contemporary texts, most of them of the news genre but also weblogs, newsgroup messages and broadcast conversations. We evaluated this model (belonging to the 3.8.0 release of CoreNLP) on our test set and then we trained a new CRF model using our training data.

4.2 Neural Approach

We adopted an implementation of BiLSTM-CRF developed from the Ubiquitous Knowledge Processing Lab (Technische Universität Darmstadt)[3]. This architecture exploits casing information, character embeddings and word embeddings; no feature engineering is required (Reimers and Gurevych, 2017a). We chose this implementation because the authors propose recommended hyperparameter configurations for several sequence labelling tasks, including NER, that we took as a reference for our own experiments. More specifically, the setup suggested by Reimers and Gurevych (2017a) for the NER task is summarised below:

- dropout: 0.25, 0.25
- classifier: CRF
- LSTM-Size: 100
- optimizer: NADAM
- word embeddings: GloVe Common Crawl 840B
- character embeddings: CNN
- miniBatchSize: 32

Starting from this configuration, we evaluated the performance of the NER classifier trying different pre-trained word embeddings. Given that the score of a single run is not significant due to the different results producing by different seed values (Reimers and Gurevych, 2017b), we run the system three times and we calculated the average of the test score corresponding to the epoch with the highest result on the development test. We used Keras version 1.0[4] and Theano 1.0.0[5] as backend; we stopped after 10 epochs in case of no improvements on the development set.

4.2.1 Pre-trained Word Embeddings

We tested a set of word vectors available online, all with 300 dimensions, built on corpora of contemporary texts and widely adopted in several NLP tasks, namely: (i) GloVe embeddings, trained on a corpus of 840 billion tokens taken from Common Crawl data (Pennington et al., 2014); (ii) Levy and Goldberg embeddings, produced from the English Wikipedia with a dependency-based approach (Levy and Goldberg, 2014); (iii) fastText embeddings, trained on the English Wikipedia using sub-word information (Bojanowski et al., 2017). By taking into consideration these pre-trained embeddings, we cover different types of word representation: GloVe is based on linear bag-of-words contexts, Levy on dependency parse-trees, and fastText on a bag of character n-grams.

In addition, we employed word vectors we developed using GloVe, fastText and Levy and Goldberg's algorithms on a a subset of the Corpus of Historical American English (COHA) (Davies, 2012) made of more than 198 million words. The chosen subset contains more than 3,800 texts belonging to four genres (i.e., fiction, non-fiction, newspaper, magazine) published in the same temporal span of our corpus of travel writings. These

[2] https://nlp.stanford.edu/software/CRF-NER.html

[3] https://github.com/UKPLab/emnlp2017-bilstm-cnn-crf

[4] https://keras.io/

[5] http://deeplearning.net/software/theano/

historical embeddings, named HistoGlove, HistoFast and HistoLevy, are available online[6].

5 Results and Discussion

Table 1 shows the results of our experiments in terms of precision (P), recall (R) and F-measure (F1): the score obtained with the Stanford NER module before and after the retraining is compared with the one achieved with the deep learning architecture and different pre-trained word embeddings.

The neural approach performs remarkably better than the CFR sequence models with a difference ranging from 11 to 14 points in terms of F1, depending on the word vectors used. The original Stanford module produces much unbalanced results with the lowest recall and F1 but a precision above 82. In all the other experiments, scores are more balanced even if in the majority of the neural experiments recall is slightly higher than precision, meaning that BiLSTM is more able to generalise the observations of named entities from the training data. Although the training data are few, compared to the corpora used for the original Stanford NER module, they produce an improvement of 13.1 and 5.9 points on recall and F1 respectively, demonstrating the positive impact of having in-domain annotated data.

As for word vectors, dependency-based embeddings are not the best word representation for the NER task having the lowest F1 among the experiments with the neural architecture. It is worth noticing that GloVe, suggested as the best word vectors by Reimers and Gurevych (2017a) for the NER task on contemporary texts, does not achieve the best scores on our historical corpus. Linear bag-of-words contexts is however confirmed as the most appropriate word representation for the identification of Named Entities, given that HistoGloVe produces the highest scores for all the three metrics.

