Sort Term Scientific Mission (STSM)
Scientific Report

Action: COST IS1310
Date of visit: 1 of November 2016 – 30 of November 2016
COST STSM Reference Number: COST-STSM-ECOST-STSM-IS1310-011116-081051
SRSM Title: Towards a methodology for the identification, extraction and analysis of place in the Republic of Letters
STSM applicant: Dr Miguel Won, INESC-ID, Lisbon, Portugal
Host: Dr Patricia Murrieta-Flores (BA, MSc), Director of the Digital Humanities Research Centre, Faculty of Humanities, University of Chester

STSM Scientific Report Contents:

1. Purpose of the STSM
2. Description of the work carried out during the STSM
3. Description of the main results obtained
4. Future collaboration with host institution
5. Foreseen publications/articles to result from the STSM
6. Confirmation by the host institution of the successful execution of the STSM

1. Purpose of the STSM

The main objective of this STSM was to bring Named Entity Recognition (NER) and Geographic Information Retrieval (GIR) techniques to the field of Spatial Humanities in the context of the historical research about the Republic of Letters (RofL). This mission focused in the collection of spatial information cited in the historical texts, as part of geographical research of Digital Humanities field. The identification of geographical places written in a document is traditionally annotated manually or with the use of historical gazetteers. This annotation process is possible for small sets of documents, but almost humanly impossible for bigger corpora, such as the thousands of letters gathered by the Reassembling the Republic of Letters project. Therefore, there is the need to use automatic computational tools programed to perform such task.
For these reasons the mission proposed to explore and evaluate automated methods to identify and retrieve geographic information from historical texts, in particular from RofL datasets. This study is in line of work with the majors WG1 (working group 1) lines of research, and should bring new insights about the course of action for the future in the context of the Reassembling the Republic of Letters project.

2. Description of the work carried out during the STSM

2.1. Data and Pre-Processing details

We have worked with two sets of letters: Mary Hamilton and Samuel Hartlib. Both corpora are written in English with the Hartlib letters mostly in Early Modern English.

The Mary Hamilton set contains a total of 161 letters in XML files annotated in TEI. The annotations contain metadata such as authorship, date, information about the transliteration project, context of the letter according to research carried out, corrections made by the transliterator and annotated words within the body text. This annotation included place and person names, which we have used to evaluate the final results.
In order to create clean text files suitable to be processed by the NER systems, all XML files went trough a cleaning step. This pre-processing consisted in the extraction of the body texts from the XML code, followed by a tokenization and tagging process, where all words and punctuation were labeled with the tags "LOC" for location, and "O" for other. For such labeling process the XML meta-data information was used.
The Samuel Hartlib corpus is composed by 5743 letters, mostly written in Early Modern English (we have not identified all written languages in the full set, but we identified letters written in French, German and Latin).
Unlike Mary Hamilton letters, the Hartlib corpus does not contain annotations that would allow the evaluation of the NER system in the full corpus. For this reason, we have selected a set of 54 letters and manually annotated, within the documents, all mentioned place-names. This set was used to measure the performance of each NER system in respect to its ability to identify the mentioned locations in the Hartlib letters.
The Hartlib corpus is coded in a set of html files that show a faithful representation of the original letters, i.e., they present the original text together with a series of comments related to the editorial process and written by the transcriber. All the comments are written within square brackets and can be simple notes that should not be part of the main text, or suggestions from the transcribers about a word that cannot longer be appreciated in the original manuscript. In this last case, the suggestion should be included in the main text. See for instance the next two examples:

1. [word/s deleted]
2. [another hand?: Mr Williamsons]

The first case is a note stating that there was a deleted word in the original manuscript, while the second example is referring to a case where the manuscript has the name "Mr Williamsons" written in the text, but by a different hand.
For the present work we want the closest form of a meaningful text, without the notes from the transcriber but with the suggestions of what should be incorporated in the final text. For the equivalent cases to the first example, a simple deletion from the original html file can be performed, because it is just a pure transcriber comment. But for the second example "Mr Williamsons" must be kept.
We were not able to find a universal pattern for all the comments that would allow the implementation of an automatic cleaning process. Therefore, in order to study the NER task performance dependency with these comments, two cleaning processes were defined: *full* and *fast clean*. We have studied the impact of each cleaning process in the same set of 54 annotated letters mentioned above.

