**Samtla-Char-NER Report**

**Implementation of Character-based Named Entity Recognition into the Samtla System**

*Student: Matthew Ralph,*

*MSc Computer Science project report, Department of Computer Science and Information, Birkbeck College University of London*

[*mralph02@dcs.bbk.ac.uk*](mailto:mralph02@dcs.bbk.ac.uk)

*Supervisor: Dr Dell Zhang*

*With thanks to Dr Martyn Harris*

*This report is substantially the result of my own work, expressed in my own words, except where explicitly indicated in the text. I give my permission for it to be submitted to the JISC Plagiarism Detection Service. The report may be freely copied and distributed provided the source is explicitly acknowledged.*

# Abstract

**Update at end.** Recent approaches to Named Entity Recognition (NER), such as that of (Kuru, Arkan Can and Deniz, 2016), demonstrate that a character-level representation of textual data can yield good results when training a deep learning model. In this project, a set of Hansard debates is aggregated, processed and labelled for use in a Bidirectional Long Short-Term Memory (BLSTM) neural network. The trained model, and the original dataset, is integrated with Birkbeck’s Samtla digital humanities text archiving system, such that the Hansard texts can be browsed in the interface, and previously unseen Named Entities are highlighted.

# Acknowledgements

I would like to express my gratitude to the people who taught me to program in Python by working on real problems, particularly to Ali Lotia and Ogonna Iwunze, whose expertise is matched only by their patience and compassion. I would also like to thank Sergio Gutierrez-Santos, whose instruction in the Java programming language was carefully designed and helped to open up a world of structured code for me, as well as demystifying unit testing.

I am grateful to Dr Martyn Harris for his help and encouragement when exploring this project and its potential integration with Samtla, and to Dr Dell Zhang for his ideas, advice on the academic landscape surrounding Named Entity Recognition, and quick responses to my queries. Petar Konovski’s quick assistance in setting up a server in Birkbeck with library dependencies greatly simplified the process of moving my code onto a suitable sized server, and I am grateful to Systems Group for the use of such a beefy machine.

Finally, I would like to thank my wife for all her help throughout this Master’s programme, while she worked on her own master’s and continued to support so many people.

*Ad Maiorem Dei gloriam*, from whom all language flows. Ps 19v14.

# List of figures and tables

Compile at end

Table of Contents

[0](#_Toc521427563)

[1. Abstract 1](#_Toc521427564)

[2. Acknowledgements 1](#_Toc521427565)

[3. List of figures and tables 1](#_Toc521427566)

[5. Introduction (including background) 3](#_Toc521427567)

[6. Overall Results (trailer) 4](#_Toc521427568)

[7. Software Architecture 4](#_Toc521427569)

[7.1. The Pipeline of tasks 4](#_Toc521427570)

[7.2. Invoke 4](#_Toc521427571)

[7.3. Named Entity Downloading 8](#_Toc521427572)

[7.4. Raw Hansard downloading 9](#_Toc521427573)

[7.5. Hansard processing 9](#_Toc521427574)

[7.6. Hansard chunking 10](#_Toc521427575)

[7.7. Hansard interpolation 10](#_Toc521427576)

[7.8. Partition into datasets and sizes 10](#_Toc521427577)

[7.9. Formation of Tensors 10](#_Toc521427578)

[7.10. Overview of files in project and what they do 10](#_Toc521427579)

[8. Implementation issues 10](#_Toc521427580)

[8.1. Wikipedia data cleanliness 10](#_Toc521427581)

[8.2. TWFY API suspect return values 10](#_Toc521427582)

[8.3. NLTK span\_tokenize bugs 10](#_Toc521427583)

[8.4. Toy dataset model – tensor sparsity 10](#_Toc521427584)

[8.5. Hansard Presentation issues 10](#_Toc521427585)

[9. Testing 10](#_Toc521427586)

[9.1. Unit testing 10](#_Toc521427587)

[9.2. Manual evaluation 10](#_Toc521427588)

[9.2.1. The ‘mini’ dataset 10](#_Toc521427589)

