Problematizing and Addressing the Article-as-Concept Assumption in Wikipedia

Yilun Lin
Northwestern University
Evanston, Illinois
allen.lin@eecs.northwestern.edu

Bowen Yu, Andrew Hall
University of Minnesota
Minneapolis, Minnesota
{bowen, hall}@cs.umn.edu

Brent Hecht
Northwestern University
Evanston, Illinois
bhecht@northwestern.edu

ABSTRACT
Wikipedia-based studies and systems frequently assume that no two articles describe the same concept. However, in this paper, we show that this *article-as-concept assumption* is problematic due to editors' tendency to split articles into *parent articles* and *sub-articles* when articles get too long for readers (e.g. "Portland, Oregon" and "History of Portland, Oregon" in the English Wikipedia). In this paper, we present evidence that this issue can have significant impacts on Wikipedia-based studies and systems and introduce the *sub-article matching problem*. The goal of the sub-article matching problem is to automatically connect sub-articles to parent articles to help Wikipedia-based studies and systems retrieve complete information about a concept. We then describe the first system to address the sub-article matching problem. We show that, using a diverse feature set and standard machine learning techniques, our system can achieve good performance on most of our ground truth datasets, significantly outperforming baseline approaches.

Author Keywords
Wikipedia; peer production; sub-article matching problem

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous;

INTRODUCTION

Over the past decade, Wikipedia has become one of the most valuable datasets for computing research and practice. As an object of analysis in social computing, Wikipedia has shed new light on computer-mediated communication and collaboration (e.g. [28–30,54]) and has helped investigate cultural perspectives in user-generated content (e.g. [11,22,45]), among many other uses. In the artificial intelligence domain, Wikipedia has proven equally beneficial: it helps to power technologies ranging from semantic web engines (e.g. [2,49,55]) to natural language understanding systems (e.g. [16,37,52]).

A key assumption in many Wikipedia-based studies and systems is that there is a one-to-one mapping between a concept and the Wikipedia article that describes the concept. This assumption, which we call the *article-as-concept assumption*, supposes that the entire description of a given concept in a given Wikipedia language edition can be found in a single Wikipedia article. For example, under the article-as-concept assumption, the entirety of the English description of Portland, Oregon should be in the "Portland, Oregon" article, and that article alone.

In this paper, we problematize the article-as-concept assumption and show that while this assumption is valid in many cases, it breaks down for a particular class of high-value concepts: concepts whose articles become too long. The Wikipedia community strongly encourages editors to divide lengthy articles into multiple articles out of a desire to facilitate readability, maximize ease of editing, sustain contributions, and make Wikipedia suitable for diverse technological contexts (e.g. slow connections) [57]. According to these guidelines, the original article (*parent article*) should contain a summary of the concept, and each split-off article (*sub-article*) should have a full treatment of an important subtopic. For example, consider the case of the Portland, Oregon concept. In the English Wikipedia, there is an article titled "Portland, Oregon" which contains a summary of the content about Portland. However, Wikipedia editors have moved more detailed content from this parent article into multiple sub-articles, including "History of Portland, Oregon", "Downtown Portland", "Tourism in Portland, Oregon" (among over a dozen other sub-articles). The sub-articles contain the bulk of the information about the aspects of Portland that they describe, and the parent article only contains brief summaries of this information.

Concepts that violate the article-as-concept assumption attract a significant share of reader interest in Wikipedia. A substantial percentage of the most-viewed articles in large language editions – e.g. "World War II" in English, "Deutschland" in German, "France" in French – have large numbers of sub-articles. Indeed, as we will show below, 71% of the page views to the 1000 most-viewed articles in the English Wikipedia belong to articles that have at least one sub-article. Moreover, among this set of 1000 most-viewed articles, the average number of sub-articles is 7.5.

CSCW '17, February 25-March 01, 2017, Portland, OR, USA
 ISBN 978-1-4503-4335-0/17/03 $15.00
DOI: http://dx.doi.org/10.1145/2998181.2998274

The separation of content about a single concept across multiple articles – especially for high-value concepts – is problematic for a wide swath of Wikipedia-based studies and systems. Consider the case of an intelligent technology that requires a text- or link-based encoding of the concept Portland, Oregon – e.g. one of the many prominent Wikipedia-based semantic relatedness algorithms [16,34,37,46,52] – or a multilingual Wikipedia study that compares the information about Portland contained in different language editions of Wikipedia (e.g. [5,11,18,22,33,36]). In both cases, simply adopting the article-as-concept assumption would result in false conclusions and missed opportunities. For the intelligent technology, large amounts of text in the sub-articles of Portland would be missing from a bag-of-words model (e.g. [16,34]) and relationships between Portland's sub-articles and related concepts would be omitted in any graph-based model (e.g. [21,37]). With regard to the multilingual Wikipedia study, while the English Wikipedia may split content about the Portland, Oregon concept into more than a dozen sub-articles, the Spanish edition may split its corresponding article into four sub-articles, and another language edition another may have no sub-articles at all. To a multilingual study or system that ignores sub-articles, it could appear as if the language editions with the *fewest* sub-articles had the most content about this concept whereas the opposite is likely true.

To address the serious challenges associated with sub-articles and their violation of the article-as-concept assumption, this paper introduces the *sub-article matching problem*. The sub-article matching problem describes the following task: for a potential parent article p in a given language edition, accurately identify all corresponding sub-articles p_s in the same language edition. For instance, solving the sub-article problem involves connecting the parent article "Portland, Oregon" with its "History of Portland, Oregon" and "Portland Fire & Rescue" sub-articles (and others), and repeating for all articles in multilingual Wikipedia. As we will show, this will mean determining whether millions of potential sub-articles are indeed sub-articles of corresponding parent articles.

In addition to defining and motivating the sub-article matching problem, this paper presents the first system that addresses the problem. We collected sub-article ground truth corpora consisting of pairs of parent articles and candidate sub-articles ($<p,p_{cs}>$ pairs) in three languages (English, Spanish, and Chinese). We then used that data to train a model that can achieve 84% classification accuracy on average, outperforming baseline approaches by 17%. Further, we show that our model works best on the articles that attract the most reader interest: it outperforms baseline accuracy by 50% on a dataset consisting only of high-interest articles.

The model's performance is achieved through the use of heterogeneous features ranging from Wikipedia editing practices to the output of semantic relatedness algorithms. Moreover, we were careful to avoid language-specific features in our model, meaning that the model should work equally well in most (if not all) major language editions.

This paper also situates the challenges to the article-as-concept assumption in a broader theoretical framework. Specifically, we discuss how this issue can be interpreted as a human-machine variation of *author-audience mismatch* [26], which was originally conceptualized for human authors and human audiences. We argue that the author-audience mismatch framework – adapted to machine audiences and human authors – is useful for understanding a growing number of problems associated with intelligent technologies' use of semi-structured peer-produced datasets like Wikipedia.

Finally, in the interest of furthering progress on the sub-article matching problem, this paper operationalizes recent calls for open data and open code in the social computing community (e.g. [39,63]). We have made our full sub-article ground truth dataset publicly available and we have released our entire sub-article model as an extension to the WikiBrain Wikipedia software library [47][1]. This will make our model immediately usable by Wikipedia researchers and practitioners. Additionally, including our model in WikiBrain will also allow researchers to make straightforward comparisons with our model's performance and, hopefully, improve upon it.

In summary, this paper makes the following contributions:

1. We identify and problematize the article-as-concept assumption, discuss its risks for Wikipedia-based studies and systems, and show that the assumption fails for a large percentage of high-interest concepts.
2. We introduce the sub-article matching problem, the solution to which is necessary to avoid adopting the article-as-concept assumption in Wikipedia-based studies and systems.
3. We describe the first model to address the sub-article matching problem. We show that our model was able to use a variety of heterogeneous features to achieve performance significantly better than baseline approaches on most ground truth datasets.
4. We have released our model in the form of a software package to enable researchers and developers to immediately begin building systems and running studies that do not adopt the article-as-concept assumption.
5. Additionally, we have provided the first three ground truth datasets for the sub-article problem. The datasets can serve as benchmarks to help the community make

[1] http://z.umn.edu/WikiSubarticles

progress towards a complete solution to the sub-article matching problem.

