Interpreting Open-Domain Modifiers: Decomposition of Wikipedia Categories into Disambiguated Property-Value Pairs

Marius Paşca
Google
1600 Amphitheatre Parkway
Mountain View, California 94043
mars@google.com

Abstract

This paper proposes an open-domain method for automatically annotating modifier constituents (*"20th-century"*) within Wikipedia categories (*"20th-century male writers"*) with properties (*"date of birth"*). The annotations offer a semantically-anchored understanding of the role of the constituents in defining the underlying meaning of the categories. In experiments over an evaluation set of Wikipedia categories, the proposed method annotates constituent modifiers as semantically-anchored properties, rather than as mere strings in a previous method. It does so at a better trade-off between precision and recall.

1 Introduction

Motivation: As Web search moves towards returning structured answers rather than flat sets of document links in response to users' queries, the need for high-quality, wide-coverage knowledge to support such answers is growing stronger. The largest of the existing knowledge repositories (Bizer et al., 2009; Hoffart et al., 2013; Nastase and Strube, 2013), whether publicly available (Bollacker et al., 2008; Vrandečić and Krötzsch, 2014) or restricted to commercial access (Wu et al., 2012), uniformly rely on data in Wikipedia for their core sets of topics and knowledge assertions. In addition to its role in Web search and information retrieval (Chen et al., 2017; Ensan and Bagheri, 2017; Ma et al., 2018; Zhang and Balog, 2018), Wikipedia and the knowledge repositories derived from it are useful in a growing variety of tasks. Such tasks pertain to text analysis (Ratinov et al., 2011; Murty et al., 2018) and, specifically, knowledge acquisition from text (Nastase and Strube, 2013; Wu et al., 2012; Hoffart et al., 2013; Gupta et al., 2019).

Millions of Wikipedia articles are connected to parent categories, which are in turn connected to their own, iteratively broader categories. For example, the article *"art:Gary Oldman"* is connected to parent categories such as *"ctg:20th-century English male actors"*, *"ctg:Alumni of Rose Bruford College"*, which are in turn connected to their own, broader categories such as *"ctg:Actors"*, *"ctg:Acting"*, *"ctg:Language"*. Some categories really correspond to individual topics, e.g., *"ctg:Rose Bruford College"*. Many other categories in Wikipedia - hundreds of thousands, by our estimates - each corresponds to a fine-grained class that groups together individual articles sharing common properties. For example, the category *"ctg:20th-century English male actors"* groups articles such as *"art:Gary Oldman"* and *"art:Jude Law"*, which conceptually share properties that could be described as *"born or living in the 20th-century"*, *"born in England"*, *"being a male"* and *"being actors"*. That Wikipedia organizes topics into hundreds of thousands of potential fine-grained classes is remarkable. Comparatively, topics in other knowledge repositories are organized into only hundreds of types, in DBpedia (Bizer et al., 2009); or thousands of collections, in Freebase (Bollacker et al., 2008). Existing applications (Ma et al., 2018) that take advantage of Wikipedia categories include the creation of large, deep, fine-grained hierarchies out of Wikipedia articles and their categories (Flati et al., 2014; Gupta et al., 2018). Unfortunately, in both Wikipedia and downstream applications, Wikipedia categories are represented as nothing more than mere strings. Their meaning is otherwise not captured. Understanding and annotating the meaning of Wikipedia categories would make them more useful and increase their impact.