The improvement obtained with the neural approach combined with historical word vectors and in-domain training data is evident when looking in details at the results over the three files constituting the test set. These texts were extracted from two travel reports, "A Little Pilgrimage in Italy" (1911) and "Naples Riviera" (1907) and one guidebook, "Rome" (1905). The text taken from the latter book is particularly challenging for the presence of many Latin place names and locations related to the ancient (and even mythological) history of the city of Rome, e.g. *Grotto of Lupercus*, *Alba Longa*. As displayed in Table 2, Neural HistoGloVe increases the F1 score of 9.8 points on the first file, 12.7 on the second and 25.3 on the third.

[6] http://bit.do/esiaS

	P	R	F1
Stanford NER	82.1	66.1	73.2
Retrained Stanford NER	78.9	79.2	79.1
Neural HistoLevy	85.3	83.3	84.3
Neural Levy	83.7	86.8	85.3
Neural HistoFast	83.9	87.4	85.6
Neural GloVe	83.7	87.9	86.0
Neural FastText	86.3	86.3	86.3
Neural HistoGlove	**86.4**	**88.5**	**87.4**

Table 1: Results of the experiments.

	Stanford NER	**Neural HistoGloVe**
	F1	**F1**
Little Pilgrimage	80.9	90.7
Naples Riviera	73.3	86.0
Rome	55.6	80.9

Table 2: Comparison of F1 in the three test files.

6 Conclusions and Future Works

In this paper we presented the application of a neural architecture to the automatic identification of place names in historical texts. We chose to work on an under-investigated text genre, namely travel writings, that presents a set of specific linguistic features making the NER task particularly challenging. The deep learning approach, combined with in-domain training set and in-domain historical embeddings, outperforms the linear CRF classifier of the Stanford NER module without the need of performing feature engineering. Annotated corpus, best model and historical word vectors are all freely available online.

As for future work, we plan to experiment with a finer-grained classification so to distinguish different types of locations. In addition, another aspect worth studying is the georeferencing of identified place names so to map the geographical dimension of travel writings in Italy. An example of visualisation is given in Figure 1 where the locations automatically identified from the test file taken from the book "Naples Riviera" are displayed: place names have been georeferenced us-

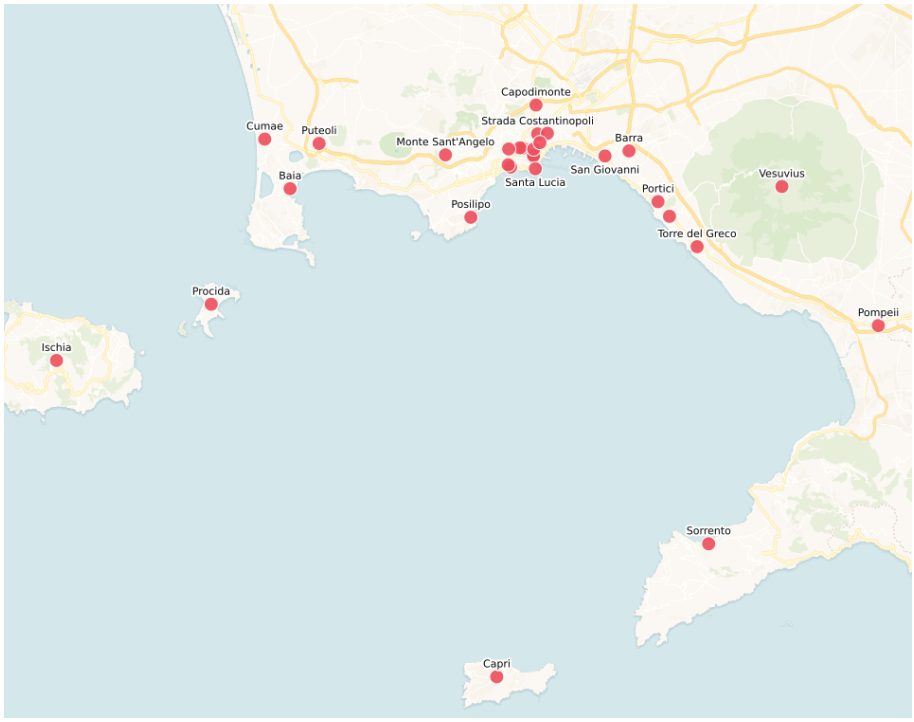

Figure 1: Map of place names in the Neapolitan area mentioned in the "Naples Riviera" test file.

ing the Geocoding API[7] offered by Google and displayed through the Carto[8] web mapping tool. Another interesting work would be the detection of itineraries of past travellers: this application could have a potential impact on the tourism sector, suggesting historical routes alternative to those more beaten and congested and making tourists rediscovering sites long forgotten.