6. We discuss how the issues with article-as-concept assumption can be understood as a human-machine version of author-audience mismatch, a framework that may describe a growing number of challenges in artificial intelligence and fields related to semi-structured peer-produced datasets

Below, we first address related work. We then discuss our efforts to build reliable ground truth datasets for the sub-article matching problem. Third, we address how we constructed our classification models and interpret model performance. Finally, we highlight the theoretical implications of our work associated with author-audience mismatch.

RELATED WORK

The primary motivation for this work arises out of research that has implicitly adopted the article-as-concept assumption. This research occurs in many areas of human-computer interaction, artificial intelligence, and related fields and can broadly be divided into work that (a) studies Wikipedia, (b) leverages Wikipedia to seed "knowledge graph"-like structures and (c) utilizes Wikipedia as a corpus for intelligent technologies. In the following, we summarize these three bodies of literature and discuss how addressing the sub-article matching problem could lead to improvements in each of them.

Wikipedia-based Studies

The article-as-concept assumption applies to the many studies which suppose that all communication and collaboration around a concept is associated with a single article (in a given Wikipedia language edition). One large area of Wikipedia-related literature for which the article-as-concept assumption is particularly problematic is the work that examines the similarities, differences, and interactions between the different language editions of Wikipedia (e.g. [1,5,11,18,22,33,36,53,62]). In this line of research, under the prevailing article-as-concept assumption, sub-articles are mistakenly treated as separate concepts. This can result in mistaken conclusions about a key variable of interest in this literature: the similarities and differences in the articles about the same concept in different language editions, e.g. "United States" (English) vs. "Estados Unidos" (Spanish) vs. "Stati Uniti d'America" (Italian); or "Portland, Oregon" (English) vs. "波特蘭_(俄勒岡州)" (Chinese).

Indeed, the most direct motivation for the research in this paper emerges from a line of work in the multilingual Wikipedia literature that seeks to surface cross-language similarities and differences. Systems associated with this line of work – e.g. Omnipedia [5] and Manypedia [36] – generally leverage visualization strategies to allow users to simultaneously view information about a concept of interest from multiple language editions. However, if a user is interested in a concept associated with sub-articles (e.g. World War II or Portland), these systems *can only show content from the parent articles* because they are unaware of the sub-articles. As a result, these systems will give inaccurate results when comparing a concept's content across language editions, which partially undermines a key goal of these systems. Additionally, these systems also often omit large amounts of information from language editions that have sub-articles about the concept of interest, as the length of sub-articles in aggregate often dwarves that of the parent article. Finally, exacerbating the situation, these systems are user-driven, so many of the queries are to popular concepts, which tend to have more sub-articles.

An unpublished release of the Omnipedia system recognized this problem and allowed users to manually specify sub-articles [20]. Our goal with this paper is to make this process automatic as well as generalizable to any Wikipedia-based system or study, not just Omnipedia. Indeed, by leveraging the WikiBrain package we developed in this project, Omnipedia and any similar system could immediately become sub-article-aware in an automated fashion.

Knowledge Graph-like Repositories

The article-as-concept assumption is also problematic for recent high-profile efforts to integrate Wikipedia into "Knowledge Graph"-like structures [49] (e.g. [24,48]), most notably those associated with the Wikimedia Foundation's newest project, Wikidata [55]. Broadly speaking, Wikidata consists of items that correspond to Wikipedia articles connected via semantically-labeled properties. Wikidata has become a critical resource for many intelligent technologies (e.g. [14,40,62]), which potentially makes the article-as-concept assumption more problematic.

Wikidata's reliance on the article-as-concept assumption dates back to its launch in 2012, when it used Wikipedia articles to seed its corpus of concepts. As a result of this approach, sub-articles are considered to be separate concepts relative to their parent articles. In other words, in Wikidata, "History of Portland" (English) and "Portland" (English) are treated as describing entirely different concepts, as is the case for hundreds of thousands of other parent article/sub-article relationships.

The problems associated with the article-as-concept assumption in Wikidata are quite apparent in the first large technology to use Wikidata information: Wikipedia itself. For instance, Wikidata is now the backend for the "Other languages" links in the Wikipedia sidebar. These links refer readers to other articles about the same concept in different languages. Because of the article-as-concept assumption in Wikidata, a reader of the "History of Portland, Oregon" (English) article (or a study using the Wikidata's connections between language editions) will not be exposed to the large "History" section in the German article on "Portland (Oregon)" that happens to not be split into a separate article. The same problems are occurring with other Wikidata integrations in Wikipedia. For instance, the Wikipedia

templates that now draw information from Wikidata will struggle to look up information about a single concept that is split across multiple Wikidata items.

This situation could potentially be addressed with the "facet of" Wikidata property that was introduced in late 2014. However, owing to the large amount of crowdsourced labor necessary to propagate this property to all appropriate concept pairs, this property is largely unused outside of a series of automatically-added temporal "facet of" relationships (e.g. connecting "2003 in film" to "film"). Indeed, none of the sub-articles for "Portland, Oregon" – e.g. "History of Portland, Oregon", "Tourism in Portland, Oregon", "List of Notable People from Portland, Oregon" – are connected to the Portland, Oregon Wikidata item by the "facet of" property.

One potential application of our work is leveraging the open-source software package we have released to power a Wikidata editing bot to help propagate "facet of" properties to more Wikidata items. In addition, our software package can also be used to help Wikidata-based applications dynamically integrate information from parent articles and sub-articles.

Wikipedia-based Intelligent Technologies

A large number of Wikipedia-based intelligent technologies adopt the article-as-concept assumption, usually doing so implicitly by integrating the assumption into core data structures and techniques. While the manner in which this is done varies from technology to technology, a useful case study comes from the Milne-Witten semantic relatedness algorithm [37], an influential Wikipedia-based contribution to natural language processing. This algorithm assesses the relatedness between two concepts by mapping them to Wikipedia articles and comparing the sets of articles that link to each of these articles (i.e. evaluating inlink overlap). However, since the algorithm treats "History of Portland, Oregon" (English) and "Portland, Oregon" (English) as describing entirely different concepts – as it does for all parent/sub-article pairs – Milne-Witten misses potentially valuable relatedness signals when comparing sets of in-linking articles (e.g. it would not consider inlinks to "History of Portland, Oregon" when assessing relatedness to the concept of Portland). Similar situations occur with Explicit Semantic Analysis [16] and other prominent semantic relatedness measures. Moreover, as is the case with Wikipedia-based studies, it is likely that multi-lingual Wikipedia-based technologies (e.g. [34,46]) will be more affected than single-language technologies due to the diversity of sub-article relationships across language editions.

It should be relatively straightforward to address these issues by integrating our sub-article matching problem model into approaches like Milne-Witten, ESA, and others. For instance, in the Milne-Witten case, preprocessing to effectively append all sub-articles to their parent articles would solve the problem.

DATASET DEVELOPMENT

Wikipedia Corpus

We downloaded XML Wikipedia data dumps in January 2015 and processed these dumps using WikiBrain [47], which is a Java software framework that provides access to a range of Wikipedia datasets and Wikipedia-based algorithms (including the semantic relatedness algorithms we use below). We focused on three language editions: Chinese, English and Spanish. These were selected because they (1) are widely spoken and (2) span the East/West spectrum, which has proven to be an important consideration across a large body of HCI literature (e.g. [5,22,25,35]). We additionally processed 22 other language editions as we required additional language editions to operationalize language-neutral features (see below). Overall, our Wikipedia datasets contains the 25 language editions with the most articles (as of January, 2015), excluding largely bot-generated editions like Cebuano and Waray-Waray. It is important to note that, in addition to utilizing WikiBrain, we also contributed back to WikiBrain by developing a sub-article matching problem extension.

Sub-article Candidates

An essential dataset in our study is our large sub-article candidate dataset. In this sub-section, we first define sub-article candidates and then describe how we mine them from the Wikipedia corpora described above.

Defining Sub-article Candidates

Without additional information, any Wikipedia article could be a sub-article of any other Wikipedia article in the same language edition. As such, a potentially intractable computational problem emerges. For instance, with its over 5,210,165 articles, the English Wikipedia alone would require examining more than 25 trillion potential sub-article relationships (5210165 * 5210165 - 5210165).

Fortunately, Wikipedia editors use a number of indicators that allow us to prune away a huge portion of these potential sub-article relationships *a priori*, resulting in a much smaller (but still large) group of *sub-article candidates*. Editors employ these indicators – which vary from language edition to language edition – to highlight for readers when content from a parent article has been split off into a sub-article.