In the *full clean* we first manually identified common patterns and the exceptions rules in the transcriber's notes in the 54 selected letters. Based in these findings, all square brackets were then correctly removed and replaced by the appropriate text. This cleaning process creates the closest text form in respect to the original. However its creation was not done automatically.
For *the fast clean* method all square brackets and its content were completely removed automatically. Therefore, in this version, the final texts have missing words, in particular the suggestions written form the transcriber. The advantage of this pre-processing, is that it can be automated and applied to the full set.
We have additionally applied standard cleaning tasks to both sets: html code, page brakes and non-alphanumeric characters, except punctuation, were removed. Also, and equivalently to Mary Hamilton set, tokenization and tagging with "LOC" and "O" tags was applied.
In addition to the possible cleaning process dependency, the NER task is also language dependent. Therefore, in order to study the potential impact of language stage difference, we have translated documents from the original Early Modern English (EME) to Modern English (ME) and parsed both language sages for results comparison.

All scripts used to pre-process and parse the analyzed texts were written in python 2.7, with special use of packages NLTK [1].The translation from EME and ME was performed with MorphAdorner tool [2].

2.2 Place-names identification

The automatic identification of locations mentioned in the letters was performed through the use of Natural Language Processing (NLP) tools developed from Name Entity Recognition (NER) field.
NER is a subtask of the NLP field and its main objective as a task is to identify named entities, such as person names or locations in a given text. For the present work we are only interested in the identification of all geographic locations mentioned in the historical letters.

There are today many free available NER tools, coded for different computer and natural languages. Equivalently to other tasks in NLP, NER tools are built through the use of supervised or unsupervised methods borrowed from the field of Machine Learning (ML). The first method relies in statistical pre-trained models fitted from annotated corpora that allow a "learning" process to take place; the second method tries to directly solve the task without requiring pre-training. Supervised methods tend to give better results but need annotated corpora, usually by humans, that usually are scarce. Up today the best NER tools based in supervised learning methods outperformed tools based in unsupervised methods, and for this reason we focus, in the current work, in NER tools that were trained by supervised methods only.
Mathematically, the "learning" process is no more than to find the best fit parameters of the underlying model used by the NER system, and therefore the main distinction between each tool is the mathematical model that it uses, as

well the parameters it receives as input. For this reason we avoid NER tools constructed with very similar mathematical models.
We have chosen five different NER tools to identify the place-names:

- Tagger [3]
- Standford NER [4]
- Edinburgh Geoparser [5]
- Spacy [6]
- Polyglot [7]

The selection followed the criteria of simplicity of its usability, good performance in standard corpora like CONLL 2003 [8] and model diversity. Parsing the texts with multiple NER tools that were trained with different statistical models allows the possibility to consider the different results as independent. In order to explore statistical independency, we have additionally implemented a voting system that combines all NER system results. Also, in order to consider information already collected, we have included a gazetteer built from the EMLO database [9].

The implemented voting system works as follows:

1. A vote per system is given for each location tagged by any of the NER systems; E.g., if the word "London", in position n, is tagged as a location by the Edinburgh Geoparser and Spacy but not by any other NER tool, it receives two votes;
2. After all votes from NER tools are distributed, a query for the location in the EMLO gazetteer is made and if there is an entry for the location an additional vote is given;
3. The candidate is tagged as a *true* location if it collected the minimum defined votes.

The NER systems performance is evaluated in terms of precision, recall and F-measure. The precision result should be interpreted as "percentage of correct entities in respect to all entities tagged as a location by the NER system". Therefore, if only one entity is tagged, and that entity is in fact a *true* location, we would then be facing a precision of 100%. For the recall, the reading should be: "percentage of corrected tagged entities in respect to the total absolute true number of entities". If 10 entities are correctly identified as a location but in fact there are 100 *true* locations in the texts, it results in a recall of 10%. Finally, F-measure is a combination of these two parameters and intends to give final balance between precision and recall. It is expressed by:

$$F = 2\frac{precision \times recall}{precision+recall}.$$

We note that in order to calculate these metrics, the position within the texts of the identified place-name was taken into account. Therefore, if e.g. "London" location is mentioned twice in one document, the NER system should identify both citations to "London", as well their respective position within the text.

