Recognising Geographical Entities in Scottish Historical Documents

Malvina Nissim
School of Informatics
University of Edinburgh
2 Buccleuch Place
EH8 9LW, UK
mnissim@inf.ed.ac.uk

Colin Matheson
School of Informatics
University of Edinburgh
2 Buccleuch Place
EH8 9LW, UK
colin@inf.ed.ac.uk

James Reid
Edinburgh University Data Library
Main Library Building
EH8 9LJ, UK
james.reid@ed.ac.uk

Keywords
Named Entity Recognition, text tagging, georeferencing

1. INTRODUCTION
The need to explicitly georeference (and thus make them inherently geographically searchable) large resource collections such as the Statistical Accounts of Scotland (SAS) which currently only contain implicit georeferences in the form of placennames is becoming more and more urgent. The crucial obvious precondition for successful georeferencing is the recognition of placename occurrences in text [3]. State-of-the-art machine learning systems for the recognition of geographical entities in newswire text achieve an f-score of over 90% [5]. We describe here an experiment with using an off-the-shelf maximum entropy tagger for recognising location names in the SAS, and show that we achieve similar performances. Section 2 briefly describes the data and its annotation, and in Section 3 we present the results of the experiment together with some brief discussion of problems and a comparison with state-of-the-art results. In Section 4 we discuss some future directions of work.

2. DATA AND ANNOTATION
Largely based on information supplied by each parish church minister, the Statistical Accounts of Scotland, covering the 1790s and the 1830s, are among the best contemporary reports of life during the agricultural and industrial revolutions in Europe.

Our dataset comprises two sets of descriptions, one concerning the parish of Edinburgh, the other concerning the parish of Dumfries, for a total of 648 documents. The original plain text files were converted into XML. Two annotators manually marked up all occurrences of locations in the whole dataset, using a customised version of the NITE XML toolkit [1]. Inter-annotator agreement was measured on 50% of the dataset. The headline precision and recall figure was that the annotators agreed 85.82% of the time on what counted as a location. Most of the disagreement was over things like the inclusion of articles with particular classes of location, and so on. Divergences between annotators were discussed and resolved in order to produce a gold-standard. The final dataset comprises 10868 sentences. Overall, there are 5692 instances of locations, for a total of 2357 different names.

3. EXPERIMENT
3.1 Tagging
We trained and tested the Curran and Clark (C&C) maximum entropy tagger [2] by using the in-built C&C standard features. These consist of a set of morphological and orthographical features, as well as information about the word itself, its part-of-speech tag, Named Entity tag history (with a window size of 2), and contextual features. We tokenised and POS-tagged the data by using the Edinburgh Language Technology Group tools (TTT, `http://www.ltg.ed.ac.uk/software/index.html`). The training and evaluation was performed by 10-fold cross-validation.

3.2 Results and Discussion
We evaluated the results by *precision* (the number of correct assignments out of all the tags assigned), *recall* (the number of retrieved entities out of all the entities in the gold-standard), and *f-score* (combined precision and recall). Table 1 shows the tagger's performance in terms of these measures.

Table 1: Results

PRECISION	RECALL	F-SCORE
93.60%	94.87%	94.23%

The results are satisfactory, and definitely comparable with state-of-the-art performances on newswire data. The latest general NER competition which included the recognition of location names was the shared task of the 2003 Conference on Computational Natural Language Learning (CoNLL-2003). Each competing team was to implement a language-independent system for the recognition and classification of four types of entities (locations, persons, organisations, and miscellaneous names) in English and German newswire texts. Sixteen systems entered the competition,

and the highest f-score on the English test data, 91.15%, was achieved by [4], who used a classifier-combination framework. On the devtest data, they obtained an f-score of 96.12%, and showed an error rate reduction of ca. 20% when integrating gazetteers in their system. The C&C tagger achieved a precision, recall, and f-score of 91.75%, 93.20%, and 92.47% on the devtest data, and of 84.97%, 90.53%, and 87.66% on the test data.

Compared to the scores just mentioned, our results are therefore very good. However, there are two caveats to consider. First, we perform a binary classification into location non-location, whereas the CoNLL task required a five-way classification. In this sense, our task is simpler. Second, our data is a transcription of old records, and much effort has been put into keeping the transcription as close as possible to the original paper text, also preserving mistakes, reporting symbols in letters, and adding comments where the text was unclear, thus making the text incoherent at times. In Example 1, for instance, several words are interposed in the middle of the word "manufactured", all referring to metatextual information, such as headings (VOL. VI. 3 M) and left and right angle brackets containing the comment "UNREADABLE".

(1) "[...] the yarn is manu VOL. VI. 3 M langle UNREADABLE rangle factured in the same manner."