To identify indicators for sub-article candidates, a single investigator fluent in English and Spanish accessed thousands of pages in all 25 language editions in our corpus, focusing on concepts that typically had sub-articles in many language editions (e.g. countries, major historical events, large cities). Although context is usually sufficient to identify a sub-article relationship, the investigator used Google Translate as an aid when necessary. Whenever the investigator encountered a potential sub-article relationship, he recorded the parent article (e.g. "Portland, Oregon"), the potential sub-article (e.g. "History of Portland, Oregon"), and, most importantly, the Wiki markup that was used to encode the relationship (Wiki markup is the markup language used by Wikipedia editors to write articles). The

Figure 1. An example of the ambiguity inherent in sub-article indicators, in this case the "{{main article}}" template in the English Wikipedia. Links in red indicate potential sub-article relationships.

final dataset consists of 3,083 such records, and is included in the release of our ground truth dataset.

Using the above process, we identified two general types of sub-article indicators. The first type is the *template-based* indicator that resembles the appearance of Figure 1, although the specific markup and prompt varies within and across languages (e.g. the "{{main article}}" template in Spanish is "{{AP}}" for *artículo principal*, and similar English templates include {{see also}} and {{further}}). Templates in Wikipedia are a type of wiki markup that editors can use to generate complex HTML just by entering a few parameters, which is illustrated in Figure 1.

The second type of indicator is significantly different. Potential sub-article relationships encoded through this type of indicator are listed at the bottom of articles under a header titled "See also" or its equivalent in other languages. Interestingly, using *see also-based* indicators for sub-articles is explicitly contrary to community guidelines in some language editions (e.g. the English Wikipedia's Manual of Style, which state that "See also" sections should be reserved for links to peripherally-related content). However, our candidate dataset reveals that editors frequently violate these guidelines (e.g. Figure 2).

More generally, while both template-based and see-also based indicators are often used to indicate sub-article relationships, they are also used for a variety of other purposes, causing large numbers of false positives to emerge. Figure 1 illustrates this phenomenon with an example highlighting one of the more common sub-article indicators in the English Wikipedia: the {{main article}} template-based indicator. The top of Figure 1 shows a situation in which this indicator is used to demarcate a true sub-article relationship ("Portland, Oregon" and "Sports in Portland, Oregon"), but the bottom shows a situation in which this is clearly not the case ("Caffeine" and History of chocolate",

Figure 2. The "See also" section of the English article about the Canadian territory of Nunavut. Some links here are clearly sub-articles, while others are more distantly related to the concept of Nunavut (e.g. "Arctic policy of Canada").

"History of coffee", "History of tea", and "History of yerba mate").

To summarize, although sub-article indicators like "{{main article}}" are ambiguous, mining them from the article text of each language edition is an essential pre-processing step. This is because (1) they can be considered to be a broad superset of all sub-article relationships and (2) they prevent us from having to compare all articles in every language edition in a pairwise fashion (a potentially intractable brute force approach). Below, we describe the procedures we use to execute this multi-faceted mining process.

Mining Sub-article Candidates

After developing our dataset of sub-article indicators, we used these indicators to write a script that parsed out all sub-article candidates across all 25 languages. In most cases, this script utilizes straightforward regular expressions, although other cases were more complicated. Our script is included in our open-source WikiBrain sub-article software library.

A quick examination of this dataset was our first indication that separating the signal (true sub-articles) from the noise (false sub-articles) will be difficult, even among the much narrower class of sub-article candidates. We originally expected that many sub-articles would follow the pattern "{something} of {parent article}" such as "Geography of the United States" (sub-article of "United States"), and the equivalent in each of the 25 languages we considered (e.g. "Geografía de Estados Unidos" in Spanish). However, it became clear in this preliminary dataset that a significant portion of sub-articles violate this pattern. For instance, this preliminary dataset contains potential sub-article relationships between parent articles p and candidate sub-articles p_{cs} such as p = "List of Chinese Inventions" and p_{cs} = "Four Great Inventions", p = "United States" and p_{cs} = "American Literature" and p = "The Silmarillion" and p_{cs} = "Valaquenta" (all from the English Wikipedia.)

Overall, we found sub-article candidates for a substantial proportion of Wikipedia articles. For instance, over a quarter of articles in English and Spanish Wikipedia contained sub-article candidates. More details about these percentages, including the share of template-based and see also-*based* indicators, is available in Table 1.

	English	Chinese	Spanish
% of articles (with templates + see also section)	20.5%	11.6%	25.2%
% of articles (with templates)	4.9%	2.3%	7.7%
% of pageviews (with template)	24.7%	11.9%	25.6%

Table 1. The percent of articles and page views associated with potential sub-articles.

Ground Truth Datasets

As is typical in many types of machine learning, our sub-article models require extensive ground truth data for both training and testing. Because no prior work has defined, let alone attempted to solve, the sub-article matching problem, it was necessary to both generate our own ground truth datasets as well as to define their character and structure. By developing these datasets and making them publicly available, we also hope to make it easier for other researchers to build on our results.

In this sub-section, we first define the high-level structure of our ground truth datasets, focusing on how they were sampled from the larger set of overall sub-article candidates. We then describe how we labeled each potential sub-article relationship in these datasets. This is an important process that, as is often the case when developing the first ground truth dataset for a new problem, led to a formalization of the definition of sub-articles.

Sampling and Structure: Selecting Sub-article Candidates

We generated three ground truth datasets, each using a different strategy to sample from the sub-article candidate population. All three datasets consist of a series of $<p, p_{cs}>$ pairs (i.e. <parent article, potential sub-article> pairs). For each dataset, we randomly selected 200 pairs from English, and 100 each from Spanish and Chinese. Each of the three sampling strategies was designed to maximize ecological validity for a different class of sub-article matching use cases. These strategies are described in more detail below:

`High-Interest`: This dataset consists of $< p, p_{cs} >$ pairs whose parent articles are sampled from the 1,000 most-viewed articles for a given language edition. We gathered page view data from the Wikimedia's Page View API[2] and aggregated it from August 2015 to February 2016. This is a user-centered dataset that is driven by the actual demand for Wikipedia content. Performance on this dataset will be a good proxy for the model's performance on most real-life systems, especially those that are directly user-facing (e.g. Omnipedia [5], Manypedia [36], etc.). This means that `High-Interest` is likely the most important of the three datasets.

`Random`: The parent articles in this dataset were sampled randomly from all articles in each language edition. Corresponding sub-article candidates were then randomly selected from the set of candidates available for each parent article (potential parent articles without sub-articles were ignored). Given the long-tail quality distribution of Wikipedia articles [58], this dataset includes large numbers of short, low-quality articles. It also contains many articles that are of very limited reader interest (relatively speaking). This dataset will give a lower-level understanding of performance across entire language editions.

`Ad-Hoc`: This dataset was generated by sampling from the 3,083 $<p, p_{cs}>$ pairs generated during the indicator identification process, focusing only on candidates from English, Spanish, and Chinese. This was the first dataset we generated, and it served as a successful feasibility test. It also provides a useful triangulation of our models' performance relative to the other two ground truth datasets.

Labeling: Evaluating Sub-article Candidates

The Wikipedia community has no precise formal definition for what separates a "true" sub-article from a "false" one. This is because the sub-article construct was invented for human readers, and human readers do not need such binary distinctions: if a human is interested in a link, s/he clicks on it, regardless of whether the link is a sub-article or not. (We return to this issue when covering audience-author mismatch in the Discussion section.)

Wikipedia-based studies and systems, on the other hand, must make this binary distinction, and must do so frequently and explicitly. As such, a major challenge becomes finding a way to codify the definition of a sub-article relationship in our ground truth datasets in a fashion that Wikipedia-based systems and studies can understand, while at the same time respecting the variability of sub-article relationships encoded by human Wikipedia editors.

To address this challenge, we adopted the approach of past Wikipedia research that has also sought to codify fluid Wikipedia constructs (e.g. [6]). Specifically, we utilized a two-part method that allows the researcher or system designer to decide how broadly or narrowly they want to define the sub-article construct, according to the needs of their application or study. The first step in this process involved coding potential sub-article relationships on an ordinal spectrum in recognition of the non-binary nature of the sub-article relationships. We then proposed several reasonable thresholds and examined our results using each of these thresholds. This multi-faceted approach allows us to understand the performance of our models at each breadth level and allows the users of our system to flexibly choose from broader and stricter definitions of sub-articles.

