ACM
SIGIR 2014
July 6-11, 2014
The 37th Annual Intl. ACM SIGIR Conf.

CharBoxes: A System for Automatic Discovery of Character Infoboxes from Books

Manish Gupta *
Microsoft, India
gmanish@microsoft.com

Piyush Bansal *
IIIT Hyderabad
piyush.bansal@research.iiit.ac.in

Vasudeva Varma
IIIT Hyderabad
vv@iiit.ac.in

Problem: Extract character centric structured information from books to ease automated generation of Wikipedia like Infoboxes.

Example: The following example is for "Harry Potter" from "Harry Potter and the Sorcerer's Stone"

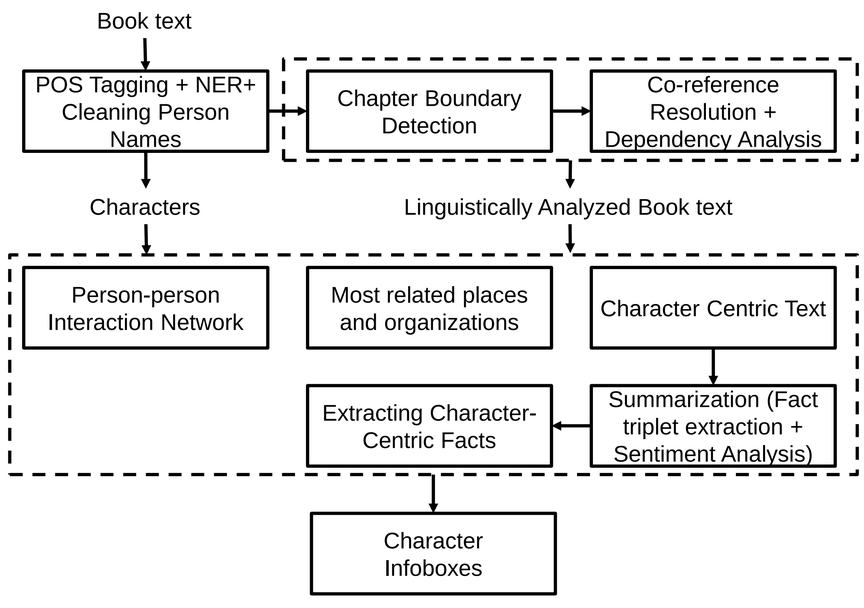
- Most related persons – Hermione Granger, Ron Weasley, Rubeus Hagrid, Snape, Dudley
- Most related places – Diagon Alley, Hogwarts, Privet Drive, Gringotts
- Most related Organizations – Gryffindor, Hogwarts, Dursleys
- Positive Personality Traits: powerful, lucky, curious, funny, bright
- Negative Personality Traits: cold, stupid, nasty, nervous, angry
- …

Related Work

- Analysis of books or multi-documents
  - Summarization of books [3, 4].
- Extracting structured data from free text
  - Extracting structured data from free text [1, 2, 5, 6].
  - The proposed system is the first to focus on extraction of character infoboxes from books.

System Components:

- **Character Extraction**
  - Post-process POS Tagged data to obtain names
  - Handle diminutives
  - Maps parts of names to canonical name
  - Maintain list of ambiguous names
- **Linguistic Analysis**
  - Co-reference Resolution
  - Sentiment Analysis
  - POS Tagging+ NER
  - Chapter Boundary Detection
  - Parse Tree Analysis
  - Understand dependencies
  - Understand subject-predicate-object triplets



Schematic diagram showing the components of system.

* Both the authors have equal contribution.

- **Person-person Interaction Network Construction**
  - Build an interaction graph between characters using non-ambiguous mentions and dialogue extraction
  - Perform disambiguation of ambiguous mentions using
    - Context words
    - Mention of full name in vicinity
    - Frequency of co-occurrence with other entities in the vicinity based on the graph
  - Use disambiguated mentions to refine interaction graph

- **Character Centric Facts Extraction**
  - Sociability of a person, a measure of his connectivity
  - Use parts of speech, dependencies and redundancy
  - Most common traits of a person
  - Personality traits (positive/negative)
  - Physical Appearance (Looks)
  - Dresses
  - Common actions

- **Character Centric Summary**
  - Consider all sentences containing the character
  - Remove sentences which also contain other characters
  - Remove sentences with quotations
  - Rank sentences with more entities higher
  - Rank longer sentences higher
  - Rank sentences which introduce a new entity higher
  - Rank sentences with dress description or looks of the character higher
  - Rank sentences with extreme sentences higher

Applications of the System:

- Automated population of Knowledge bases with Character specific information
- Effective summarization
- Effective marketing of books
- Aid understanding in classroom teaching
- Interesting for readers and viewers of movies based on books.

References:

[1] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. Toward an Architecture for Never-Ending Language Learning. In AAAI, 2010.

[2] Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. Web-scale Information Extraction in KnowItAll: (Preliminary Results). In WWW, pages 100–110, 2004.

[3] Anna Kazantseva and Stan Szpakowicz. Summarizing Short Stories. Computational Linguistics, 36(1):71–109, Mar 2010.

[4] Rada Mihalcea and Hakan Ceylan. Explorations in Automatic Book Summarization. In EMNLP, pages 380–389, 2007.

[5] Stephen Soderland. Learning Information Extraction Rules for Semi-Structured and Free Text. Machine Learning, 34(1-3):233–272, 1999.

[6] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A Large Ontology from Wikipedia and Wordnet. Web Semantics, 6(3):203–217, 2008.

Microsoft

Paper ID: de18

IIIT, HYDERABAD