WikiTalk corpus: a resource for defining "online discussion" genre

Ho-Dac L.-M.(1) and Laippala V.(2)

(1) CLLE-ERSS: CNRS & University of Toulouse,
(2) University of Turku

Abstract

Wikipedia constitutes a popular and extremely useful resource for studies in both linguistics and natural language processing (see [Yano and Kang 2008]). This paper introduced a new language resource based on the French Wikipedia online discussion pages, the WikiTalk corpus. The corpus includes 159,578,279 words and 3,022,240 posts structured into 1,024,351 thematic sections and has automatic syntactic analyses produced with the Talismane analyser ([Urieli 2013]), and is made publicly available. In this paper, we present the different steps of construction process and a linguistic analysis on the discussions and on the language used in them.

1 Introduction

This paper presents a new language resource based on the French Wikipedia online discussion pages, the WikiTalk corpus. The corpus includes 159,578,279 words and 3,022,240 posts structured into 1,024,351 thematic sections and has automatic syntactic analyses produced with the Talismane, a statistical transition-based dependency parser ([Urieli 2013, Urieli and Tanguy 2013]). It is also made publicly available at the publication of this paper.

In the following, after describing the advantages of Wikipedia as a source of language data, we present the different steps of the corpus constitution and evaluate the method applied in the identification of the different elements of the talks. Finally, we analyse the linguistic characteristics of the corpus by comparing the WikiTalk corpus to other web genres and by combining hypothesis-driven and data-driven approaches.

2 Wikipedia as corpus


This study focuses on Wikipedia talk pages associated to article pages. The main advantages of Wikipedia talk pages are the following: comparing to other kinds of online discussions, Wikipedia gives free data which are fairly well-formed and associated with rich meta-data including topics, shared knowledge (given by the content of the collaborative written article), writer information but also e.g. on the type of the discussion, i.e. whether it has been staying neutral or it the writers need to keep calm.

In contrast, one of the main disadvantages is that comparing to other kinds of online interactions, the Wikipedia talk timeline is very difficult to follow because posts can be inserted anywhere in the talk pages without any chronological order. Moreover, users may modify a previous post without changing the timestamp which makes the timeline somehow artificial.

3 WikiTalk corpus constitution

The WikiTalk corpus is composed of talk pages extracted from the French Wikipedia snapshot from 12th may 2015 which contains 3,487,480 talk pages among which only 366,326 are kept as a part of the released WikiTalk Corpus.

1Distributed under Creative Commons Attribution-ShareAlike 3.0
57% (1,990,927) talk pages associated to user pages were removed. Even if these talk pages belong to online discussions genre, they differ from talk pages associated to specific article according to topic homogeneity and shared knowledge access. Among the remaining talk pages only 24% (366,326) contains more than two words. The 366,326 remaining talk pages were structured into threads and posts on the basis of wikicode. Threads correspond to division delimited by (sub)headings signalled with the wiki traditional syntax: /==.*==/. Posts are delimited according to

1. timestamp and eventually user signature such as: Viking59 10 mai 2009 à 17:16 (CEST)
2. a change of interactonnal level indicated with zero, one or more semi-colon (:) at the beginning of the post.

Once threads and posts delimited, all discussions were formatted according to TEI-P5 norm. Meta-data are encoded in the teiHeader as illustrated below with the <classDecl> element ("discipline" indicates associated thematic portals, "avancement" corresponds to project’s quality scale and "interaction" gives information about "staying neutral" or "keep calm" (i.e. lightgreen)).

```
<category type="discipline">!
  <catDesc>Politique française</catDesc>
</category>
<category type="avancement">
  <catDesc>B</catDesc>
</category>
<category type="interaction">
  <catDesc>{{Appel au calme|lightgreen}}</catDesc>
</category>
```

Discussion structure is encoded according to the following TEI elements:

- <div> for threads
- <head> for topic titles and
- <sp who="user" when="timestamp" interactionalLevel="#"> for posts.