With regard to the ordinal coding stage, for each language of each dataset (English, Spanish, Chinese), we recruited two

[2] https://wikimedia.org/api/rest_v1/

coders who were fluent in the corresponding language. Coders were asked to assign each <p, p_{cs}> a code along an ordinal scale from 0 (definitely not a sub-article) to 3 (definitely a sub-article), with each code being defined as follows[3]:

- **3**: The *only* reason the sub-article candidate exists is to split the corresponding parent article into more manageable subtopics. The potential sub-article really *does not deserve its own page*, and the corresponding parent article is the best place to put the sub-article's content.
- **2**: Same as above, but the topic of the sub-article candidate is significant enough to warrant its own page.
- **1**: The sub-article candidate contains information that would be useful to have in the parent article, but also contains its own, *unrelated (non-overlapping) content.*
- **0**: The sub-article candidate is about a topic that is *trivially related* to the parent article or has a large amount of *non-overlapping content*.

The inter-rater reliability on our datasets as computed by Weighted Cohen's Kappa [12] ranged from 0.56 to 0.78, which is considered a "moderate" to "substantial" agreement [32]. We used Weighted Cohen's Kappa since it is the most appropriate for our ordinal codes [3].

After examining our ground truth ratings data, we determined three reasonable thresholds that researchers and practitioners may want to use to separate sub-articles from non-sub-articles. The strictest definition requires an average score of 3.0 from two coders, meaning that both gave the relationship a '3'. Next, we considered a threshold at an average rating of 2.5, which is more flexible but still required one coder to give a candidate relationship a '3' and the other to give it a '2'. Finally, we also considered an average score of 2.0 as a threshold, which is the broadest definition and can come from various rating configurations.

MODELING

Overview

The goal of our modeling exercise was to accurately address the sub-article matching problem using machine learning techniques. In other words, our aim was to build classification models that can accurately predict whether a parent article/sub-article candidate pair <p, p_{cs}> represents a true sub-article relationship (i.e. a <p, p_s>). As we discussed above, we defined this classification problem along three dimensions: (1) dataset {`high-interest, random, ad-hoc`}, (2) sub-article definition/threshold {average rating = 2.0, 2.5, 3.0} and (3) language {English, Spanish, Chinese}. Our machine learning experiments, described below, allow us to assess our models' performance along each dimension.

We experimented with a variety of well-known machine learning algorithms including SVM, Random Forest, Decision Tree, Naïve Bayes, Logistic Regression, and Adaboost. In the body of this section, we report results from the algorithm with the best performance. Our emphasis here is to demonstrate that one can successfully address the sub-article matching problem using popular machine learning algorithms instead of providing a detailed performance analysis of each specific algorithm. However, we supplement this discussion with detailed classification accuracies for each machine learning algorithm and dataset configuration in Appendix A. Interestingly, in most situations, the choice of the best machine learning algorithm is consistent across languages and definitions of sub-articles within a given dataset type (i.e. `high-interest, random or ad-hoc`).

For evaluation, we followed standard practice [7] and conducted 10-fold cross validation, reporting the average accuracy across all folds. Because this paper is the first to define and attempt to solve the sub-article matching problem, we cannot compare our models' accuracies with any prior work. This is a relatively common situation when applying machine learning in HCI research [23,43]. When it occurs, the best practice is to one compares one's results to straightforward baseline approaches (e.g. [19,44,61]). In this paper, we utilized the baseline approach that we found to be most powerful: always predicting the most frequent label in the training set.

Because our models are only as powerful as the features they leverage, before describing our results, we first describe each feature in detail. The features we use are diverse, drawing from techniques ranging from simple syntax comparisons, to network metrics, to advanced natural language processing algorithms. However, nearly all of our features have one key property in common: language neutrality. This means that they can be utilized to help predict whether a sub-article candidate is really a sub-article of a given parent article, regardless of the language edition of the parent and candidate.

A frequent technique we use to make a feature that would be otherwise language-specific into one that is language neutral is immediately converting language-specific <p, p_{cs}> pairs to language-neutral concepts using Wikidata cross-language mappings. For instance, comparing the number of characters or tokens shared by the parent and sub-article candidate article titles is a feature whose output and effectiveness varies extensively across language editions (i.e. it is more useful in Western languages than Eastern languages). However, by using cross-language mappings, when considering articles from Eastern language editions, our models can take advantage of the power of this approach in Western languages by examining the titles of the equivalent

[3] Our full coding instructions and codebook are included in the code repository, linked above.

concepts in English, Spanish, and so on. Our typical approach to implementing this at scale is to calculate the value of each feature in a language-specific fashion for all articles about the same concepts as the input <p, p_{cs}> pair. We then aggregate the output of these values, e.g. using maximums, averages, ratios, or summaries.

Features

PotSubLangsRatio

Since assigning sub-article indicators is a manual process, we expect that if an article is a sub-article candidate for a given parent article in many different language editions, this will increase the likelihood that the candidate is a true sub-article. For instance, "History of the United States" (English) is a sub-article candidate (as indicated by the "{{main article}}" template) of "United States" (English), and the same relation is true for their corresponding articles in Spanish Wikipedia (with the {{AP}} template). We operationalize this feature by calculating the ratio between the number of languages in which there is a potential sub-article relationship and the number of languages in which the parent articles and sub-article candidate both have corresponding articles.

MaxTokenOverlap

This feature focuses on article titles only and considers the percentage of tokens in the parent article's title contained within the sub-article candidate's title. It takes the maximum token overlap of the equivalent articles in all languages. A high value signifies that the parent article and sub-article candidate share a large portion of words in their titles (in at least one language) and we hypothesized that this would represent a higher-likelihood sub-article relationship. When computing this feature, we tokenized Eastern languages that are written in a *scriptio continua* pattern (no spacing or other dividers) and to match characters between Traditional Chinese and Simplified Chinese.

NumLangsRatio

This feature measures the relative "globalness" of the parent article and the sub-article candidate across all 25 languages. It is computed as the number of language editions in which the parent article has foreign language equivalents, divided by the same number for the sub-article. For instance, for the <"Portland, Oregon", "Sports in Portland, Oregon"> pair in English Wikipedia, "Portland, Oregon" has corresponding articles in all 25 languages while "Sports in Portland. Oregon" only has an article in the English Wikipedia. We hypothesized that a higher value would indicate a higher likelihood of a parent/sub-article relationship because other languages might not yet have split the relevant content into two articles.

MainTemplatePct

This feature leverages one of the most prominent sub-article indicators: the main template. Main templates can be seen in Figure 1, and all 25 language editions have a version of this template. Although ambiguously used in many cases, we hypothesized that main templates had the highest precision of all the indicators. Moreover, in many languages, the usage guide for this template corresponds well with the notion of sub-articles (e.g. [59]). We calculated this feature as follows: the number of language editions in which the sub-article candidate appears in parent article's main template divided by the number of language editions in which there is any sub-article indicator between the two articles. In other words, the feature is the share of the potential sub-article relationships between two concepts defined using a main template.

MaxSectionTokenOverlap

This feature specifically considers the template-based sub-article indicators. Note that in Figure 1, these indicators almost always appear below a section sub-head. This feature is the direct analogue to *MaxTokenOverlap*, but uses the title of the preceding section sub-head rather than the titles of the articles.

MaxMainTFInSub

In all language editions, most Wikipedia articles begin with a summary of the content of the article. This feature calculates the term frequency of the parent article's title in the summary paragraph of the sub-article candidate and takes the maximum across all languages. We hypothesized that as part of the natural editing process, when editors spin off a sub-article, they refer back to the parent article in the introduction. As such, we expected a higher value would lead to a higher likelihood of a sub-article relationship.

IndegreeRatio

This feature describes the relative centrality of the parent article and sub-article candidate in the article graph of a given language edition. We hypothesized that true sub-article relationships would more often involve a central/important parent and a less central/important sub-article than vice versa. This feature is calculated by taking the ratio of the indegree of the parent article (i.e. the number of Wikipedia articles that contain a hyperlink to this article) and the indegree of the sub-article, each of which is summed across all languages. Indegree is commonly used as a straightforward metric of network centrality/importance in large graphs like the Wikipedia article graph [27].

