Adventures of Sherlock Holmes Sample Annotation

Mercedes Hohenberger
Indiana University
mmhohenb@iu.edu

1 Introduction

Having a complete, annotated file of the canon Sherlock Holmes stories would be really cool to linguistically analyze. It's roughly (6?) million words long, so efficient annotation is essential.

Timely, accurate annotation of large texts, and corpora, takes a mix of fairly accurate automation and manual correction of what errors arise from the automation. Before manually correcting a complete, automatically-annotated text, it is ideal to first manually correct a sample of the text. By doing so, one can identify a great deal of what systematic errors exist in the automatic annotation. When the manual annotation correcter(s) know of the systematic errors, they can hone in on the environments in which the errors occur, and actively seek out those contexts when correcting the annotation of the larger text.

I applied the first step of this annotation process, annotating a sample of the text, to a short story collection, The Adventures of Sherlock Holmes, using an annotation software called MorphAdorner.

1.1 The Adventures of Sherlock Holmes

The Adventures of Sherlock Holmes was written by Sir Arthur Conan Doyle and first published as individual stories in The Strand Magazine between 1891 and 1892 (Mulder, 2012). The stories were published together as a book later in 1892. (Mulder, 2012) Various editions of the book have been published since then, including an open source, electronic copy on Project Gutenberg in various formats[1].

The edition on Project Gutenberg contains metadata about the Project Gutenberg eBook edition, the original title and author lines, a table of contents, titles above each short story, roman numerals separating the subsections within each short story, and the Project Gutenberg license for its edition.

1.2 MorphAdorner

MorphAdorner is an annotation software created as part of the MONK project (Mueller and Unsworth, 2009). It was created to lightly, and accurately, linguistically annotate English texts from the ”English language texts from the early Modern English period[2] to the start of the twentieth century” (Burns, 2013). In the second version, MorphAdorner 2.0, the creators ”sought to improve MorphAdorner's processing of several Text Creation Partnership corpora” to, in combination with Abbott, ”turn the TCP texts into the foundation for a ”Book of English...” (Burns, 2013).

It is capable of annotating words in a plain text file, or a Text Encoding Initiative-compliant (Consortium, 2018) XML file of texts. The training corpus's matching time frame and genre, ability to take plain text files (which the entirety of canon Sherlock Holmes stories are publicly available in) as input, and inclusion of both POS tags and lemmas in the annotation made MorphAdorner a logical choice for the automatic annotation of the sample text from ”The Adventures of Sherlock Holmes”.

The annotations for each word in the input text include the spelling of the word as published, its standard spelling, its lemma, a sentence-end boolean, and its Part-of-Speech (POS) tag.

Maintaining the spelling of the word as published allows for faithfulness to the original text.

The standard spelling is a MorphAdorner-specific standard spelling of a given lexeme. By creating a standard form for each lexeme, Mor-

[1] The URL with links for all available formats of the book is http://www.gutenberg.org/cache/epub/1661 .

[2] Here, the beginning of the early Modern English period is defined as the time immediately after the Great Vowel Shift of England, in the late fourteenth century.

phAdorner is able to reduce the data sparsity accrued by having alternate versions of the same lexeme.

In general, a lemma is the dictionary form of a given word. For example, the lemma of the noun 'passions' is 'passion'. The lemma of the verb 'lighting' is 'light'. In the case of a lemma annotation via MorphAdorner, the lemma is the dictionary form of the printed word as found in the lexicon created as part of training, the 'current word lexicon'.

[Lemmas typically mean loss of inflectional morphology, but MA also strips derivational morphology, but marks the derivation with a 'used as' POS tag]

If the word does not occur in the current word lexicon, Morphadorner uses a lemmatizer which contains both words with irregular forms and a rule-based grammar.

The Part-of-Speech Tagger is a smoothed Hidden Markov Model (HMM) Trigram Tagger which implements the Viterbi algorithm. MorphAdorner can POS tag with any given arbitrary tagset, given a annotated training file which is tagged with the arbitrary tagset.

1.3 NUPOS Tagset

MorphAdorner's default tagset is the NUPOS tagset (Mueller, 2009). This tagset is similar to the extended Penn Treebank tagset, but makes finer distinctions than that tagset. The NUPOS tagset also accounts for morpho-syntactic expressions which existed during/after the Early Modern English period but are no longer in everyday use.

This tagset contains 241 tags, not including punctuation. They are a combination of 'word classes', and tense, mood, case, person, number, degree, and negativity, where applicable.

The NUPOS tagset consists of 17 major word classes: adjective, adv/conj/pcl/prep, adverb, conjunction, determiner, foreign word, interjection, negative, noun, numeral, preposition, pronoun, punctuation, symbol, undetermined, verb, and wh-word. Each of these classes is divided further to make 34 word classes. `Differences at word class level` non-splitting of genitive/possessive markers

`Differences at tag-level` The default current word lexicon treats words with negation (ex: cannot, can't, won't, didn't, never, none) as their own lexemes, not splitting the negation off. A consequence of this is the addition of 45 tags to the tagset, which are identifiable by the inclusion of 'x' at the end of the tag.