Contributions: The main contributions of this paper are as follows. First, it provides a precise, semantically-anchored understanding of the role of the constituents in defining the underlying meaning of Wikipedia categories. For this purpose, it pro-

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 6218–6228,
November 16–20, 2020.

poses an open-domain method for annotating modifier constituents within categories with properties and values referred to by modifier constituents. Previous work (Paşca, 2017) annotates categories with isolated, ambiguous strings such as *"century"* for the constituent *"20th-century"* within the category *"ctg:20th-century English male actors"*. In contrast, the method proposed here annotates categories with non-ambiguous values that are Wikipedia articles or Wikidata topics, such as *"val:20th century"* for the same constituent *"20th-century"* within *"ctg:20th-century English male actors"*. More importantly, it also annotates categories with properties with well-defined descriptions and semantic meaning in Wikidata (Vrandečić and Krötzsch, 2014), thus annotating the same constituent *"20th-century"* with the property *"prp:P569 (date of birth)"* from Wikidata. Second, the paper is the first to investigate the role of Wikidata in automatically enriching Wikipedia categories. In contrast, previous methods for open-domain information extraction mostly rely on unstructured or semi-structured text (Sun et al., 2018), Wikipedia (Tsurel et al., 2017) and repositories other than Wikidata (Hoffart et al., 2013; Qu et al., 2018; Moniruzzaman et al., 2019) with few exceptions (Chisholm et al., 2017). Third, in experiments over the gold set of Wikipedia categories, in comparison to a previous method (Paşca, 2017), the proposed method automatically annotates constituent modifiers as semantically-anchored properties and topics, rather than mere strings. It does so at a better trade-off between precision and recall.

2 Annotating Categories

Notations: The following prefixes distinguish among the various kinds of items: *art* as in *"art:Gary Oldman"*, for a Wikipedia article (http://en.wikipedia.org/wiki/Gary_Oldman); *ctg* as in *"ctg:20th-century English male actors"*, for a Wikipedia category (http://en.wikipedia.org/wiki/Category:20th-century_English_male_actors); *tpc* as in *"tpc:Gary Oldman"*, for a Wikidata topic (http://www.wikidata.org/wiki/Q83492), which often has an equivalent Wikipedia article (*"art:Gary Oldman"*); *prp* as in *"prp:P569 (date of birth)"*, for a Wikidata property (http://www.wikidata.org/wiki/Property:P569); *val* as in *"val:20th century"*, for the value of a property of a topic in Wikidata.

Goal: Finer-grained categories in Wikipedia often take the form of compositional noun phrases. Within such categories, individual modifier constituents refer to values that implicitly allude to explicit properties applying to, and shared by, the descendant Wikipedia articles located under the categories. For example, the modifier constituent *"English"* alludes to an explicit property regarding the *place of birth* applying to *"art:Gary Oldman"*, *"art:Jude Law"* and other descendant Wikipedia articles located under the category *"ctg:20th-century English male actors"*. Categories are represented simply as strings in Wikipedia. Understanding the role played by as many of their individual constituents as possible, by accurately identifying the explicit properties to which their constituents implicitly refer, would go a long way in understanding the overall meaning of the categories. It is the main goal of this paper.

Sources of Annotations: As suggested in (Paşca, 2017), Wikipedia itself can serve as the source for annotations of modifier constituents within Wikipedia categories. If Wikipedia connects a child category *"ctg:20th-century English male actors"* to a parent category *"ctg:English male actors by century"*, such a connection can be taken as evidence that the modifier *"20th-century"* within the child category plays a certain role *"century"*. Relying on data within Wikipedia itself is elegant but has shortcomings. First, the extracted annotations are strings. They are ambiguous. Whether the annotation *"century"* refers to a unit of measuring time, a 1981 novel or a cruise ship launched in 1995 is not encoded or clarified. Second, generic or underspecified annotations, like the string *"type"* for *"Zoology"* in *"ctg:Zoology museums"*, do not add much towards understanding the meaning of categories. Third, the annotations often reveal only the type of the value of the property alluded to by the modifier constituent, which is insufficient for understanding the explicit property. To illustrate, the annotation *"century"* for the modifier *"20th-century"* is arguably insufficient; *lived during* or *born in* would be more desirable.