MilneWitten

This feature is the MilneWitten semantic relatedness (SR) measurement [37,38] between the parent article and sub-article candidate. We hypothesized that a higher SR between the two articles would mean that these two articles are more likely to be in true sub-article relationship. For example, in the English Wikipedia, "History of chocolate" and "Caffeine" are less related than "Sports in Portland, Oregon" and "Portland, Oregon".

Other Features

Besides the features described above, we also tested features that consider the structural complexity of p and p_{cs}. For example, the ratio between the number of templates in a

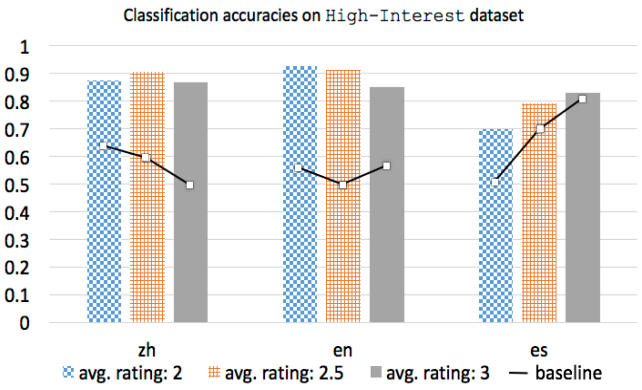

Figure 3.1

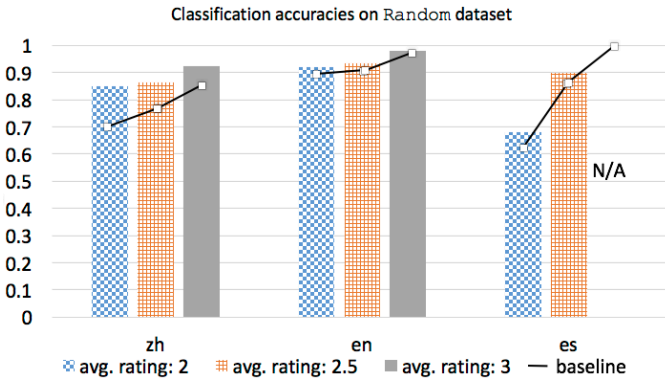

Figure 3.2

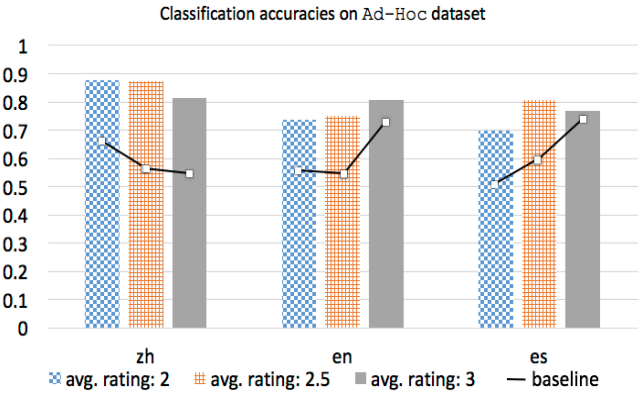

Figure 3.3

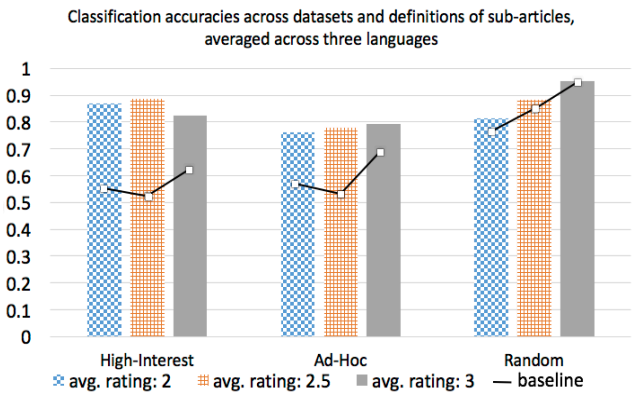

Figure 3.4

Figure 3.1 – 3.4. Classification accuracies across datasets, language editions and different thresholds of sub-article ratings. Each colored vertical bar shows the best accuracies among the machine learning algorithms considered, and the black line indicates baseline performance.

parent article and candidate sub-article and the ratio between the number of references in a parent article and a candidate sub-article. These features provided only a marginal improvement to the classification accuracy. With parsimony in mind, we did not include them in the final model construction.

Results

High-Interest Dataset

Figure 3.1 shows the ability of our model to correctly distinguish true sub-article relationships from false ones in the `High-Interest` ground truth dataset according to all three definitions of sub-articles (2.0, 2.5, 3.0). Recall that the `High-Interest` dataset focuses on high-demand concepts that are frequently accessed by Wikipedia readers, making our model's results on this dataset particularly important. For simplicity, although we tried the multiple machine learning algorithms described above, we only report the one with the highest accuracy in the figure (more results are available in Appendix A). For `High-Interest`, Linear SVM and Random Forest alternated as the best techniques.

As can be seen in Figure 3.1, in all cases, our models outperformed the baseline method by a substantial margin. On average, our model exceeded baseline performance by a factor of 1.45, and this number went up to around 1.8 in specific cases. The highest absolute accuracy was 90% (English, 2.5-average) and the lowest absolute accuracy was 70% (Spanish, 2.0-average). Overall, Figure 3 shows that for concepts that are in high demand, the features we have defined make it a relative straightforward task for a learned classification model to determine whether a sub-article candidate is a "true" sub-article.

Random Dataset

Figure 3.2 shows the classification results on the `Random` dataset. Immediately visible in the figure is that the baseline accuracies for this dataset are much higher than those for the other datasets. In the most extreme case – the 3.0-threshold Spanish dataset – the baseline accuracy reaches 100%, which makes the classification results meaningless in the context of this study. These higher baselines occur because in this dataset, the vast majority of sub-article candidates are negative examples. The necessary result of this situation is that a baseline approach that always guesses 'no' will be

Feature	Average Importance Rank
MaxMainTFInSub	2.7
MaxTokenOverlap	2.8
MaxSectionTokenOverlap	3.7
NumLangsRatio	3.8
PotSubLangsRatio	4.5
MilneWitten	5.4
IndegreeRatio	5.6
MainTemplatePct	7.3

Table 2. Feature importance averaged across all languages and all definitions of sub-articles on the `High-Interest` dataset. Feature importance is computed using a Random Forest model.

accurate most of the time, and this is a tough baseline for any algorithm to beat.

Indeed, while their absolute accuracies are quite high, our models in general only marginally improve upon the baselines with respect to this dataset. The highest classification accuracy relative to the baseline is on Chinese sub-article candidates with the 2.0-average threshold. English with a 3.0-average threshold was the only configuration in which the classifier failed to improve upon the baseline at all.

Ad-hoc Dataset

Figure 3.3 shows the classification accuracies on the `Ad-Hoc` dataset. Among the machine learning algorithms, Linear SVM and Random Forest alternate as the best algorithm across different languages and definitions of sub-articles. While absolute accuracy levels on this dataset are comparable with those for `High-Interest`, the baseline accuracies are generally higher. This means that a smaller improvement was made by our models on this dataset relative to the baseline.

Summary of Results

Figure 3.4 compares the average classification performance for each dataset across all sub-article thresholds and languages. This figure further reinforces several high-level trends mentioned above:

- On the `High-Interest` dataset that contains articles in high-demand by readers and the `Ad-Hoc` dataset that is sampled from more meaningful concepts, our classification results outperform the baseline consistently and substantially across all three languages (See Figure 3.1, 3.3 and 3.4).
- On the `Random` dataset that contains articles typically of lower interest, shorter length, and lower quality, our models generally do not make a substantial improvement compared to the baseline method (See Figure 3.2 and 3.4).

Feature Analysis

In order to understand which features were the most important to the success of our models, we examined the Random Forest versions of our models. These models have the advantage of (1) being the top or close to the top performing models in most configurations of our experiments (see Appendix A) and (2) they afford straightforward analysis of feature importance. Specifically, in Random Forest models, feature importance can be evaluated by adding up the weighted impurity decrease for all trees in the forest using an impurity function such as the Gini index or Shannon entropy [8,9]. This approach has been used for feature selection in various domains including but not limited to bioinformatics [50,51], image classification [17], and ecology [13].