[9.2.2. The ‘toy’ dataset 10](#_Toc521427590)

[9.3. Model cross-validation 10](#_Toc521427591)

[9.4. Overall evaluation 11](#_Toc521427592)

[10. Summary and Conclusions 11](#_Toc521427593)

[10.1. Pre-processing is hard 11](#_Toc521427594)

[10.2. Labelling is hard 11](#_Toc521427595)

[10.3. Sentence tokenization is hard 11](#_Toc521427596)

[11. References 11](#_Toc521427597)

[12. User Manual 11](#_Toc521427598)

[13. Appendix: Code 11](#_Toc521427599)

[14. What’s My Work 11](#_Toc521427600)

# Introduction (including background)

The brief for this project was to demonstrate Named Entity Recognition, using the approach cited in (Kuru, Arkan Can and Deniz, 2016), and the Keras implementation of this provided by GitHub user 0xnurl.[[1]](#footnote-1) The target dataset was the Hansard, the record of debates in both of the houses of Parliament in the United Kingdom.[[2]](#footnote-2) This dataset is now available via the Parliament UK Data API,[[3]](#footnote-3) however this API is largely undocumented and was not available at the start of this project. Instead, I used the They Work For You API,[[4]](#footnote-4) which has all debates from 1919 onwards available for download in a parsed XML format annotated with metadata about the speaker. I did not have time to use this high-quality metadata during the project and found a few issues with the API (detailed in section 8.2). However, I am grateful for free use of this API which certainly made data preparation easier for me.

In order to implement the model, I had to produce labelled Hansard data. To manually label a few thousand debate documents, required to train even a very basic model, would have been too time-consuming for a project of a few months. So, I used a form of automatic labelling I refer to henceforth as ‘interpolation’, the algorithm for which is explained in section 7.5. Interpolation relied on me having a very large set of Named Entities in my chosen categories or locations, organizations and people. I used the DBPedia SPARQL endpoint[[5]](#footnote-5) and Python’s excellent SPARQLWrapper library[[6]](#footnote-6) to download all the Named Entities on Wikipedia in these categories. There were some data cleanliness issues with this dataset that I never fully overcame, which are detailed in section 8.1.

I used the interpolated (labelled) Hansards to generate a Y tensor. The processed Hansard debates themselves were segmented (or, as I refer to it below, “chunked”) into sentences, and then each character was converted to a number, to create the X tensor. I then used the 0xnurl implementation to train the BLSTM model. An overview of results is given in section 6.

I chose this project because of my interest in linguistics and in humanities texts. In my first degree, Classics, I was fortunate to study linguistic change from Classical Greek to *koine*, the language of the New Testament. I also studied some phenomena of Latin that are markers of a particular gender or class. This project is a tiny step, greatly helped by the labours of TheyWorkForYou, towards mining the value of the Hansard records.

This project is also in part politically motivated. It is vitally important that democratic citizens re-engage with the task of using factual analysis and solid statistics to make important decisions, rather than being emotionally stirred by the language of tyrants. This is a far older problem than the Romans and Ancient Greeks. Learning which companies, places and people we spend most energy talking about as a democracy seems to me, in its own small way, a part of that enormous and essential task.

# Overall Results (trailer)

Do at end

# Software Architecture

## The Pipeline of tasks

This project is, in essence, a data pipeline. Data is sourced from Hansard debates and from Named Entities, combined using a variety of algorithms, and then stored in a format that can then be used to predict unseen named entities. As such, it is best visualised using a pipeline flow (see Figure 1). Each element of the pipeline is introduced in more detail in the sections below, along with details of the algorithms and data storage mechanisms used. Implementation difficulties are discussed in section 8.

The outputs of each pipeline step were persisted to disk, either as simple text files, Python pickle objects in binary format, or Keras’ on H5 binary output format. Such persisting is essential when working with a large amount of data, firstly to allow each stage of the output to be validated and checked, and also to ensure the whole pipeline would not need to be run (which takes several days) every time a bug is discovered. As most of the functions in the pipeline do not return pure values, but write their results out to disk (using ‘print’ statements only to inform the user of their progress), a unit testing approach was needed that could fake a UNIX filesystem in order to validate the functions were working as expected. This is explained in section 10.1.