2.3 Geographic Information

During the work done under this STSM no research has been done about the extraction of the geographic information, such as latitude and longitude. This information was extracted, whenever possible, from the Geonames database. For every location identified from the NER systems explained above, a query to the Geonames database was requested, and saved in case a positive response. However, no study has been performed to evaluate the correctness of such responses, namely, if the geographic locations obtained correspond to the physical location mentioned in the text. These queries were made using *python* module *geopy* [10].

3. Description of the main results obtained

Tables 1-4 show the results when considering individually each NER system, and when combining them with the voting system explained above. For this last case we have set 5 different minima voting thresholds, ranging from 2 to 5 votes. Table 1 shows the results for Mary Hamilton letters and Tables 2, 3 and 4 for the selected Samuel Hartlib letters. Table 2 shows the evaluation when these letters were pre-processed by the *full clean* rules followed by a translation from the original text to Modern English. Table 3 shows the equivalent results for the documents that have been *fully cleaned* but not translated to ME. Finally, in table 4 we show the results when only *fast clean* pre-processed is applied. For all tables we have signaled (in bold) the best F-Measure result.

Hamilton			
Model	Precision	Recall	F-Measure
tagger	43.8	58.3	50
Stanford NER	55.9	63.1	59.3
Geoparser	53.9	71.6	61.4
Spacy	66.7	56.6	61.2
Polyglot	60.7	64.1	62.3
Comb+EMLO 1 vote	34.9	84.6	49.4
Comb+EMLO 2 votes	58.9	79.9	67.8
Comb+EMLO 3 votes	**74.9**	**68.5**	**71.6**
Comb+EMLO 4 votes	84.2	57.7	68.5
Comb+EMLO 5 votes	90.3	46.7	61.6

Table 1: Precision, Recall and F-Measure results for Mary Hamilton letters.

Hartlib (full clean and ME)			
Model	Precision	Recall	F-Measure
tagger	53.8	51.8	52.8
Stanford NER	73.6	67.1	70.2

Geoparser	67	50.6	57.6
Spacy	61.8	54.9	58.1
Polyglot	65.7	57.1	61.1
Comb+EMLO 1 vote	43.6	78.8	56.2
Comb+EMLO 2 votes	**72.1**	**76**	**74**
Comb+EMLO 3 votes	81.2	60.9	69.6
Comb+EMLO 4 votes	87.8	48.9	62.8
Comb+EMLO 5 votes	92.4	36.5	52.4

Table 2: Precision, Recall and F-Measure results for the 50 selected letters of Samuel Hartlib corpus, with full clean pre-processing and translated from the original to Modern English.

Hartlib (full clean and EME)			
Model	Precision	Recall	F-Measure
tagger	53.2	52.8	53.3
Stanford NER	73.1	68	70.4
Geoparser	53.4	51.8	52.4
Spacy	58.9	56.3	57.5
Polyglot	52.7	58.9	55.6
Comb+EMLO 1 vote	35.4	81.1	49.3
Comb+EMLO 2 votes	**69.3**	**78**	**73.4**
Comb+EMLO 3 votes	80.5	61.4	69.7
Comb+EMLO 4 votes	87.7	49.4	63.2
Comb+EMLO 5 votes	93.4	37.3	53.3

Table 3: Precision, Recall and F-Measure results for the 50 selected letters of Samuel Hartlib corpus, with full clean pre-processing only.

Hartlib (fast clean and EME)			
Model	Precision	Recall	F-Measure
tagger	53.6	53	53.3
Stanford NER	73.2	68.2	70.6
Geoparser	53.5	52	52.7
Spacy	58.8	56.3	57.6
Polyglot	52.8	59.2	55.8
Comb+EMLO 1 vote	41.3	80.6	54.6
Comb+EMLO 2 votes	**69.3**	**78.3**	**73.5**
Comb+EMLO 3 votes	81.1	61.7	70.1
Comb+EMLO 4 votes	87.3	49.6	63.3
Comb+EMLO 5 votes	93	37.5	53.4

Table 4: Precision, Recall and F-Measure results for the 50 selected letters of Samuel Hartlib corpus, with fast clean pre-processing only.