Table 2 presents the importance rank for each feature on the `High-Interest` dataset (averaged across language and sub-article definition). Note that *MainTemplatePct* is the *least* important feature. Recall that this feature is motivated by the fact that it reflects community-defined rules for linking parent articles and sub-articles. As such, we originally expected *MainTemplatePct* to be a strong predictor of true sub-article relationships. However, even though we account for human error to some extent (by aggregating across all languages), *MainTemplatePct* remains relatively non-predictive. Closer examination of the data reveals that although these guidelines are properly documented, human editors failed to consistently follow the guidelines. For example, in the "Donald Trump" (English) article, which is one of the top ten most-viewed English articles in our page view data, Wikipedia editors correctly tagged the true sub-article "Donald Trump presidential campaign, 2016" with the template {{main article}} while they incorrectly tagged the true sub-article "Donald Trump presidential campaign, 2000" with the template {{see also}}.

Table 2 also shows that *MaxSectionTokenOverlap, MaxTokenOverlap,* and *MaxMainTFInSub* are relatively important features. Unlike *MainTemplatePct,* which relies on editors explicitly indicating sub-article relationships as defined by community guidelines, these natural language features are *implicit*: they capture the lexical, linguistic and semantic relationships between parent articles and sub-articles that emerge through the natural editing process. For instance, when Wikipedia editors move content to a new sub-article, it is natural for them to add a lead sentence that points back to the parent article [60]. This natural editing process is captured by *MaxMainTFInSub.* We believe that the variation in effectiveness between the explicit standard-based features and the implicit features may point to author-audience mismatch issues, which will be detailed in the Discussion section.

The Impact of the Article-as-Concept Assumption

Our trained models allow us to quantify the impact of the article-as-concept assumption. Specifically, the models

Impact Metric	Statistics
% of articles w/ sub-articles	70.8%
% of page views to articles w/ sub-articles	71.0%
Avg # of sub-article per article	7.5

Table 3. The impact of applying our model on the top 1000 most-viewed articles in English Wikipedia.

allow us to ask: for how many articles and page views is the article-as-concept assumption invalid due to sub-articles?

To address this question, we deployed our `High-Interest, 2.5`-average threshold model on the 1,000 most-viewed articles in English Wikipedia. Table 3, which contains the results of this analysis, shows that sub-articles cause violations of the article-as-concept assumption in a large percentage of cases. For instance, over 70% page views to this set of critical articles go to articles that contain at least one sub-article. Table 3 also reveals that on average, each of the top 1000 English Wikipedia articles has 7.5 sub-articles.

This result has important implications for user-centric Wikipedia-based technologies such as Omnipedia, Manypedia, and others. Based on the findings in Table 3, designers of these technologies should assume that users will frequently engage with articles that have sub-articles. Indeed, it appears that at least for the most popular articles, sub-articles are not the exception, they are the rule.

DISCUSSION

Audience-Author Mismatch

Immediately above, we showed that the article-as-concept assumption, which is central to many Wikipedia-based studies and systems, fundamentally breaks down for a substantial proportion of high-value Wikipedia articles. In this section, we describe how this issue may be just one in a growing series of problems for Wikipedia-based studies and systems associated with *author-audience mismatch* [26].

The author-audience mismatch framework was originally intended to explain problems associated with human authors failing to sufficiently customize their content for a given audience (e.g. in a cross-cultural context). However, this framework may also be helpful for understanding the root cause of the article-as-concept assumption and its resultant problems. Namely, Wikipedia editors write Wikipedia articles for the needs of human audiences, but increasingly, Wikipedia has two additional audiences as well: Wikipedia-based studies and, in particular, Wikipedia-based systems. These audiences often have fundamentally different needs than human audiences.

It is important for Wikipedia's studies and systems audience that all content about a concept in a given language edition be contained in a single article. However, for Wikipedia editors' intended audience – other humans – doing so would violate Wikipedia's guidelines to break up long articles into parent and sub-articles. These guidelines emerge from clear human needs. For instance, lengthy articles take a long time to load with the slow Internet connections used by many Wikipedia readers [57]. Additionally, long-standing notions of web usability clearly establish that lengthy web pages result in poor user experiences [10,41,42].

The tension between the needs of human audiences and those of Wikipedia studies and systems is exacerbated by a few additional findings in this paper. In our analysis of feature performance, we identified that while Wikipedia has developed a number of structured mechanisms to link sub-articles to their corresponding parent articles (e.g. {{main article}} templates), they are misused so extensively that our models found them only minimally helpful when separating true sub-article relationships from false ones. We also observed the inverse. For instance, structured constructs like "See also" sections are not supposed to be used for strong relationships like those between parent articles and sub-articles, but editors use them this way anyway. It is not hard to see why these problems have emerged: human authors know that their human audience can click on a sub-article link if they are interested, regardless of whether it is properly encoded.

Author-audience mismatch problems may create major challenges for machines and studies beyond the article-as-concept assumption, and even in peer production datasets other than Wikipedia. For example, consider the tagging infrastructure in OpenStreetMap (OSM) – the "Wikipedia of Maps" [15]. OpenStreetMap has a robust and well-defined set of tagging practices, but recent evidence suggests that OSM editors do not follow standards when they are not perceived as necessary for their specific human audience [31]. For instance, if a human has to decide between tagging a restaurant that serves both coffee and donuts with either "coffee shop" or "donut shop", it is unlikely they will spend time reading and strictly following the guidelines. Instead, the most likely thought process is "I'll choose one and people know that they can probably get coffee and donuts at both types of places." However, for an OSM-based location-based recommender system using this information as training data (e.g. [4]), this is a potentially serious source of noise if it generalizes across many tags.

Returning to Wikipedia, another example comes from how different language editions treat synonyms. Recent research by Wulczyn et al. [62] has found that different language editions opt to create articles for different synonyms of the same concept. For instance, the English Wikipedia decided to create article only for "Neoplasm" while German Wikipedia chose to create article only for "Tumor". While this may suit the human audiences of each individual language edition, it requires additional intelligence for a machine to establish a linkage.

Limitations and Future Work

While we were able to address the sub-article matching problem with good accuracy for most datasets, our solution has a few important limitations that serve as starting points

for future work. First, our models were trained only on potential sub-article relationships that are explicitly encoded by Wikipedia editors. We believe this approach is appropriate as it respects the decisions of editors and almost certainly captures the vast majority of sub-article candidates. That said, it would be interesting to try to discover implicit sub-articles in an automated fashion. A system that can execute implicit sub-article discovery successfully may be useful to Wikipedia editors [56] in addition to system-builders and researchers who work with Wikipedia.

Another limitation is that our models are trained on only a small set of data from just three language editions. There are possible nuances in sub-article usage that might be missed with this limited view. A third limitation is that although our models work well on the articles that attract the most reader interest, they fail to work equally well on the large number of articles that are of lower quality and shorter length. Future work should involve designing a model focused on these articles.

Finally, since we made our features language-independent by aggregating over all language editions, we observed some scalability problems for structurally complex articles. For example, feature generation for the "United States" (English) page, which contains over 200 potential sub-articles, can take minutes. While this problem can be easily addressed by caching the classification results, future work may want to improve the scalability of our approach to reduce pre-processing and updating time.

CONCLUSION

In this paper, we identified and problematized the article-as-concept assumption that is widely adopted in Wikipedia-based studies and systems. We showed that this issue will impact the performance and accuracy of these studies and systems for a large percentage of high-interest Wikipedia articles, and we formulated the sub-article matching problem as a way to mitigate this situation. By developing models that draw on a diverse feature set, we addressed the sub-article matching problem with relatively high accuracy, especially for the Wikipedia articles that attract the most attention. Finally, in order to help researchers immediately address the sub-article matching problem in their own systems and studies and push this line of research forward, we have made our model and our gold standard sub-article datasets freely available for download[4].

ACKNOWLEDGMENTS

The authors would like to thank Stephanie Hernandez, Geovanna Hinojoza, Federico Peredes, Stephanie Hecht, Patti Bao, and Darren Gergle for their valuable contributions to this project. This project was funded by NSF IIS-1552955, NSF IIS-1526988, and NSF IIS-1421655.