As an alternative to reliance of previous work on Wikipedia itself, a novel aspect of the method proposed here is taking advantage of data available within Wikidata. Like Wikipedia, Wikidata is an actively developed, growing resource that benefits from editing by human contributors. For millions of topics, many of which are explicitly mapped to a corresponding Wikipedia article, Wikidata asserts knowledge about the topics as property-value

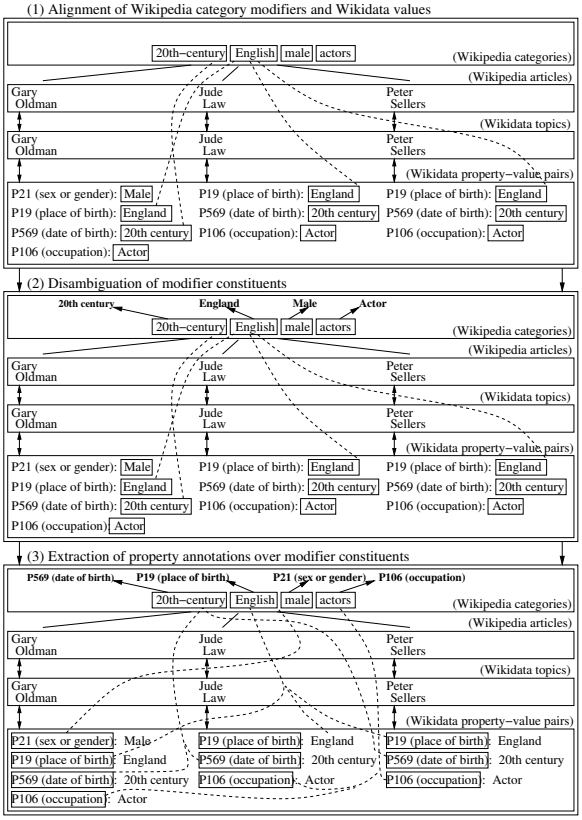

Figure 1: Overview of annotation of modifier constituents within Wikipedia categories based on Wikidata properties and values

pairs. Examples are the pairs *"prp:P19 (place of birth)"* and *"tpc:England"*; or *"prp:P21 (sex or gender)"* and *"tpc:Male"*, for the Wikidata topic *"tpc:Gary Oldman"* mapped to the Wikipedia article *"art:Gary Oldman"*. Values are usually other Wikidata topics. Properties (or predicates) are themselves special topics with explicit semantics in Wikidata. For example, *"prp:P19 (place of birth)"* is described in Wikidata as *"[..] birth location of a person, animal or fictional character"*. Unlike strings, Wikidata properties are semantically-anchored and therefore preferable as annotations over Wikipedia categories.

Acquisition from Wikidata: As illustrated in Figure 1, the proposed method annotates Wikipedia categories with Wikidata-based properties and values in three stages: (1) align arbitrary ngrams from Wikipedia categories, on one hand, to values of Wikidata properties of descendant Wikipedia articles, on the other hand; (2) as a side effect of the alignment, disambiguate the aligned ngram (string) to the aligned value (which, in most cases, is a Wikidata topic that also has an equivalent Wikipedia article); and (3) among Wikidata properties whose values were aligned to ngrams, extract one property per ngram, as a property annotation for the modifier constituent represented by the ngram.

(1) **Alignment of Modifiers and Values**: Rather than requiring modifier constituents of a Wikipedia category to be separately identified in advance, modifier constituents are identified and extracted simultaneously along with the annotations (first box from the top in Figure 1). The set of all contiguous spans (ngrams) within a Wikipedia category constitutes the initial, very noisy set of candidate modifier constituents of the category. Candidate alignments for each ngram come from the Wikidata property-value pairs of descendant Wikipedia articles of the category. The ngram is compared to the name of each value. In the first (top) box in Figure 1, one of the descendant Wikipedia articles of the category *"ctg:20th-century English male actors"* is *"art:Gary Oldman"*. It has the property *"prp:P21 (sex or gender)"* in Wikidata, whose value *"tpc:Male"* matches (after string normalization) the ngram *"male"* from the category.

