Bauhaus-Universitaet Winter 2015/2016

Behind the Article:

Recognizing Dialog Acts in Wikipedia Talk Pages

A paper by Oliver Ferschke, Iryna Gurevych and Yevgen Chebotar

Presented by:
Roxanne El Baff

Overview

- Abstract
- Introduction
- Related Work
- Annotation Schema
- Corpus Creation and Analysis
- Automatic Dialog Act Classification
- Conclusions

Overview

- **Abstract**
- Introduction
- Related Work
- Annotation Schema
- Corpus Creation and Analysis
- Automatic Dialog Act Classification
- Conclusions

Abstract

- Propose an annotation schema for the discourse analysis of Wikipedia Talk Pages aimed at the coordination efforts for article improvement
- Perform automatic dialog act classification on Wikipedia discussions and achieve an average of F_1-score of 0.82

Overview

- Abstract
- **Introduction**
- Related Work
- Annotation Schema
- Corpus Creation and Analysis
- Automatic Dialog Act Classification
- Conclusions

Introduction

- **Paradigm of information sharing**: Participatory and collaborative content production
  - Collaborative writing process differs considerably from the way individual writing is done.
- Wikipedia supports collaborative writings:
  - Wikipedia offers a communication platform, the **Talk pages**, where they can discuss the ongoing writing process with other users

Introduction - Goal

How the huge online community around Wikipedia regulates and enforces standards of behavior and article quality?

→ Main goal of the paper:

Analyze the content of the discussion pages of the English Wikipedia with respect to the dialog acts aimed at the coordination efforts for article improvement

Introduction - Goal

How the huge online community around Wikipedia regulates and enforces standards of behavior and article quality?

→ Main goal of the paper:

Analyze the content of the discussion pages of the English Wikipedia with respect to the dialog acts aimed at the coordination efforts for article improvement

A dialog act label characterizes the intention and the role of a contribution in a dialog

Introduction - Primary Contributions

Primary contributions in this paper:

1. An annotation schema for dialog acts reflecting the efforts for coordinating the article improvement
2. The Wikipedia discussion corpus, consisting of 100 segmented and annotated Talk pages
3. Dialog act classification pipeline (using machine learning algorithms & feature selection techniques) and achieve F_1-score of .82 on the corpus

Overview

- Abstract
- Introduction
- **Related Work**
- Annotation Schema
- Corpus Creation and Analysis
- Automatic Dialog Act Classification
- Conclusions

Related Work

1962 (John Austin) Use of language as a tool for performing actions

1969 - 1976 (Searle) Systemized the Speech act theory + his classification of illocutionary act is still used as a starting point for creating dialog act classification schemata for NLP

2007 -2011 (Viegas et al. then Schneider et al.): Analyze how talk pages are used for planning the work on articles and resolving disputes among the editors

2011 (Bender et al.): Analyze how the participants in Wikipedia discussions establish their credibility and how they express agreement and disagreement toward other participants or topics

Related Work

2008 (Stvilia et al.) Analyze how information quality in Wikipedia articles is assessed on the Talk pages and which type of IQ problems are identified by the community

2011 (Laniado et al.) Examine Wikipedia discussion network in order to capture structural patterns of interaction

→ No corpus that reflects the efforts of article improvement in Wikipedia discussions