[4] http://z.umn.edu/WikiSubarticles

REFERENCES

1. Sisay F. Adafre and Maarten de Rijke. 2006. Finding Similar Sentences Across Multiple Languages in Wikipedia. 62–69.
2. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web*, Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber and Philippe Cudré-Mauroux (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 722–735. Retrieved February 16, 2012 from http://www.springerlink.com/content/rm32474088w54378/
3. Roger Bakeman and John M. Gottman. 1997. *Observing Interaction: An Introduction to Sequential Analysis*. Cambridge University Press.
4. Andrea Ballatore, Gavin McArdle, Caitriona Kelly, and Michela Bertolotto. 2010. RecoMap: An Interactive and Adaptive Map-based Recommender. In *Proceedings of the 2010 ACM Symposium on Applied Computing* (SAC '10), 887–891. https://doi.org/10.1145/1774088.1774273
5. Patricia Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. 2012. Omnipedia: Bridging the Wikipedia Language Gap. In *CHI '12: 30th International Conference on Human Factors in Computing Systems*.
6. Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. 2012. Omnipedia: Bridging the Wikipedia Language Gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12), 1075–1084. https://doi.org/10.1145/2207676.2208553
7. Christopher Bishop. 2007. *Pattern Recognition and Machine Learning*. Springer, New York.
8. Leo Breiman. 2001. Random forests. *Machine learning* 45, 1: 5–32.
9. Leo Breiman. 2002. Manual on setting up, using, and understanding random forests v3. 1. 2002. *URL: http://oz. berkeley. edu/users/breiman/Using_random_forests_V3* 1.
10. Tom Brinck, Darren Gergle, and Scott D. Wood. 2001. *Usability for the Web: Designing Web Sites that Work*. Morgan Kaufmann.
11. Ewa S Callahan and Susan C Herring. 2011. Cultural bias in Wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology* 62, 10. https://doi.org/10.1002/asi.21577
12. Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* 70, 4: 213.

13. D. Richard Cutler, Thomas C. Edwards, Karen H. Beard, Adele Cutler, Kyle T. Hess, Jacob Gibson, and Joshua J. Lawler. 2007. Random Forests for Classification in Ecology. *Ecology* 88, 11: 2783–2792. https://doi.org/10.1890/07-0539.1
14. Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. 2014. Introducing Wikidata to the Linked Data Web. In *The Semantic Web – ISWC 2014*, Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz and Carole Goble (eds.). Springer International Publishing, 50–65. https://doi.org/10.1007/978-3-319-11964-9_4
15. Killian Fox. 2012. OpenStreetMap: "It"s the Wikipedia of maps'. Retrieved October 31, 2016 from https://www.theguardian.com/theobserver/2012/feb/18/openstreetmap-world-map-radicals
16. Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *IJCAI '07: Twentieth Joint Conference for Artificial Intelligence*.
17. Pall Oskar Gislason, Jon Atli Benediktsson, and Johannes R. Sveinsson. 2006. Random Forests for land cover classification. *Pattern Recognition Letters* 27, 4: 294–300. https://doi.org/10.1016/j.patrec.2005.08.011
18. Scott A. Hale. 2015. Cross-language Wikipedia Editing of Okinawa, Japan. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 183–192. https://doi.org/10.1145/2702123.2702346
19. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* 11, 1: 10–18. https://doi.org/10.1145/1656274.1656278
20. Brent Hecht. 2013. The Mining and Application of Diverse Cultural Perspectives in User-Generated Content. Northwestern University, Evanston, IL.
21. Brent Hecht, Samuel Carton, Mahmood Quaderi, Johannes Schöning, Martin Raubal, Darren Gergle, and Doug Downey. 2012. Explanatory Semantic Relatedness and Explicit Spatialization for Exploratory Search. In *SIGIR '12*.
22. Brent Hecht and Darren Gergle. 2010. The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context. In *CHI '10: 28th International Conference on Human Factors in Computing Systems* (CHI '10), 291–300. https://doi.org/10.1145/1753326.1753370
23. Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. 2011. Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11), 237–246. https://doi.org/10.1145/1978942.1978976
24. Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence* 194: 28–61.
25. Lichan Hong, Gregorio Convertino, and Ed H. Chi. 2011. Language Matters In Twitter: A Large Scale Study. In *ICWSM*. Retrieved August 6, 2016 from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2856/3250/
26. Jos Hornikx and Daniel J. O'Keefe. 2011. Conducting Research on International Advertising: The Roles of Cultural Knowledge and International Research Teams. *Journal of Global Marketing* 24, 2: 152–166. https://doi.org/10.1080/08911762.2011.558813
27. Jaap Kamps and Marijn Koolen. 2009. Is Wikipedia Link Structure Different? In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (WSDM '09), 232–241. https://doi.org/10.1145/1498759.1498831
28. Aniket Kittur and Robert E. Kraut. 2008. Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (CSCW '08), 37–46. https://doi.org/10.1145/1460563.1460572
29. Aniket Kittur and Robert E. Kraut. 2010. Beyond Wikipedia: Coordination and Conflict in Online Production Groups. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (CSCW '10), 215–224. https://doi.org/10.1145/1718918.1718959
30. Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He Says, She Says: Conflict and Coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '07), 453–462. https://doi.org/10.1145/1240624.1240698
31. Marina Kogan, Jennings Anderson, Leysia Palen, Kenneth M. Anderson, and Robert Soden. 2016. Finding the Way to OSM Mapping Practices: Bounding Large Crisis Datasets for Qualitative Investigation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 2783–2795. https://doi.org/10.1145/2858036.2858371
32. J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1: 159–174. https://doi.org/10.2307/2529310
33. Paul Laufer, Claudia Wagner, Fabian Flöck, and Markus Strohmaier. 2015. Mining cross-cultural relations from Wikipedia - A study of 31 European food cultures. In *ACM WebScience '15*.
34. Xiaozhong Liu, Tian Xia, Yingying Yu, Chun Guo, and Yizhou Sun. 2016. Cross Social Media Recommendation. In *AAAI ICWSM '16*.
35. Paolo Massa and Federico Scrinzi. 2011. Exploring Linguistic Points of View of Wikipedia. In *Proceedings*

of the 7th International Symposium on Wikis and Open Collaboration (WikiSym '11), 213–214. https://doi.org/10.1145/2038558.2038599