In case of a match between a category ngram and a Wikidata value, the Wikidata property of the matched value becomes a candidate property annotation of the category ngram. Simultaneously, the category ngram becomes a candidate modifier constituent of the category. For the category *"ctg:20th-century English male actors"*, the property *"prp:P21 (sex or gender)"* becomes a candidate property annotation of the modifier constituent *"male"* from the category.

(2) **Disambiguation of Modifier Constituents**: The values of Wikidata properties are usually Wikidata topics which, in turn, have equivalent Wikipedia articles. A side effect of the alignment is the disambiguation of the category's modifier constituents (aligned ngrams) in terms of unambiguous Wikidata topics. For example, the ambiguous ngram *"English"* from the category *"ctg:20th-century English male actors"* is aligned to the value *"tpc:England"*, in the second box from the top in Figure 1. The value is a Wikidata topic. Therefore, the ambiguous modifier constituent *"English"* from the category is effectively disambiguated to the Wikidata topic *"tpc:England"* and its equivalent Wikipedia article *"art:England"*.

(3) **Extraction of Property Annotations**: The alignment may produce multiple candidate properties for the same modifier constituent of a category. For each modifier constituent, the candidate prop-

erty with the largest article support set is selected as the preferred property annotation (third box from the top in Figure 1). The article support set is the set of descendant Wikipedia articles of the category that, based on the alignment of their Wikidata values, contribute towards extracting a particular Wikidata property for a given modifier constituent. For example, the candidate properties for the modifier constituent *"Armin van Buuren"* of *"ctg:Armin van Buuren albums"* are *"prp:P676 (lyrics by)"*, *"prp:P162 (producer)"* and *"prp:P175 (performer)"*. The last is extracted via alignment to values of more descendant Wikipedia articles (e.g., *"art:Imagine (Armin van Buuren album)"*, *"art:10 Years (Armin van Buuren album)"*, *"art:Intense"*) than the other candidates. It is selected as the property annotation for *"Armin van Buuren"*.

Overall Annotations of Modifier Constituents: Given a category from Wikipedia, the method and its associated stages described above produce annotations for zero or more of its modifier constituents. If a modifier constituent is annotated, it is annotated with a topic that disambiguates it; and/or a property annotation. For the category *"ctg:20th-century English male actors"*, the modifier constituent *"20th-century"* is annotated with: the topic *"tpc:20th century"*, which disambiguates it; and the property *"prp:P569 (date of birth)"*. Note that the method is not limited to annotating modifier constituents. Depending on data available in Wikidata, the method might also annotate head constituents (without distinguishing them as such), although less frequently. For example, the method successfully annotates *"novels"* and *"women"* in *"ctg:Zombie novels"* and *"ctg:17th-century Norwegian women"* with *"prp:P136 (genre)"* and *"prp:P21 (sex or gender)"* respectively. But it fails to annotate *"games"* in *"ctg:Zombie Studios games"*.

3 Experimental Setting

Data Sources: The experiments operate over English snapshots of Wikipedia and Wikidata from June 2018. As in previous work (Ponzetto and Strube, 2007; Paşca, 2017), Wikipedia articles are automatically discarded if they are disambiguation or redirect pages. Similarly to (Ponzetto and Strube, 2007; Piccardi et al., 2018), Wikipedia categories are automatically discarded if they are meant for Wikipedia's internal bookkeeping, as approximated by the presence of the subphrases *article(s)*, *category(ies)*, *infobox(es)*, *pages*, *redirects*, *stubs*, *templates*, *wikiproject*, *use mdy dates*, *lists*, *stubs* or *wikidata* in their names. Finally, as in (Hoffart et al., 2013; Paşca, 2017; Gupta et al., 2018), categories are automatically discarded if they likely correspond not to classes but rather to individual topics. In this case, such categories are approximated by the absence of any plural-form tokens (based on lemmatization data in WordNet), thus discarding, e.g., *"ctg:Rose Bruford College"*. Alternatively, previous work on distinguishing Wikipedia articles that are classes (Paşca, 2018) could be extended to Wikipedia categories. The filtered Wikipedia snapshot connects 5,101,643 articles to 1,124,679 categories.