From the shown results we can see that the combination system with a minimum of 2 to 4 votes was able to constantly outperform the individual NER systems. Also, all experiments resulted in a best minimum F-Measure of 70, which gives consistency to the analysis.

The best F-Measure result was obtained for the set of Hartlib Letters when *full clean* pre-processing and translation to Modern English was performed. However, we didn't observe a significant difference between the *fast clean* set and between EME and ME. In fact, the gain in F-Measure between *fast clean* pre-processing and *full clean* pre-processing together with EME to ME translation, is only from 73.5 to 74.
The F-Measure results for the Hartlib set are higher for all scenarios (Table 2-4) in respect to the Mary Hamilton set. This result was surprising, because the original files from Hamilton set were cleaner and written in a closer form of Modern English.
In respect to the individual NER systems, Standford NER gave the best F-Measure for all Hartlib scenarios. Its performance is constantly close to the voting system, which allows one to infer that a combinatorial system of Standford NER and EMLO gazetteers should give similar results. We have not studied possible individual combinations.

4. Future collaboration with host institution

The present work opened a path for new possible collaborations, with special focus to continue to develop tools from the NLP field to study and analyze historical texts from the Republic of Letters. From the results obtained in this mission, we concluded that there is scope for improvement in respect to place-names identification.

Also, the geographic identification was not studied in deep detail, and there is the need to explore methods developed by the field of Geographic Information Retrieval on the subject of place reference resolution with Deep Neural Networks, which could be one of the next tasks to research.
As for now a meeting in February 2017, with the people of interest in the field, will take place in the Digital Humanities Research Centre of the University of Chester, where the results from the present work and possible future collaborations will be discussed.

5. Foreseen publications/articles to result from the STSM

Won, M., Murrieta-Flores, P., and Martins, B. (2017-in preparation). Towards a methodology for the identification and geocoding of place-names in historical documents. To be submitted to *Geohumanities* or *Digital Humanities Quarterly*.

6. Confirmation by the host institution of the successful execution of the STSM

The STSM was accomplished with very satisfactory results. This work enabled us to identify the next lines of research that WG1 need to pursue and allowed the identification of further challenges to be addressed in relation to 'space and place' in the RofL project.

[1] NLTK Project, “Natural Language Toolkit — NLTK 3.0 documentation,” 2016. [Online]. Available: http://www.nltk.org/. [Accessed: 22-Dec-2016].
[2] P. R. Burns, “MorphAdorner v2: A Java Library for the Morphological Adornment of English Language Texts.,” Evanston, IL., 2013.
[3] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, “Neural Architectures for Named Entity Recognition,” *CoRR*, vol. abs/1603.0, 2016.
[4] J. R. Finkel, T. Grenager, and C. Manning, “Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling,” in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005, pp. 363–370.
[5] B. Alex, K. Byrne, C. Grover, and R. Tobin, “Adapting the Edinburgh Geoparser for Historical Georeferencing,” *Int. J. Humanit. Arts Comput.*, vol. 9, no. 1, pp. 15–35, Mar. 2015.
[6] Explosion AI, “spaCy - Industrial-strength Natural Language Processing in Python.” [Online]. Available: https://spacy.io/. [Accessed: 22-Dec-2016].
[7] R. Al-Rfou, B. Perozzi, and S. Skiena, “Polyglot: Distributed word representations for multilingual nlp,” *arXiv Prepr. arXiv1307.1662*, 2013.
[8] E. F. Tjong Kim Sang and F. De Meulder, “Introduction to the CoNLL-2003 shared task,” in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* -, 2003, vol. 4, pp. 142–147.
[9] Cultures of Knowledge, “Early Modern Letters Online,” 2009. [Online]. Available: http://emlo.bodleian.ox.ac.uk. [Accessed: 22-Dec-2016].
[10] “GeoPy 1.10.0.” [Online]. Available: https://geopy.readthedocs.io/en/1.10.0/. [Accessed: 22-Dec-2016].