36. Paolo Massa and Federico Scrinzi. 2012. Manypedia: Comparing Language Points of View of Wikipedia Communities. In *WikiSym '12: 8th International Symposium on Wikis and Open Collaboration*.
37. David Milne and Ian H Witten. 2008. An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In *WIKIAI '08: First AAAI Workshop on Wikipedia and Artificial Intelligence*.
38. David Milne and Ian H Witten. 2008. Learning to link with wikipedia. In *CIKM '08: 17th ACM Conference on Information and Knowledge Management* (CIKM '08), 509–518. https://doi.org/10.1145/1458082.1458150
39. Dan Morris. 2015. Improving the replicability of HCI research with supplementary material. Retrieved May 27, 2016 from http://sigchi.tumblr.com/post/129729411430/improving-the-replicability-of-hci-research-with
40. M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. 2016. A Review of Relational Machine Learning for Knowledge Graphs. *Proceedings of the IEEE* 104, 1: 11–33. https://doi.org/10.1109/JPROC.2015.2483592
41. J. Nielsen. 2011. Mini-IA: Structuring the Information About a Concept. Retrieved July 20, 2016 from https://www.nngroup.com/articles/mini-ia-structuring-information/
42. Jakob Nielsen. 1997. Be succinct!(Writing for the Web). Retrieved July 20, 2016 from http://www.useit.com/alertbox/9703b.html
43. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10* (EMNLP '02), 79–86. https://doi.org/10.3115/1118693.1118704
44. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, Oct: 2825–2830.
45. Ulrike Pfeil, Panayiotis Zaphiris, and Chee Siang Ang. 2006. Cultural Differences in Collaborative Authoring of Wikipedia. *Journal of Computer-Mediated Communication* 12, 1: 88–113.
46. Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. 2010. Cross-language plagiarism detection. *Language Resources and Evaluation* 45, 1: 45–62. https://doi.org/10.1007/s10579-009-9114-z
47. Shilad Sen, Toby Jia-Jun Li, WikiBrain Team, and Brent Hecht. 2014. WikiBrain: Democratizing Computation on Wikipedia. In *Proceedings of The International Symposium on Open Collaboration*, 27:1–27:10. https://doi.org/10.1145/2641580.2641615
48. Wei Shen, Jianyong Wang, Ping Luo, and Min Wang. 2012. LINDEN: Linking Named Entities with Knowledge Base via Semantic Knowledge. In *Proceedings of the 21st International Conference on World Wide Web* (WWW '12), 449–458. https://doi.org/10.1145/2187836.2187898
49. Amit Singhal. 2012. Introducing the Knowledge Graph: things, not strings. *Google: Official Blog*. Retrieved from http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html
50. Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9: 307. https://doi.org/10.1186/1471-2105-9-307
51. Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8: 25. https://doi.org/10.1186/1471-2105-8-25
52. Michael Strube and Simone Paolo Ponzetto. 2006. WikiRelate! Computing Semantic Relatedness Using Wikipedia. In *AAAI '06: The Twenty-First National Conference on Artificial Intelligence*, 1419–1424.
53. Ramine Tinati, Paul Gaskell, Thanassis Tiropanis, Olivier Phillipe, and Wendy Hall. 2014. Examining Wikipedia Across Linguistic and Temporal Borders. In *Proceedings of the 23rd International Conference on World Wide Web* (WWW '14 Companion), 445–450. https://doi.org/10.1145/2567948.2576931
54. F. B. Viegas, M. Wattenberg, J. Kriss, and F. V. Ham. 2007. Talk Before You Type: Coordination in Wikipedia. In *40th Annual Hawaii International Conference on System Sciences, 2007. HICSS 2007*, 78–78. https://doi.org/10.1109/HICSS.2007.511
55. Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM* 57, 10: 78–85. https://doi.org/10.1145/2629489
56. Daniel S. Weld, Fei Wu, Eytan Adar, Saleema Amershi, James Fogarty, Raphael Hoffmann, Kayur Patel, and Michael Skinner. 2008. Intelligence in Wikipedia. In *AAAI*, 1609–1614. Retrieved May 27, 2016 from http://www.aaai.org/Papers/AAAI/2008/AAAI08-274.pdf
57. Wikipedia contributors. 2015. Wikipedia:Article size. *Wikipedia, the free encyclopedia*. Retrieved December 20, 2015 from https://en.wikipedia.org/w/index.php?title=Wikipedia:Article_size&oldid=686412509
58. Wikipedia editors. Wikipedia:100,000 feature-quality articles. Retrieved from https://en.wikipedia.org/wiki/Wikipedia:100,000_feature-quality_articles

59. Wikipedia Editors. Template:Main article. Retrieved October 31, 2016 from https://en.wikipedia.org/wiki/Template:Main_article
60. Wikipedian Editors. Wikipedia:Splitting. Retrieved from https://en.wikipedia.org/wiki/Wikipedia:Splitting#How_to_properly_split_an_article
61. Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann.
62. Ellery Wulczyn, Robert West, Leila Zia, and Jure Leskovec. 2016. Growing Wikipedia Across Languages via Recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, 975–985. Retrieved July 20, 2016 from http://dl.acm.org/citation.cfm?id=2883077
63. ICWSM Data Sharing Initiative. Retrieved May 27, 2016 from http://www.icwsm.org/2016/datasets/datasets/

APPENDIX A: DETAILED CLASSIFICATION ACCURACIES FOR FIGURE 3.1- 3.4

Detailed classification results for the `High-Interest` Dataset (Figure 3.1)

	zh			en			es		
	Rating:2	**Rating:2.5**	**Rating:3**	**Rating:2**	**Rating:2.5**	**Rating:3**	**Rating:2**	**Rating:2.5**	**Rating:3**
Baseline	63.54%	59.37%	50%	56.03%	50.24%	58.93%	50.98%	70.58%	81.37%
Linear SVM	85.44%	90.81%	84.66%	88.92%	91.38%	83.09%	64.81%	67.45%	81.18%
Random Forests	85.33%	86.44%	86.77%	92.69%	90.80%	82.95%	70.00%	74.18%	81.18%
Naïve Bayes	85.44%	87.66%	83.66%	89.35%	89.88%	82.57%	65.90%	73.36%	70.45%
Logistic	85.33%	88.07%	83.44%	90.33%	89.90%	82.07%	67.72%	71.27%	78.27%
KNN	72.88%	74.00%	71.88%	80.54%	77.21%	78.16%	69.54%	79.18%	83.18%
Adaboost	83.33%	85.44%	82.66%	88.88%	90.83%	82.97%	59.18%	70.72%	75.18%
Decision tree	79.00%	81.00%	75.88%	88.28%	86.42%	78.07%	67.72%	62.81%	72.27%

Detailed classification results for the `Random` Dataset (Figure 3.2)

	zh			en			es		
	Rating:2	**Rating:2.5**	**Rating:3**	**Rating:2**	**Rating:2.5**	**Rating:3**	**Rating:2**	**Rating:2.5**	**Rating:3**
Baseline	69.60%	76.47%	85.29%	89.10%	91.58%	98.01%	63.00%	86.00%	100%
Linear SVM	83.18%	82.27%	91.18%	91.52%	92.02%	98.02%	66.99%	85.00%	N/A
Random Forests	79.27%	86.18%	86.18%	91.04%	93.02%	97.52%	63.00%	90.00%	N/A
Naïve Bayes	73.45%	69.36%	85.18%	79.61%	75.26%	59.04%	67.99%	77.00%	N/A
Logistic	84.18%	83.27%	92.095	91.52%	92.02%	98.02%	66.99%	88.00%	N/A
KNN	81.45%	83.09%	86.09%	89.52%	92.02%	98.02%	52.00%	85.00%	N/A
Adaboost	81.18%	81.18%	90.09%	87.59%	90.02%	98.02%	62.99%	83.00%	N/A
Decision tree	82.27%	76.36%	92.18%	84.09%	87.57%	97.04%	64.99%	86.00%	N/A

Detailed classification results for the `Ad-Hoc` Dataset (Figure 3.3)

	zh			en			es		
	Rating:2	**Rating:2.5**	**Rating:3**	**Rating:2**	**Rating:2.5**	**Rating:3**	**Rating:2**	**Rating:2.5**	**Rating:3**
Baseline	67.96%	58.25%	55.33%	56.71%	55.72%	74.12%	51.15%	59.59%	75.75%
Linear SVM	85.63%	83.63%	80.45%	73.66%	73.66%	77.09%	69.44%	80.55%	71.88%
Random Forests	86.63%	82.54%	81.36%	70.57%	72.14%	80.61%	68.66%	73.66%	76.88%
Naïve Bayes	79.72%	78.72%	81.36%	73.14%	69.69%	37.30%	67.44%	71.66%	67.77%
Logistic	86.45%	85.54%	81.45%	71.19%	73.66%	77.57%	66.44%	73.55%	75.00%
KNN	78.81%	82.63%	71.72%	71.28%	66.28%	69.61%	59.55%	66.55%	66.77%
Adaboost	82.54%	73.90%	75.63%	70.07%	73.11%	78.11%	64.55%	70.44%	71.66%
Decision tree	84.72%	81.36%	77.63%	64.66%	65.51%	71.09%	60.55%	67.55%	64.77%

Detailed overall classification accuracies (Figure 3.4)

	`High-Interest`			`Ad-Hoc`			`Random`		
	Rating:2	**Rating:2.5**	**Rating:3**	**Rating:2**	**Rating:2.5**	**Rating:3**	**Rating:2**	**Rating:2.5**	**Rating:3**
Baseline	56.04%	52.83%	62.46%	58.31%	53.10%	69.72%	77.77%	86.38%	95.27%
Linear SVM	86.09%	88.59%	81.44%	74.00%	77.70%	79.12%	81.10%	86.57%	94.78%
Random Forests	83.86%	85.87%	78.73%	73.46%	74.70%	74.62%	78.16%	88.34%	94.28%
Naïve Bayes	84.12%	85.62%	81.20%	73.50%	74.26%	74.21%	74.70%	80.18%	86.39%
Logistic	84.87%	87.60%	80.70%	75.46%	76.45%	79.37%	81.10%	88.32%	94.28%
KNN	77.47%	77.75%	74.82%	72.03%	72.28%	72.92%	80.60%	87.83%	94.52%
Adaboost	84.17%	84.66%	80.96%	74.99%	73.74%	75.63%	80.39%	86.59%	94.40%
Decision tree	85.13%	80.17%	74.57%	67.03%	67.01%	72.42%	74.45%	79.90%	92.34%