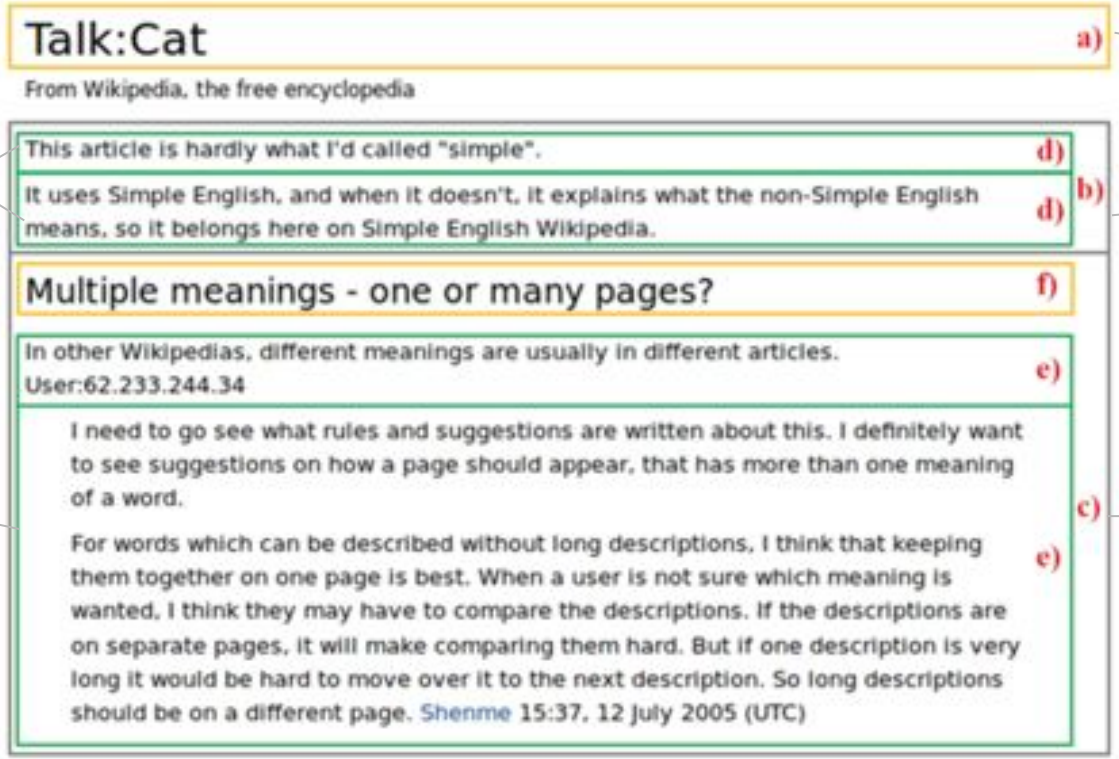
Overview

- Abstract
- Introduction
- Related Work
- **Annotation Schema**
- Corpus Creation and Analysis
- Automatic Dialog Act Classification
- Conclusions

Annotation Schema

- Main purpose of **Wikipedia Talks**
  - Coordination of the editing process with the goal of improving and sustaining the quality of the article
- Criteria for article quality are defined in the guidelines
  - Good article
  - Very good article
  - ⇒ Well written in simple English, comprehensive, stable, accurate, verifiable and follow the Wikipedia style guidelines

Annotation Schema



(a) Talk page title

(b) untitled discussion topic

(c) titled discussion topic

(d) Unsigned turns

(e) Signed turns

Annotation Schema

- Composed an annotation schema that reflects the coordination efforts for article improvement →
  - Manually analyzed a set of 30 Talk pages from the English Wikipedia to **identify the types of article deficiencies that are discussed and the way article improvement is coordinated**
- Used Stvilia's findings of IQ: which identifies 12 types of quality problems

⇒ Resulting tagset: 17 labels which can be divided into 4 categories

Annotation Schema

- Category 1: Article Criticism
  - Comments that identifies deficiencies in the article
- Category 2: Explicit Performative
  - Announce, report or suggest an edit
- Category 3: Information Content
  - Communicate new information, request for information or suggest changes to an established fact
- Category 4: Interpersonal
  - attitude towards other participants

Annotation Schema

- Category 1: Article Criticism

Label	Description	Example
Article Criticism		
CM	Content incomplete or lacking detail	*It should be added (1) that voters may skip preferences, but (2) that skipping preferences has no impact on the result of the elections.*
CW	Lack of accuracy or correctness	*Kris Kringle is NOT a Germanic god, but an English mispronunciation of Christkind, a German word that means "the baby Jesus".*
CU	Unsuitable or unnecessary content	*The references should be removed. The reason: The references are too complicated for the typical reader of simple Wikipedia.*
CS	Structural problems	*Also use sectioning, and interlinking*
CL	Deficiencies in language or style	*This section needs to be simplified further; there are a lot of words that are too complex for this wiki.*
COBJ	Objectivity issues	*This article seems to take a clear pro-Christian, anti-commercial view.*
CO	Other kind of criticism	*I have started an article on Google. It needs improvement though.*

