Detecting Meaning with Sherlock Holmes[*]

The Annotated Holmes

How can we read better?

Francis Bond
Division of Linguistics and Multilingual Studies
http://www3.ntu.edu.sg/home/fcbond/
bond@ieee.org

Location: LT25

Outline

- An online edition of the Canon
- Teaching through Tagging
  — Interactive Lexical Semantics
- Sense Distributions in NTU-MC
- Word Sense Disambiguation
- Sentiment
- Where do we go from here?

An online edition of the Canon

The Adventure of the Readable Texts

For my students (for this class), I wanted texts

- faithful to the original
- easy to read on various devices
- aesthetically pleasing
- linked to the rich world of information of the Great Game

Sadly, no one site has given us what we want, so like many before us, we ended up producing our own.

Data! Data! Data!

First, I took a look at what was already out there:

Edition	Name
Gute	The Project Gutenberg HTML
ACD	Arthur Conan Doyle Encyclopedia
BSW	Baker Street Wiki
SS	Short Stories
SHC	The Complete Sherlock Holmes Canon
Lit2Go	Lit2Go
Camden	Camden House: The Complete Sherlock Holmes
MoonFind	MoonFind: Searching for Sherlock

"Data! data! data!" he cried impatiently. "I can't make bricks without clay." The Adventure of the Copper Beeches (ACD 1892) 4

Selection Criteria

- Useful Metadata: When and where published, Author, Copyright, …
- Annotation: links to TV and film versions, chronologies, definitions of words and Sherlockian scholarship
- Spoilers! Does the annotation reveal the villain?
- Font: Is it a nice serif font (like in the Strand Magazine)
- Resizability: can you read it on your phone
- Pictures: does it have the illustrations nicely embedded?
- Miscellaneous: does it have search?

‘*It has long been an axiom of mine that the little things are infinitely the most important.*’ A Case of Identity

Results

Table 1: Rating Online Texts of the Canon

Edition	Meta	Annotation	Spoilers	Font	Resize	Pics	Misc
ACD	A	A	F	C	F	C	Search
Gute	C	F	-	A	A	F	Book
BSW	A	A	F	B	F	F	Ads
SS	C	F	-	C	A	F	Ads
SHC	C	F	-	A	B	F	ToC
Lit2Go	A	F	-	C	F	F	Audio
Camden	A	F	-	C	F	A	ToC
MoonFind	F	F	-	C	F	F	Search

Every site had some good points, but no one is perfect.

Our Approach

https://github.com/fcbond/sh-canon/

- Resizable, nice font with tufte-css
- Full table of contents, linked from each story
- Metadata in the margin: the date and place of publication, story number, which collection, *date of action?*
- Illustrations in the text ToDo
- Links to annotation at the end
  - Wikipedia
  - Arthur Conan Doyle Encylopedia
  - Bill Dolan's Study Guide

Annotation (new + ToDo)

- Locations link to google maps
  - What should we do for made-up places?
  - Should link locations to geonames

ToDo

- Currency links to amount in today's currency
  done for SPEC
- Person + Organization should link to entry in ACD Encylopedia
- Texts are searchable with a site-specific google search
  metadata is hidden from indexing by the `<!--googleoff: all-->` command.

Annotation ToDo

- The sense level annotations of the *NTU Multilingual Corpus*, which links each open class word to its definition in Wordnet
  - We can also show translations and/or hypernyms
- Content level annotation like the Annotated Sherlock Holmes
- We have syntactic annotation (treebanks) for *The Adventure of the Speckled Band* which it would be good to display for those who are interested
- More from the Beacon Society's award winners, including information about word meaning, rhetorical devices, …
- Links to facsimiles of the original stories

Translations and Glosses

- The Canon has been widely translated: we would like to also prepare editions of foreign languages, and consider the issue of linking the translations. E.g., *InLéctor*, who have produced a nice bilingual English-Spanish edition of *The Adventures of Sherlock Holmes*.

- With the multilingual wordnets, we can present glosses of the correct sense, in any of 34 languages (although not all will be complete)

- We can investigate how to choose the words to highlight
  - Low frequency (rare) word
  - Lower frequency sense of a word (uncommon meaning)
  - No direct translation — give definition
  - Word selected by e.g., Bill Dolan, my students, …

  it is quite hard to do this nicely, ...

Some Technical Issues

- There is a lot of duplication in online resources (and I have only made it worse)

- How can we share information?

  - Put the code and meta-data on github

    ```
    SPEC PUBLISHED The Strand Magazine in February 1892
    SPEC ORDER 10
    SPEC COLLECTION The Adventures of Sherlock Holmes
    SPEC DORN http://www.beaconsociety.com/uploads/3/7/3/8/37380505/dorn
    SPEC TITLE The Adventure of the Speckled Band
    SPEC IMAGE She lifted her veil 104.jpg
    ```

    are there standard ways to refer to lines?

- Is everyone happy to share information?

  - We have to make sure not to use data without permission

Conclusion

- We now have some stories with every (open-class) word linked to an entry in a dictionary, and its grammatical category
  What else can we do with this?

- For SPEC we also have this in Chinese, Japanese, Indonesian, Italian (and are working on Abui, Bulgarian, Dutch, Spanish and Polish). And these are linked to the English
  What we can do with this?

- If you have any ideas (or want the data, or want to help me make more) please let me know.