By traversing chains of Wikidata property-value pairs whose property is *"prp:P131 (located in the administrative territorial entity)"*, some of the existing location-based data in Wikidata is automatically expanded. For example, given the existing property-value pair *"prp:P19 (place of birth)"* and *"tpc:London"* for some Wikidata topic, additional property-value pairs like *"prp:P19 (place of birth)"* and *"tpc:Greater London"*, or *"prp:P19 (place of birth)"* and *"tpc:London"*, are added to the same Wikidata topic. Some of the temporal values in Wikidata are also automatically expanded. For any property-value pairs whose values are encoded as dates in Wikidata, additional property-value pairs are generated by a) selecting only the years; and also b) replacing the years with corresponding decades and centuries. For example, additional values generated starting from the value *"21 March 1958"* include *"tpc:1958"*, *"tpc:1950s"* and *"tpc:20th century"*.

Extraction Parameters: Property annotations extracted by the proposed method are discarded if they are one of the two properties used in Wikidata for organizing topics hierarchically, namely *"prp:P31 (instance of)"* or *"prp:P279 (subclass of)"*. During the alignment of modifier constituents from categories to values of properties from Wikidata, the two strings being compared are considered to match if, after conversion to lowercase, their lemmas (Fellbaum, 1998) or stems are either identical or one is an adjectival form and the other is the corresponding nominal form in WordNet, e.g., *"English"* vs. *"England"*.

Experimental Runs: The method from (Paşca, 2017) exploits the Wikipedia category network. It is a method available specifically for annotating modifier constituents within Wikipedia cate-

gories. Based on connections in Wikipedia from a child category of the form *"Z X"* to a parent category of the form *"X by Y"*, it extracts the annotation *"Y"* for the modifier constituent *"Z"* in the child category. It serves as a baseline run (denoted $\mathbf{B}_{wcn}$) in our experiments. For example, the modifier constituent *"20th-century"* in the child category *"ctg:20th-century actors"* is annotated as *"century"*, based on the presence of the parent category *"ctg:Actors by century"*. Besides the baseline run B_{wcn}, the method proposed here is evaluated through an experimental run denoted $\mathbf{R}_{prp}$. It extracts Wikidata properties such as *"prp:P569 (date of birth)"*, as property annotations.

Evaluation Set: The gold evaluation set for our experiments is a random sample of 700 target Wikipedia categories that are classes rather than individual topics. The set is created by inspecting random Wikipedia categories manually and either discarding them, if they correspond not to classes but instead to individual topics such as *"ctg:Association for Computing Machinery"* and *"ctg:Mille Lacs County, Minnesota"*; or retaining them, until the desired number of categories have been retained. Choosing the number of categories to retain is a balance between the desire to create a large evaluation set, on one hand; and the reality of the high cost (duration) of manual assignment of gold annotations, on the other hand. The retained target categories form the evaluation set for which modifier constituents and associated gold annotations must be compiled. The evaluation set covers a diverse range of domains of interest including art and entertainment for *"ctg:12 Stones albums"*, sports for *"ctg:Bulgarian arm wrestlers"*, religion for *"ctg:Buddhist temples in Southeast Asia"* or technology for *"ctg:3D platform games"*.

The manual assignment of annotations to Wikipedia categories from scratch would be a daunting task. It would require the analysis of hundreds of candidate Wikidata topics (properties), in order to select the correct or best annotation for each possible modifier constituent within each category. Even assuming unlimited human annotation resources of the highest quality, the task would be cumbersome and time-consuming, if not infeasible.

In information retrieval, it is not uncommon to assess the relevance of documents selected not from the entire underlying document collection, but rather from documents automatically retrieved by any of the retrieval methods being evaluated.