Annotation Schema

- Category 2: Explicit Performative

Explicit Performative		
PSR	Explicit suggestion, recommendation or request	*This section needs to be simplified further*
PREF	Explicit reference or pointer	*Got it. The URL is* `http://www.dmbeatles.com/history.php?year=1968`
PFC	Commitment to an action in the future	*Okay, I forgot to add that, I'll do so later tonight.*
PPC	Report of a performed action	*I took and hopefully simplified the "[[en:Prehistoric music—Prehistoric music]]" article from EnWP*

Annotation Schema

- Category 3: Information Content

Information Content		
IP	Information providing	*"Depression" is the most basic term there is.*
IS	Information seeking	*So what kind of theory would you use for your music composing?*
IC	Information correcting	*In linguistics and generally speaking, when Talking about the lexicon in a language, words are usually categorized as 'nouns', 'verbs', 'adjectives' and so on. The term 'doing word' does not exist.*

Annotation Schema

- Category 4: Interpersonal

Interpersonal		
ATT+	Positive attitude towards other contributor or acceptance	*Thank you.*
ATTP	Partial acceptance or partial rejection	*Okay, I can understand that, but some citations are going to have to be included for [[WP:V]].*
ATT−	Negative attitude towards other contributor or rejection	*Now what? You think you know so much about everything, and you are not even helping?!*

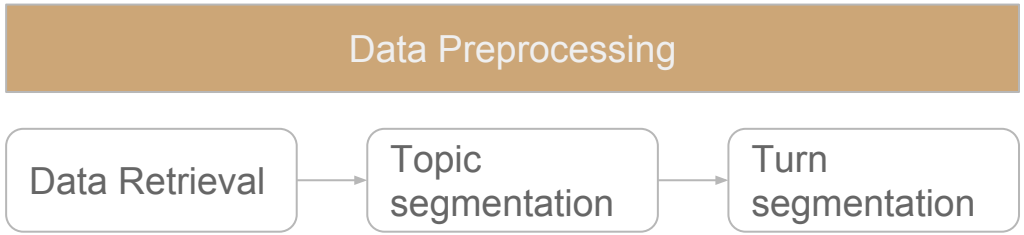
Overview

- Abstract
- Introduction
- Related Work
- Annotation Schema
- **Corpus Creation and Analysis**
- Automatic Dialog Act Classification
- Conclusions

Corpus Creation and Analysis

- The corpus consists of 100 annotated Talk pages extracted at the 4th of Apr. 2011:
  - Pages with less than 4 contributions were disregarded
  - Pages were selected according to the number of turns and the pages were divided into 3 classes:
    - (i) Discussion pages with 4-10 turns → 50 randomly extracted pages
    - (ii) Discussion pages with 11-20 turns → 40 randomly extracted pages
    - (ii) Discussion pages with > 20 turns → 10 randomly extracted pages

Corpus Creation and Analysis



Corpus Creation and Analysis

Data Preprocessing

Annotation Process

- Used MMAX2
- 2 annotators were introduced to the annotation schema by an instructor and trained on a an extra set of 10 discussion pages.
- During the annotation of the corpus, the annotators were allowed to discuss difficult cases
- Expert decided all cases which the annotations does not match

Corpus Creation and Analysis

Analysis:

- Corpus contained 313 discussions: 1367 turns by 337 users
- Average length of turns: 42 words
- 290 of 337 contributors are registered users and 129 wrote aninymously
- Most frequent labels:
  - Information Providing (IP), requests (PSR) and reports of performed edits (PPC)
    - IP : 78% of the turms
    - > ¼ : labeled PSR and PPC
    - PSR ~= PPC and PFC
- Most common label adjancency pairs: PSR → PPC
- Article criticism labels have been assigned to 39.4% of all turns. ½ are assigned to the first run