'un-' words This added 18 tags to the tagset. These are identifiable by the inclusion of '-u' at the end of the tags.

'Used as' class LORD A MERCY The tag labels follow a template.

Addition of second-person singular verbs, both past and present (ex: thou art) Addition second plural imperative Addition first and third plural present

2 Methodology

The annotation process consisted of four general steps: preprocessing, sample acquisition, automated annotation, and manual correction of the automated annotation.

2.1 PreProcessing

In order to transform the plaintext file of the novel from Project Gutenberg[3] into clean input to feed into MorphAdorner, a series of preprocessing steps were needed.

First, I removed all text which came before the first word of the first short story, and all text which came after the last word of the last short story. This included removing Project Gutenberg-specific information, the anthology title and author's name, and the table of contents.

Next, the headers, whose purpose is to help readers navigate through the anthology, needed to be removed. This involved removing the title of each short story, as well as removing the roman numerals used to number each subsection of each short story.

After that, all that remained were novel lines and many empty lines, so any empty lines were removed.

2.2 Sentence Sampling

After preprocessing the book, I obtained a random sample of roughly ten percent of the anthology's sentences. This required splitting the preprocessed text into sentences, and then appending a random collection those sentences to the text file being read into MorphAdorner. Because "The Adventures of Sherlock Holmes" contains roughly six

[3] "The Adventures of Sherlock Holmes" may be found as a plain text file on Project Gutenberg at http://www.gutenberg.org/cache/epub/1661/pg1661.txt

thousand sentences, I collected six hundred of its sentences.

2.3 Automated Annotation

After collecting the six hundred random sentences into a plain text file, they were fed into MorphAdorner. I used the ’adornplaintext’ function, which reads plain text files as input and uses MorphAdorner’s nineteenth century literature training corpus to annotate the input text.

3 Manual Corrections

Erroneous POS tags and lemmas were manually corrected. All known errors are described here.

3.1 Lemma Corrections

The lemma corrections namely fall into one of two categories: the lower-casing of certain nouns, and the combination of lemmas due to mis-tagged symbols.

3.1.1 Undone Capitalization

During lemmatization, capitalization was undone on peoples’ titles, possessive proper nouns, and road names.

People’s titles found in the data included ’Mr.’, ’Mrs.’, ’Miss’, and ’Doctor’. All of these titles’ first letters were lowercased in the automated lemmatization process.

The mis-lemmatized names, McCarthy and Hankey, were in the form of possessive proper nouns. They were lemmatized as lowercased versions of their names: mccarthy and hankey. MorphAdorner correctly lemmatized possessive use of ”Arthur’s”, as ”Arthur”. This suggests that the training corpus may have contained example(s) of ”Arthur’s”, but not enough possessive proper nouns to apply its lemmatization rule to other possessive nouns.

When road terms such as ’road’, ’street’, were part of compound nouns, such as ’Oxford Street’, ’Baker Street’, ’Regent Street’, and ’Edgeware Road’, the compound noun was treated as two separate entities. The road terms were treated as common nouns, and subsequently lowercased when lemmatized.

3.1.2 Symbol and Period Combinations

If a capitalized symbol, such as ’C’, occurred at the end of a statement, it was mistaken as a proper noun. When MorphAdorner identifies a set of at least one alphabetic character(s), followed by a period, as a proper noun, it annotates the string of alphabets and the period as one entity. When this identification is incorrect, it results in two erroneous lemma annotations. The symbol’s lemma contained an additional character, the period, while the period lacked its own lemmatization.

3.2 POS Tag Corrections

There were few erroneous Part-of-Speech tagging errors. For those errors which existed, there were two types of erroneous POS taggings: those which occurred when tokens were part of a compound noun but treated a separate entities, or tokens were separate entities and treated as a single entity.

As was mentioned in the symbols section of lemmatization errors, if a letter was used as a symbol and located immediately before a sentence-ending period, the symbol and period were tagged as a proper noun. The correction was made by manually separating the symbol and the period, and then giving the correct Part-of-Speech tag to each of them.

All road terms were tagged as common nouns, including when those terms were part of a specific road name. Each road name was identified, and each tag was corrected to ’np-n1’, the NUPOS tag for common nouns being used as a proper noun.

4 Discussion

Using the NUPOS tagset was a bad idea. It takes forever to manually correct a file with potentially 241 different tag options!!!!!!!!

5 Further Work

After having annotated a sample of the Adventures of Sherlock Holmes, there is now a list of errors to actively seek out while correcting errors in the whole collection.

References

Philip R. Burns. 2013. Morphadorner v2: A java library for the morphological adornment of english language texts. Evanston, IL. Northwestern University.

The TEI Consortium. 2018. Tei p5: Guidelines for electronic text encoding and interchange.

Martin Mueller. 2009. Nupos: A part of speech tag set for written english from chaucer to the present.

Martin Mueller and John Unsworth. 2009. The monk project final report. *Northwestern University*.

Megan Mulder. 2012. The adventures of sherlock holmes, by arthur conan doyle (1892). *Wake Forest University*.