Acknowledgments

The author wants to thank Manuela Speranza for her help with inter-annotator agreement.

References

Tita Beaven. 2007. A life in the sun: Accounts of new lives abroad as intercultural narratives. *Language and Intercultural Communication*, 7(3):188–202.

David J Bodenhamer. 2012. The spatial humanities: space, time and place in the new digital age. In *History in the Digital Age*, pages 35–50. Routledge.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Lars Borin, Dimitrios Kokkinakis, and Leif-Jöran Olsson. 2007. Naming the past: Named entity and animacy recognition in 19th century Swedish literature. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, pages 1–8.

Peter Burke. 1997. *Varieties of cultural history*. Cornell University Press.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, 7(2):121–157.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. 2004. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In *LREC*, volume 2, pages 837–840.

Maud Ehrmann, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. 2016. Diachronic evaluation of NER systems on old newspapers. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, number EPFL-CONF-221391, pages 97–107. Bochumer Linguistische Arbeitsberichte.

Safaa Eltyeb and Naomie Salim. 2014. Chemical named entities recognition: a review on approaches and applications. *Journal of cheminformatics*, 6(1):17.

[7] https://developers.google.com/maps/documentation/geocoding/start

[8] https://carto.com/

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.

Ian Gregory, Christopher Donaldson, Patricia Murrieta-Flores, and Paul Rayson. 2015. Geoparsing, GIS, and textual analysis: Current developments in spatial humanities research. *International Journal of Humanities and Arts Computing*, 9(1):1–14.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, volume 1.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Alison Jones and Gregory Crane. 2006. The challenge of Virginia Banks: an evaluation of named entity analysis in a 19th-century newspaper collection. In *Digital Libraries, 2006. JCDL'06. Proceedings of the 6th ACM/IEEE-CS Joint Conference on*, pages 31–40. IEEE.

Frédéric Kaplan. 2015. A map for big data research in digital humanities. *Frontiers in Digital Humanities*, 2:1.

Omer Levy and Yoav Goldberg. 2014. Dependency-Based Word Embeddings. In *ACL (2)*, pages 302–308.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1064–1074.

Sunghwan Mac Kim and Steve Cassidy. 2015. Finding names in trove: named entity recognition for Australian historical newspapers. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 57–65.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Clemens Neudecker, Lotte Wilms, Wille Jaan Faber, and Theo van Veen. 2014. Large-scale refinement of digital historic newspapers with named entity recognition. In *Proc IFLA Newspapers/GENLOC Pre-Conference Satellite Meeting*.

Clemens Neudecker. 2016. An Open Corpus for Named Entity Recognition in Historic Newspapers. In *LREC*.

Lucia C Passaro, Alessandro Lenci, and Anna Gabbolini. 2017. INFORMed PA: A NER for the Italian Public Administration Domain. In *Fourth Italian Conference on Computational Linguistics CLiC-it 2017*, pages 246–251. Accademia University Press.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2017a. Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. *arXiv preprint arXiv:1707.06799*.

Nils Reimers and Iryna Gurevych. 2017b. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348.

Yannick Rochat, Maud Ehrmann, Vincent Buntinx, Cyril Bornet, and Frédéric Kaplan. 2016. Navigating through 200 years of historical newspapers. In *iPRES 2016*, number EPFL-CONF-218707.

Rachele Sprugnoli, Giovanni Moretti, Sara Tonelli, and Stefano Menini. 2016. Fifty years of European history through the lens of computational linguistics: the De Gasperi Project. *Italian Journal of Computational Linguistics*, pages 89–100.

Rachele Sprugnoli, Sara Tonelli, Giovanni Moretti, and Stefano Menini. 2017. A little bit of bella pianura: Detecting Code-Mixing in Historical English Travel Writing. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*.

Rachele Sprugnoli. 2018. "Two days we have passed with the ancients...": a Digital Resource of Historical Travel Writings on Italy. In *Book of Abstract of AIUCD 2018 Conference*. AIUCD.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Seth Van Hooland, Max De Wilde, Ruben Verborgh, Thomas Steiner, and Rik Van de Walle. 2013. Exploring entity recognition and disambiguation for cultural heritage collections. *Digital Scholarship in the Humanities*, 30(2):262–279.