Table 1 gives final numbers of the WikiTalk corpus. Finally, seven extracted talk pages were manually evaluated

<table>
<thead>
<tr>
<th>discussions</th>
<th>sections</th>
<th>posts</th>
<th>mots</th>
</tr>
</thead>
<tbody>
<tr>
<td>366,326</td>
<td>1,024,351</td>
<td>3,022,240</td>
<td>132,406,816</td>
</tr>
</tbody>
</table>

Table 1: Quantitative overview of the WikiTalk corpus

i.e. 413 posts and 47,284 token. Results show that 23 posts were not extracted and 33 posts were wrongly delimited, among which 25 merged several posts in one. As a result, extraction process have a precision of 0.92 and a recall of 0.95.

4 Linguistic characteristics

Different modes of Computer-Mediated Communication (CMC) vary significantly between each other and between other genres. In order to describe the particular mode(s) of CMC present in the WikiTalk corpus, we propose a set of contrastive analyses combining hypothesis-driven and data-driven approaches. The former concerns the evaluation of readability, subjectivity and interactional discourse structures. The latter aims at identifying syntactic n-grams which have the capacity to classify segments as belonging to WikiTalk corpus.

Table 2 shows corpora used for comparing WikiTalk to other computer-mediated communications. "Rue89" and "AgoraVox" are online journalism corpora: "Rue 89" is composed of articles written by professional journalists while "AgoraVox" is made up of articles written by internet users (cf. [Küppers and Ho-Dac 2010]). "WikiArticle" is composed with article pages extracted from a Wikipedia 2013 version. "Health Fora" gather together four online fora about health care. All corpora were parsed automatically using Talismane.

---

2 1,013,791 (68%) talk pages were just activated without fill of posts and 116,432 (8%) were redirected to another talk pages.
3 <sp> is defined in TEI-P5 guidelines as "An individual speech in a performance text, or a passage presented as such in a prose or verse text."
4 Soon available http://redac.univ-tlse2.fr/
<table>
<thead>
<tr>
<th>Corpora</th>
<th>tokens</th>
<th>genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rue89</td>
<td>2,192,995</td>
<td>online professional press</td>
</tr>
<tr>
<td>AgoraVox</td>
<td>4,099,662</td>
<td>online citizen press</td>
</tr>
<tr>
<td>WikiArticle (2013)</td>
<td>226,207,672</td>
<td>encyclopedia articles</td>
</tr>
<tr>
<td>Health Fora</td>
<td>236,368,151</td>
<td>online discussion</td>
</tr>
<tr>
<td>WikiTalk (2015)</td>
<td>132,406,816</td>
<td>online discussion</td>
</tr>
</tbody>
</table>

Table 2: Corpora used for contrastive analyses

4.1 WikiTalk main characteristics

Three aspects of linguistic characteristics were analysed through all five corpora:

- writing level and readability by measuring the amount of unknown words,
- subjectivity by projecting an affect lexicon ([Augustyn et al. 2008]),
- discourse structures by extracting posts’ opening.

Figure 1 shows the percentage of unknown words according to Talismane automatic lemmatization\(^5\). As expected, health fora have the highest percent of unknown words: over 25% of selected POS are unknown, up to 38% for nouns. WikiTalk also contains unknown words (7% for all POS), but very few comparing to health fora. Not so much differences between talk and article pages were observed, except for adverbs (7% unknown in talk pages vs. 3.5% in article pages and online press). These findings agreed with recent studies on Wikipedia readability ([Elia 2009, Walton 2009, Myers 2010]).

As for subjectivity, figure 2 shows the amount of common nouns, verbs and adjectives referring to emotion according to an affect lexicon ([Augustyn et al. 2008]). There again, we found no significant difference between talk and articles pages: WikiTalk as WikiArticle show under 2% of ”affect words”, that is less than online news (3.3%) and health fora (4.5%). These results fit with ”neutral point of view” which is one of the Wikipedia’s three core content policies\(^6\), as [Myers 2010] also observed. Last analysis concerns post opening which were studied by collecting all n-grams made of 1, 2 or 3 words that occur at the beginning of posts in the two online

\(^5\) Lemma remains unknown if it is not found in the Leff lexicon [Sagot et al. 2006]
\(^6\) the other two are ”Verifiability” and ”No original research”
discussion corpora: wikiTalk and health Fora. As a result, WikiTalk posts frequently begin with argumentative structures associated with consensus building such as It seems to me that, I agree with, If you are, Actually, etc. On the contrary, posts in health Fora are full of salutations and thanks openings (Hello girls or Thanks).