In contrast, the CoNLL data comes from newswire text, the standard kind of text on which NER has been performed to date. In this respect, our task is more difficult, as the data is noisy and text type new.

Eleven out of the sixteen teams who participated in the CoNLL shared task integrated gazetteer information in their systems, and all obtained an improvement in performance, even though [5] noted that currently available external resources still need a lot of manual processing in order for them to be used really successfully in NER systems. We collected location names from the geoXwalk database (`www.geoXwalk.ac.uk`), removed head words, and thus created a specific gazetteer containing a total of nearly 195000 Scottish names. We then added a binary feature (gaz=yes, gaz=no) in the system by using a simple lookup method.

It should be noted that the extract from the geoXwalk gazetteer used in this experiment does not exploit the unique features of the gazetteer. geoXwalk is more than just a simple lookup facility as every geographic feature stored in the gazetteer has its detailed geometry stored with it. The ability to derive the relationships between features implicitly by geometric computation is significant and provides more accurate results than can be ascertained by simple lookups based on hierarchical thesauri methods as in traditional gazetteers. When candidates are referenced against the gazetteer, geoXwalk provides a means to access its alternate geographies (of which there are many in the UK) as well as a standard footprint. For example a candidate placename Knowsley could be resolved as parish code BX003 as well as grid reference 340900, 392300 - 347217, 397660. The result is that more powerful geographical based search strategies can be applied e.g. find me all documents about Gaelic songs that do not reference the Western Isles. In the context of geographical entity recognition the gazetteer affords the possibility of deriving spatial contiguity measures that might assist in disambiguation tasks (see Future Work below).

The naive approach here however, did not obtain any improvement at all over the previous results. Table 2 shows the precision, recall, and f-score of a simple match of gazetteer entries on the whole dataset. The second line in the table shows the performance of a simple match to a name list obtained from the training data, using 10-fold cross-validation.

Table 2: Gazetteer Performance

RESOURCE	PRECISION	RECALL	F-SCORE
specific gazetteer	72.17%	70.20%	71.17%
name list from trainset	56.30%	54.19%	55.23%

The figures in Table 2 on the one hand show that our current approach to using a gazetteer to assist in the task offers little benefit, since given its precision and recall it would be expected to help. On the other hand, they suggest that the annotated entities are often ambiguous (and this is especially true for some lowercase strings), since the precision of the name list extracted from the training data is far from satisfactory. The fact that human annotators disagree on about 15% of the cases only partially accounts for this problem. Clearly, there is scope in further exploring the use of ancillary gazetteer sources as well as more sophisticated matching techniques as means to improving the measures of precision and recall being attained currently.

4. CONCLUSION AND FUTURE WORK

We have shown that an off-the-shelf maximum entropy tagger for named entity recognition can be successfully trained to recognise location names in data other than newswire text, namely historical descriptions of Scotland. We believe that a more sophisticated integration of specific external lexical resources can yield a further improvement on our current system. Given that due to the strategies adopted in the original transcription of the old paper records the data is intrinsically quite noisy, it is likely that even better results can be achieved by putting some effort in to further cleaning the texts.

The data relative to the parish of Edinburgh has an additional layer of annotation, where for each location, 4 subtypes can be specified, namely *boundaries* (countries, counties, parishes, etc.), *hydrographic features* (rivers, estuaries, lakes, etc.), *man-made features* (cities, towns, villages, etc.), and *physiographic features* (mountains, plains, etc.). A preliminary experiment in recognising these specific types of locations yielded a drop in performance of about 20%. Although this is partially to be expected, given that the annotators agreed 77.75% of the time on both the occurrence of a location and its sub-class, additional features and more explicit reference to spatial contiguity by exploiting the geoXwalk gazetteer are likely to lead to significant improvement. A two-stage process involving recognising entities as locations in a first pass, and subsequently classifying them into the 4 subclasses seems to be the most promising avenue. We will report on such experiments in the final version of

this paper.

5. REFERENCES

[1] J. Carletta, S. Evert, U. Heid, J. Kilgour, J. Robertson, and H. Voormann. The NITE XML Toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers*, 35(3):353–363, 2003.

[2] James R. Curran and Stephen Clark. Language Independent NER using a Maximum Entropy Tagger. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 164–167. Edmonton, Canada, 2003.

[3] Ian Densham and James Reid. A geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service. In *Proceedings of the Workshop on the Analysis of Geographic References held at HLT/NAACL 2003*, Edmonton, Canada, 2003.

[4] Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. Named Entity Recognition through Classifier Combination. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 168–171. Edmonton, Canada, 2003.

[5] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada, 2003.