Were this system to be ‘productionised’, then all the stages of this flow would be run through a Continuous Integration system such as Jenkins or GoCD which would run the different stages in the correct sequence; in such an arrangement, new Named Entity data and newly produced Hansard records could be fed into the pipeline and used to continuously update the model to account for new data.

## Invoke

As so much of this project’s effort was in collecting data for pre-processing, a command-line driven front-end was preferred over building a Graphical User Interface (GUI) just for the internal tasks of gathering, processing and aggregating data. Given the pipeline structure of the project, it was essential to have a tool that would allow code execution to start at any point in the pipeline, with all the correct dependencies in place, having run any prerequisite tasks required. Invoke[[7]](#footnote-7) was chosen, after some experimentation with Argh, Shovel and Doit.

Invoke was found to support arbitrary library imports from the Python global library and the current project, whereas Doit manipulated the user’s PYTHONPATH and so could not be integrated with a project structured into modules. Invoke also supports the basic features for which one might use a Makefile – a simple command line front-end providing many possible entrypoints into an application, with listed prerequisite tasks which could be called with specified, or default, arguments. In contrast to using Make, the task file itself (tasks.py in the code listing, given in section 13) is in pure python and does not require tab characters for delineation. Most calls were simply Python library imports and function executions, but some required separate command-line invocations e.g. to start PyTest or MyPy (for unit testing and static type analysis, respectively), which Invoke natively supports much more elegantly than Argh or Shovel. A full list of Invoke tasks and their descriptions is found in section 15.



Figure 1 pipeline data processing model

## Named Entity Downloading

**Step 1.** Firstly, named entities must be accrued. This is a prerequisite for any automated labelling approach. For locations, the CONLL2003 English dataset was used, together with DBPedia resources of type ‘dbo:Place’. For Organizations, the Amex, Nasdaq and NYSE Stock Exchange company listings were downloaded in Comma Separated Value (CSV) format, as was similar data from the London Stock Exchange, the CONLL2003 English dataset, and DBPedia’s ‘dbo:Organisation’ type. For people, the CONLL2003 English dataset and DBPedia ‘dbo:Person’ type were used, and the New York City Most Popular Baby Names data from Kaggle.[[8]](#footnote-8) Biography-center.com, which was suggested as a naming source by (Klein *et al.*, 2003), no longer has lists of names in an easily-parseable format. However, I suspect that the amount of data in DBPedia has hugely increased since 2003. As Table 2 shows, the size of the DBPedia datasets dwarf the other datasets for all three Named Entity (NE) types.

Table 2 % of NE data from DBPedia

|  |  |
| --- | --- |
| **Dataset** | **% from DBPedia** |
| Locations | 99.8 |
| Organizations | 96.8 |
| People | 99.7 |

**Step 2.** The resulting data had to be cleaned to remove stopwords and some of the more obvious junk data. The data quality issues with the NE datasets are discussed in section 8.1.

**Step 3.** Simple UNIX utilities ‘cat’ and ‘sort’ were used to deduplicate the aggregated NE lists, and sort them into a large text file for each NE type.

## Raw Hansard downloading

**Step 4.** To download the Hansards in a programmatic manner, the TheyWorkForYou API was chosen. This was on the basis of its high quality documentation, and the availability of Hansard debates in XML format, with enriched metadata tags (created using a Perl parser which ran over the source PDF Hansard documents originally available on the UK Parliament website) naming each speaker and detailing their constituency and party. Unfortunately, there was not time in this project to make use of this extra metadata.

Python’s concurrent.futures.ThreadPoolExecutor implementation was chosen to increase the speed of downloads as this activity is mainly bound by network I/O. Only function invocations, with their required parameters, needed to be provided in order for Python’s concurrent library modules to parallelise the downloads – no manual thread handling code is needed.

