Index of Middle English Prose: A search tool based on language modelling

Honkapohja, Alpo
alpo.honkapohja@ilos.uio.no
University of Oslo, Norway

Thaisen, Jacob
jacob.thaisen@ilos.uio.no
University of Oslo, Norway

Nøklestad, Anders
anders.noklestad@iln.uio.no
University of Oslo, Norway

Access to medieval manuscripts, which form a major part of our collective heritage, is dependent on the search tools available for us. The *Index of Middle English Prose* (IMEP) seeks, for the first time, to locate and identify all surviving English prose texts composed between ca. 1200 and 1500, through its printed catalogues. A crucial weakness of the IMEP is the lack of a satisfactory digital search tool for interested parties to search the collection, especially since the database records (incipits and explicits) preserve linguistic variation inherent to a non-standardised vernacular like Middle English. The poster/presentation demonstrates a web-based search tool for the IMEP developed as a part of an MSCA-IF project , which is capable of handling specifically this type of variation.

Table 1: Linguistic variation in Middle English: incipits from the *Brut* chronicles

In the nobele lande of syrrye there was a nobele kyng and	Rylands Eng 103 [1]
In the noble land of surey þer was a noble king	BL Add 10099 [1]
In the noble land of surre ther was a noble kyng and myhty and a	Lambeth 84 [1]
Off the noble land of surrye ther was a royal kynge	Peniarth 343 [1]
In the noble land of surrye ther was a worthi kynge	Lei UL 47 [1]
Some tyme in the lande of surre ther was a myghty & a ryall	Oxf Un 154 [1]

The challenge lies in the fact that variation exists in orthography, syntax and lexicon and in the title as well as incipits and explicits. For example, Table 1 includes examples of variation in orthography (*there* vs. *þer*, *syrrye* vs. *surey* vs. *surre* vs. *surrye*), lexicon (*nobele kyng*, *royal kynge*, *worthi kynge; In the noble land vs. off the noble land*), word order and incomplete noun phrases (*a noble kyng and myhty* vs. *a myghty & a ryall* [king]). A simple search for a string will find all occurrences of that specific string in a dataset and no other. Searches using regular expressions will also fail to meet the target since they presuppose the user can predict every possible linguistic variation.

The search tool developed for IMEP makes it possible to search for texts regardless of the wide range of linguistic variation. It consists of:

- a database of incipits and explicits in the *IMEP*,
- A set of language character-based n-gram models, one for each incipit or explicit, built using the SRILM language modelling toolkit.
- A fuzzy search Python script for evaluating a search text against the set of language models.

The tool is optimised for recall but not precision. It will retrieve several possible matches for a search string but it will not attempt to select the 'correct' one among them. When the user inputs a search string, the web application looks it up in the database to find exact matches. Simultaneously, it calls the fuzzy search script with the search string. The script:

- uses SRILM to create a language model from the search string and matches each manuscript text against it.
- Selects 100 best matches, recalls the language model for each, and matches the search string against each one.
- Selects 20 best matches from the previous step and returns their database IDs and perplexity values.

The web application, developed in Django, then combines the results of the direct database lookup with the results from the fuzzy search script and presents them in the shape of a list with exact matches at the top, followed by the fuzzy matches in order of increasing perplexity.

This two-step process increases efficiency. Since the IMEP contains thousands of incipits and explicits, matching the search string against each for optimum precision leads to processing times of up to several minutes per search. On the other hand, SRILM can match a large number of texts against a single language model quickly, so testing all manuscript texts against the language model for the search string and reverse matching only for the 100 best matches keeps search times manageable. The tool works well enough and has definite potential for application to the IMEP. Since it is not reliant on a word list, it can be adapted to other datasets.

Bibliography

The Index of Middle English Prose, vols I-XXIII (1984-present). Cambridge: D.S. Brewer.

Stolcke, Andreas & Zheng, Jing & Wang, Wen & Abrash, Victor (2011): "SRILM at sixteen: update and outlook" in: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop.