Building and Annotating Corpora of Collaborative Authoring in Wikipedia

Johannes Daxenberger, Oliver Ferschke and Iryna Gurevych

Workshop: Building Corpora of Computer-Mediated Communication: Issues, Challenges, and Perspectives
Research Focus

Web User -> Production

Collaboration

Line-up

- Wikipedia Article Revisions
  - Motivation
  - Corpus
  - Annotation

- Wikipedia Talk Pages
  - Motivation
  - Corpus
  - Annotation
Revision History: Edits

Revision as of 17:53, 5 December 2005 (edit)
RJII (talk | contribs)
m (→Etymology: grammar)
← Previous edit

Revision as of 18:27, 5 December 2005 (edit)
(undo)
Infinity0 (talk | contribs)
(what is capitalism without wage labour? where is there capitalism without wage labour? how can capitalism exist without wage labour?)
Next edit →

Line 2:

{{wikiquote}}

In common usage, the word "capitalism" means an economic system in which all or most of the [means of production] are [private property privately owned] and [management operated] (usually for [profit]), and the [investment] of [capital (economics) capital] and the [production], [logistics distribution] and [prices] of [commodities] ([good (economics) goods] and [services]) are determined mainly in a [free market].

Line 2:

{{wikiquote}}

In common usage, the word "capitalism" means economic system in which all or most of the [means of production] are [private property privately owned] and [management operated] (usually through the employment of [wage labour], and for [profit]), and the [investment] of [capital (economics) capital] and the [production], [logistics distribution] and [prices] of [commodities] ([good (economics) goods] and [services]) are determined mainly in a [free market].
Motivation: Wikipedia Revision History as data source for NLP applications

- Wikipedia-related usage
  - quality assessment of articles
  - vandalism detection
  - author behavior

- Wikipedia as source for linguistic applications
  - error detection/correction
  - paraphrasing
  - textual entailment
  - information retrieval
Edits vs. Revisions

Each pair of adjacent revisions \( (r_{v-1}, r_v) \) creates a set of \( n \) edits \( e_{v-1,v}^k, k \in \{0,1,\ldots, n\} \).
Wikipedia Quality Assessment Corpus

- Balanced collection of 10 Featured and 10 Non-Featured Articles from the English Wikipedia
- For each Featured Article (FA), there is a Non-Featured Article (NFA) of comparable length and edit frequency
- 1995 edits in 891 revision pairs
  - divided into 4 groups from different revision history stages
Wikipedia Edit Category Taxonomy

SURFACE
  - MARKUP
    - Insert
    - Delete
    - Modify
  - Paraphrase
  - Grammar/Spelling
  - Relocation

WIKIPEDIA POLICY
  - INFORMATION
    - Revert
    - Vandalism
  - FILE
  - REFERENCE
  - TEMPLATE
    - Insert
    - Delete
    - Modify
Annotation Study

- 3 Expert Annotators (students)
- Multi-labeling: each edit is assigned a set of categories \( Y \subset L \), where \( L \) is the set of categories, hence \( |L| = 21 \)
- Gold standard: majority votes
- Data reliability (inter-annotator agreement):
  - \( A_O = 0.96 \), \( \kappa_{pool} = 0.65 \)
  - Fleiss’ Kappa per Top-Level Categories:
    - Surface: \( \kappa = 0.61 \)
    - Text-Base: \( \kappa = 0.66 \)
    - Wikipedia Policy: \( \kappa = 0.79 \)

Corpus available for download
Collaboration and Quality: Distribution of Edit Categories for Groups

Absolute Number of Edits labeled with top level categories
Discussion spaces in Wikipedia

**Article**

Natural language processing

This article has multiple issues. Please help improve it or discuss these issues on the talk page.

- It needs additional citations for verification.
- It may be in need of reorganization to comply with Wikipedia's layout guidelines. Tagged since July 2008.

Natural language processing (NLP) is a field of computer science, artificial intelligence (also called machine learning) and linguistics concerned with the interactions between computers and human (natural) languages. Specifically, the process of a computer extracting meaningful information from natural language input and/or producing natural language output. In theory, natural language processing is a very attractive method of human-computer interaction. Natural language understanding is sometimes referred to as an AI-complete problem because it seems to require extensive knowledge about the outside world and the ability to manipulate it.

Whether NLP is distinct from, or identical to, the field of computational linguistics is a matter of perspective. The Association for Computational Linguistics defines the latter as focusing on the theoretical aspects of NLP. On the other hand, the open-access journal "Computational Linguistics", styles itself as "the longest running publication devoted exclusively to the design and analysis of natural language processing systems" (Computational Linguistics (Journal)).

Modern NLP algorithms are grounded in machine learning, especially statistical machine learning. Research into modern statistical NLP algorithms requires an understanding of a number of disparate fields, including linguistics, computer science, and statistics. For a discussion of the types of algorithms currently used in NLP, see the article on pattern recognition.

**Discussion**

Talk: Natural language processing

Incorporate stemming?

I suggest adding a link to stemming in the see also or subtasks or challenges. I am not sure who is responsible for editing this article though, and I don't want to edit it myself without asking. Is stemming too detailed, or a subtask of another subtask only like IR? Not sure. I thought it was a pretty popular problem. Josh Froelich 19:46, 13 December 2006 (UTC)

"I am not sure who is responsible for editing this article though. You are, feel free to edit any wikipedia page. Yes it feels very wrong the first few times, but your fine to do so. Someone will fix it if your wrong anyhow. Scott A Herbert (talk) 13:56, 24 February 2011 (UTC)"

External links

I think everyone would agree the external links section is a complete mess and full of spam, vanity links, and other links that don't add anything to the article. I count 47 external links. I'm sure there is someone out there who supports each one, but I think we all can agree that 47 is too many and there is certainly some redundancy.

I know it can be hard to part with large chunks of an article, but I propose the following: we assume that we are going to delete all of them and anyone who wants a link kept should nominate it here on the talk page. We can then discuss whether it actually adds something unique. Please keep in mind WP-EL also.

--Selket 22:50, 1 February 2007 (UTC)

The implementations links seem alright. However the R & D groups links are way too many. Unfortunately, each group would want their own link up there. Also, there were a few links to blogs. Am I right in believing that those links should be deleted?

Ummonk 22:06, 4 February 2007 (UTC)

I think the implementations links should be removed per WP-EL, WP:SPAM, and
Motivation: Why analyze Talk pages?

- Quality assessment
  - Which quality flaws are discussed on the Talk page?

- Article augmentation
  - Incorporate a broad range of opinions about controversial topics
  - Identify and mark incorrect article content

- Insights into the collaborative writing process

- **Focus of our work:**
  Coordination efforts for article improvement
Pre-processing of Talk pages

1. Segmentation of Talk page into discussions

Easy:
- MediaWiki Markup
- One section per discussion
- Discussion title explicitly defined

2. Segmentation of discussions into turns

Explicit Markup? ✗
User Signatures? ✗
Revision History? ✓

Only 67% of turns signed

3. Identify correct authors for turns

- \(\text{œ}^\text{TM}\) 04:41, 20 December 2010 (UTC)
- [NerdyScienceDude](#) 13:24, 17 December 2010 (UTC)
- [Greatorangepumpkin](#) 18:06, 16 December 2010 (UTC)
- [Puffin](#) Let's talk! 17:33, 2 July 2011 (UTC)
- [Dylan620](#) 21:20, 13 January 2011 (UTC)
- Taku
Corpus Creation

Selection according to discussion size

100 Talk pages

- 50 small pages: 4-10 turns
- 40 middle-sized pages: 11-20 turns
- 10 large pages: >20 turns

Total 1367 turns for annotation

69.595 articles
5783 active Talk pages
683 relevant* Talk pages

* Talk pages with more than 3 contributions
Annotation Schema

17 labels in 4 categories

Focus: Coordination efforts for article improvement

e.g. missing info, spelling error

Dialog Act

Explicit Performative

Information Content

Interpersonal

e.g. report of error correction

e.g. info providing, info seeking

e.g. positive attitude, negative attitude
Annotation Study

- Two annotators
  - trained on separate set of 10 Talk pages
  - assign multiple labels per turn
  - allowed to discuss difficult cases

- Gold Standard
  - Disagreements decided by expert annotator

- Dataset reliability
  - Inter-rater agreement: $A_O = 0.94$, $\kappa_{pool} = 0.67$

http://www.ukp.tu-darmstadt.de/data/wikidiscourse
A Survey of NLP Methods and Resources for Analyzing the Collaborative Writing Process in Wikipedia

Oliver Ferschke and Johannes Daxenberger and Iryna Gurevych


- Wikipedia Revisions
  - The Concept of Revisions in Wikipedia
  - NLP Applications
  - Article Trust, Quality and Evolution
  - Vandalism Detection
- Discussions in Wikipedia
  - Technical Overview
  - Work Coordination and Conflict Resolution
  - Information Quality
  - Authority and Social Alignment
  - User Interaction
- Tools and Corpora
Sources

Corpora

- Wiki-Edits (Gold Standard annotations, Annotation Guidelines)
  - http://www.ukp.tu-darmstadt.de/data/wiki-edits (CSV)
- Wiki Discussions (Annotations)
  - http://www.ukp.tu-darmstadt.de/data/wikidiscourse (XMI, MMAX)