## Hansard processing

The files downloaded from TheyWorkForYou are XML files with a lot of markup and metadata which would distract from the Named-Entity-learning task. After failed attempts with bleach.clean,[[9]](#footnote-9) the lxml library’s etree module[[10]](#footnote-10) was successfully used to remove all markup and preserve just the text of the debates. In order for lxml to accept the XML files and process them, the encoding of the files and the lxml library’s config had to be set to use UTF-8. Hansard debates use a wide range of characters, including accented letters like é as well as abbreviations like ¾, so it makes sense to pick the most widely-used Unicode encoding standard.

## Hansard chunking

Filesystem issues.

## Hansard interpolation

Interpolation algorithm.

## Partition into datasets and sizes

Bucketing.

## Formation of Tensors

Choosing a good sentence size - medians

## Overview of files in project and what they do

# Implementation issues

## Wikipedia data cleanliness

## TWFY API suspect return values

## NLTK span\_tokenize bugs

## Toy dataset model – tensor sparsity

## Hansard Presentation issues

E.g. No speaker information due to XML processing

# Planning

Lack of detail in the plan. Steps missing from plan, actual timeline used.

# Testing

## Unit testing

Pyfakefs

## Manual evaluation

### The ‘mini’ dataset

History



## The ‘toy’ dataset

History

## The full dataset

## Model cross-validation

The limitations of the labelled data

Sensible baseline: assume everything is NULL.

## Overall evaluation

Something in context of the GUI

# Summary and Conclusions

## Pre-processing is hard

## Labelling is hard

Interpolation algorithm won’t abandon one interpolation for a better one – first-found bias, location bias

## Sentence tokenization is hard

Taught specific abbreviations to the tokenizer. Still bugs outstanding.

# References

# User Manual

# Appendix A: Code

# Appendix B: List of Invoke tasks

Table 1 comprises a list of Invoke tasks which could be started within the project, along with a description of the work that they did. The tasks are a combination of environment setup, running automated tests, and the ‘business logic’ of the code – the downloading and processing of Named Entities and Hansard debates. Having one clean, uniform interface for all these tasks greatly simplified the workflow when parts of the pipeline had to be re-run, without the overhead of creating a GUI or integrating with one. Also, having pre-requisites in code avoids the need for repetition. For example, the unit tests are not run unless the virtual environment is set up, and the static type-checker has run already. These tasks are both listed as pre-requisites of the ‘test’ task in Invoke.

Table 1 List of Invoke tasks used to drive the pipeline

|  |  |
| --- | --- |
| **Task** | **Description** |
| char-ner-create-x | Create an X tensor of Numpy arrays from numerified Hansard data |
| char-ner-create-y | Create a Y tensor of Numpy arrays from onehot vectors from interpolated (labelled) Hansard debates |
| char-ner-display-median-sentence-length | Get the median sentence length of a given dataset |
| char-ner-display-pickled-alphabet | Display the CharBasedNERAlphabet object pickled to disk by char-ner-pickle-alphabet |
| char-ner-pickle-alphabet | Use a small subset of the Hansard debates data to union together all characters used and create a CharBasedNERAlphabet object with a number-to-character mapping |
| char-ner-rehash-datasets | Rehash all the debate data into a different number of buckets – discussed in section 7.7. |
| compile | Run py\_compile on all python files to find compile-time static code problems |
| enable-venv | Enable the Virtual Environment (i.e. a segregated location for pip installs) for this project. A required prerequisite for several other tasks. |
| hansard-chunk-all | Use the chunker on all Hansard debates in the collection – described in section 7.6 |
| hansard-chunk-one | Use the chunked on one Hansard debate, to allow manual validation |
| hansard-display-chunked | Display one Hansard debate, tagged with all its sentence-boundaries, to validate the sentence chunking algorithm. |
| hansard-display-interpolated-file | Display one Hansard debate, with every character tagged by the interpolator as 0 (null), 1 (location), 2 (organization) or 3 (person) |
| hansard-download-all | Concurrently download all Hansard debates from both houses, from a given starting date. |
| hansard-fix-uninterpolated | Find all Hansard debates which did not correctly interpolate due to an NLTK bug with span\_tokenize, and re-interpolate |
| hansard-interpolate-all | Interpolate (label) all Hansard debates using Named Entity data |
| hansard-interpolate-one | Interpolate (label) one Hansard debate for manual validation |
| hansard-numerify-one-to-file | Numerify one Hansard debate – i.e. replace each of its characters with the equivalent integer for this CharBasedNERAlphabet, and store in a file for manual validation |
| hansard-process-all | Do pre-processing steps on all Hansards – discussed in more detail in section 7.5. |
| hansard-process-one | Do pre-processing steps on one Hansard for manual validation. |
| hansard-write-total-number-of-sentences-to-file | Count how many sentences there are in each dataset, and write out to file for easy retrieval. This information is used to estimate to the user how long the tensor creation will take. |
| model-minify-toy | Take the toy dataset and truncate the 1st dimension of all the X and Y tensors to the first 4000 samples, to create a mini dataset |
| model-train-mini | Run model.fit() on the Keras BLSTM model, with a batch of the first 4000 samples from the toy dataset. This is to test the end-to-end process of saving and retrieving the model. |
| model-train-toy | Run model.fit() on the Keras BLSTM model, with a toy dataset of 1 320th of the Hansard debates. This is to get an initial indication of the model’s learning capability. |
| ne-data-companies-download-process | Both download and process companies data from DBPedia and other sources |
| ne-data-companies-process | Only do post-processing, data cleansing tasks on companies data, to assist with iteratively improving the cleaning algorithm |
| ne-data-people-download-process | Both download and process people data from DBPedia and other sources |
| ne-data-people-process | Only do post-processing, data cleansing tasks on people data, to assist with iteratively improving the cleaning algorithm |
| ne-data-places-download-process | Both download and process people data from DBPedia and other sources |
| ne-data-places-process | Only do post-processing, data cleansing tasks on places data, to assist with iteratively improving the cleaning algorithm |
| print-debate-titles | Both download and process people data from DBPedia and other sources |
| python-type-check | Run mypy, Python’s static type checker, over all files I wrote in the project which contain type annotations |
| test | First run python-type-check, then run pytest unit tests on the project |

Each task in invoke called out to something else; most tasks invoked a library function from elsewhere in the code base, while some invoked shell commands, for example to de-duplicate and sort the Named Entity lists. In this case, it is faster for the shell to call a GNU C binary and use the ‘sort’ and ‘uniq’ commands, than to use similar functionality in Python. Using the tasks.py file as a dispatcher, without it containing any processing logic itself, ensured that it remained easy to understand and reason with as the project grew.

Once the project reached a scale where the model processing had to be done on cloud compute nodes, the use of a virtualenv also proved worthwhile. By use of a pip freeze file (requirements-freeze.txt in the project), the DigitalOcean Droplet virtual machine used could be configured with exactly the same python libraries and python interpreter, even though the system python was different from the laptop used for development work.

# What’s My Work

1. <https://github.com/0xnurl/keras_character_based_ner> [↑](#footnote-ref-1)
2. <https://hansard.parliament.uk/> [↑](#footnote-ref-2)
3. <http://www.data.parliament.uk/dataset/12> and <http://api.data.parliament.uk/> [↑](#footnote-ref-3)
4. <https://www.theyworkforyou.com/api/> [↑](#footnote-ref-4)
5. <https://dbpedia.org/sparql> [↑](#footnote-ref-5)
6. <https://rdflib.github.io/sparqlwrapper/> [↑](#footnote-ref-6)
7. http://www.pyinvoke.org/ [↑](#footnote-ref-7)
8. <https://www.kaggle.com/new-york-city/nyc-baby-names> [↑](#footnote-ref-8)
9. <https://github.com/mozilla/bleach> [↑](#footnote-ref-9)
10. <https://lxml.de/api/lxml.etree-module.html> [↑](#footnote-ref-10)