4.2 WikiTalk vs. WikiArticles automatic classification

This section presents a data-driven comparison of the global linguistic characteristics of WikiTalk to WikiArticle. While traditional approaches often concentrate on the lexical differences between texts (see e.g. the keyword analysis in [Scott and Tribble 2006] reflecting thematic and stylistic features, we propose to describe the corpora by their structural characteristics using unlexicalized syntactic N-grams ([Kanerva et al. 2014, Goldberg and Orwant 2013]). This method enables a more robust analysis on text characteristics that does not depend on the text topic but attempts to generalize the level of description beyond individual lexical topics to typical structures ([Laippal et al. 2015a, Laippal et al. 2015b]).

The definition of the typical structures is done as a text classification task using the Vowpal Wabbit linear classifier ([Agarwal et al. 2011]). In addition to being fast and easily adjustable to large corpora, it has the advantage of generating the list of the significant text class features and their weights. We use this list as the source of the text class characteristics to be analysed. The classification is accomplished using the stochastic gradient method with a 50%/50% split on the training and text corpora and text segments of three sentences to be classed. As features we use unlexicalized tri-arcs composed of three syntax dependencies between tokens with the actual lexical information deleted but with all other information on the syntactic dependency, Part-of-Speech Class and other morphological features (see Figure 3 for an example.)

Table 3 gives the most typical syntactic n-grams used by classifier to distinguish WikiTalk to WikiArticle. In order to evaluate the significance of the syntactic N-grams as text class characteristics, we compared our syntactic n-grams method to classification based on lexical features. Table 4 gives the F-scores for these different methods of classification. Results show that our method based on syntactic n-grams is certainly better for discovering structural features than for classifying WikiTalk vs. WikiArticle. Nevertheless, as the work presented here is still in progress, the next step consists in testing other classifier settings, especially concerning the training/test
Table 3: WikiTalk typical syntactic n-grams

<table>
<thead>
<tr>
<th>syntactic n-gram</th>
<th>description</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>I+verb+that</td>
<td>I think that</td>
<td>I can add that</td>
</tr>
<tr>
<td>I+aux+verb</td>
<td>I have withdrawn</td>
<td>I have taken the opportunity to</td>
</tr>
<tr>
<td>I+verb+adverb</td>
<td>I also xxx</td>
<td>I therefore xxx</td>
</tr>
<tr>
<td>conditionnal</td>
<td>It would be better to</td>
<td>It should be more explicit in</td>
</tr>
<tr>
<td>It+extraposition</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: F-score comparing lexical methods and syntactic methods for automatic classification

<table>
<thead>
<tr>
<th></th>
<th>F-score</th>
<th>talk pages</th>
<th>articles pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag-of-lemmata</td>
<td>.92</td>
<td>.92</td>
<td></td>
</tr>
<tr>
<td>Lexical 3-grams</td>
<td>.88</td>
<td>.89</td>
<td></td>
</tr>
<tr>
<td>Syntactic N-grams</td>
<td>.86</td>
<td>.85</td>
<td></td>
</tr>
</tbody>
</table>

5 Conclusion

In this paper, we have presented a new language resource composed of the online discussions related to the encyclopedia articles. Our analysis shows that, compared to other CMC, the language used in WikiTalk is relatively standard and neutral: the vocabulary does not include frequent colloquialisms, affect vocabulary and the typical sentence structures are complete and not e.g. telegraphic as could be imagined in a CMC corpus. The rich meta-date and the clear structuring of the pages (e.g. encyclopedia articles, discussions, noticeboards, etc.) open many further application possibilities for the data. In future work, we aim at benefiting from it to build a web genre detection system by using the typical features associated to different web genres in the Wikipedia corpus and then applying this information to detect similar genres in even larger web corpora.

References


