Normalizing Early English Letters for Neologism Retrieval

Mika Hämäläinen, Tanja Säily and Eetu Mäkelä
Department of Digital Humanities
University of Helsinki

1 Introduction

Our project studies social aspects of innovative vocabulary use in early English letters. In this abstract we describe the current state of our method for detecting neologisms. The problem we are facing at the moment is the fact that our corpus consists of non-normalized text. Therefore, spelling normalization is the first step we need to solve before we can apply automatic methods to the whole corpus.

2 Corpus

We use CEEC (Corpora of Early English Correspondence) [9] as the corpus for our research. The corpus consists of letters ranging from the 15th century to the 19th century and it represents a wide social spectrum, richly documented in the metadata associated with the corpus, including information on e.g. socio-economic status, gender, age, domicile and the relationship between the writer and recipient.

3 Finding Neologisms

In order to find neologisms, we use the information of the earliest attestation of words recorded in the Oxford English Dictionary (OED) [10]. Each lemma in the OED has information about its attestations, but also variant spelling forms and inflections.

How we proceed in automatically finding neologism candidates is as follows. We get a list of all the individual words in the corpus, and we retrieve their earliest attestation from the OED. If we find a letter where the word has been used before the earliest attestation recorded in the OED, we are dealing with a possible neologism, such as the word *monotonous* in (1), which antedates the first attestation date given in the OED by two years (1774 vs. 1776).

(1) How I shall accent & express, after having been so long cramped with the *monotonous* impotence of a harpsichord! (Thomas Twining to Charles Burney, 1774; TWINING_017)
The problem, however, is that our corpus consists of texts written in different time periods, which means that there is a wide range of alternative spellings for words. Therefore, a great part of the corpus cannot be directly mapped to the OED.

4 Normalizing with the Existing Methods

Part of the CEEC (from the 16th century onwards) has been normalized with VARD2 [3] in a semi-automated manner; however, the automatic normalization is only applied to sufficiently frequent words, whereas neologisms are often rare words. We take these normalizations and extrapolate them over the whole corpus. We also used MorphAdorner [5] to produce normalizations for the words in the corpus. After this, we compared the newly normalized forms with those in the OED taking into account the variant forms listed in the OED. NLTK’s [4] lemmatizer was used to produce lemmas from the normalized inflected forms to map them to the OED. In doing so, we were able to map 65,848 word forms of the corpus to the OED. However, around 85,362 word forms still remain without mapping to the OED.

5 Different Approaches

For the remaining non-normalized words, we have tried a number of different approaches.

– Rules
– SMT
– NMT
– Edit distance, semantics and pronunciation

The simplest one of them is running the hand-written VARD2 normalization rules for the whole corpus. These are simple replacement rules that replace a sequence of characters with another one either in the beginning, end or middle of a word. An example of such a rule is replacing yes with ies at the end of the word.

We have also trained a statistical machine translation model (with Moses [7]) and a neural machine translation model (with OpenNMT [6]). SMT has previously been used in the normalization task, for example in [11]. Both of the models are character based treating the known non-normalized to normalized word pairs as two languages for the translation model. The language model used for the SMT model is the British National Corpus (BNC) [1].

One more approach we have tried is to compare the non-normalized words to the ones in the BNC by Levenshtein edit distance [8]. This results in long lists of normalization candidates, that we filter further by their semantic similarity, which means comparing the list of two word appearing immediately after and before the non-normalized word and the normalization candidates picking out
the candidates with largest number of shared contextual words. And finally, filtering this list with Soundex pronunciation by edit distance. A similar method [2] has been used in the past for normalization which relied on the semantics and edit distance.

6 The Open Question

The above described methods produce results of varying degrees of success. However, none of them is reliable enough to be trusted above the rest. We are now in a situation in which at least one of the approaches finds the correct normalization most of the time. The next unsolved question is how to pick the correct normalization from the list of alternatives in an accurate way.

Once the normalization has been solved, we are facing another problem which is mapping words to the OED correctly. For example, currently the verb *to moon* is mapped to the noun *mooning* recorded in the OED because it appeared in the present participle form in the corpus. This means that in the future, we have to come up with ways to tackle not only the problem of homonyms, but also the problem of polysemy. A word might have acquired a new meaning in one of our letters, but we cannot detect this word as a neologism candidate, because the word has existed in the language in a different meaning before.

References