Label	Score	Ignored?		Description
		Ip?	Ir?	
c	1.0	No	No	Correct annotation
i	0.0	No	No	Incorrect annotation
s	0.0	No	Yes	Incorrectly identified modifier
d	0.0	Yes	No	Modifier with unspecified annotation

Table 1: Correctness labels assigned to triples of a target Wikipedia category, modifier constituent and annotation in the evaluation set (Label=correctness label; Score=score of correctness label; Ip?=ignored during computation of precision?; Ir?=ignored during computation of recall?)

Target Category: Modifier Constituent→Annotation	Label
19th-century French politicians: French→country	i
19th-century French politicians: French→P27 (country of citizenship)	c
19th-century French politicians: politicians→profession	c
Plautdietsch-language films: Plautdietsch-language→P364 (original language of work)	c
Artists from Liverpool: Liverpool→city	i
Artists from Liverpool: Liverpool→P19 (place of birth)	c
Courts in Sweden: Sweden→P27 (country of citizenship)	i
People from Yozgat Province: Yozgat→city	s
People from Yozgat Province: Yozgat Province→(unspecified annotation)	d

Table 2: Examples of entries from the evaluation set. An entry is tuple of a target category, a modifier constituent, an extracted (or unspecified) annotation and a correctness label (Label=correctness label)

Similarly, the practical alternative pursued here is to manually label the correctness of automatically extracted annotations. For each of the target categories in the evaluation set, the annotations extracted for its modifier constituents by the experimental runs are manually labeled with one of a set of correctness labels, according to the perceived correctness of the annotations. Shown in Table 1, the correctness labels quantify the correctness of an annotation extracted for a modifier constituent within a category. They assess whether an annotation captures a property and, if so, whether it does so correctly. Table 2 illustrates correctness labels assigned to a sample of extracted annotations from the evaluation set. Each entry in the evaluation set is a tuple of a target category, a modifier constituent, an extracted (or unspecified) annotation and its correctness label. In the evaluation set, an-

notations that capture the desired property correctly are assigned the correctness label *c* (*Correct annotation*). Correct annotations by run B_{wcn} must capture the desired property lexically, such as the string *"profession"* for the modifier *"politicians"* within *"ctg:19th-century French politicians"*. Comparatively, correct annotation by run R_{prp} must capture the desired property semantically, such as the Wikidata property *"prp:P106 (occupation)"* for the same modifier. It is not sufficient that the string name *occupation* of the Wikidata property lexically capture the desired property. Thus, the assignment of correctness labels is relatively more lenient for run B_{wcn} and comparatively stricter for run R_{prp}. Annotations deemed incorrect are assigned the correctness label *i*. For example, the annotation *"country"* extracted for *"French"* in *"ctg:19th-century French politicians"* is incorrect because it does not reveal whether the underlying property might be *visited* or perhaps *born in* or possibly *made in*.

Target categories from the evaluation set may contain relevant modifier constituents for which none of the experimental runs extract any annotations. Such modifier constituents do not receive any manual correctness label and do not become part of the evaluation set so far. The resulting evaluation set would still be well suited for computing relative recall among the various experimental runs; but less suited for computing absolute recall. To alleviate the problem, modifier constituents that should have some (unspecified) annotation that has not yet been extracted by any of the experimental runs are annotated as such. For this purpose, a special "unspecified" annotation is added in the evaluation set for those modifier constituents. They are assigned the correctness label *d* from Table 1. An example is the modifier constituent *"Yozgat Province"* in *"ctg:People from Yozgat Province"* in Table 2. Consequently, the evaluation set can be used to compute not just the precision but also the recall of a given experimental run. Overall, the evaluation set contains one or more annotations for each of 1,316 unique pairs of a target category and a modifier constituent. Note that the count is larger than the number of entries in evaluation sets previously introduced for the evaluation of tasks related to compositionality analysis (Hendrickx et al., 2013; Paşca and Buisman, 2015; Paşca, 2017). The target categories in the evaluation set each consist of just above 4 tokens on average. Entries containing annotations extracted by different experimental runs

Fraction of Categories with Extracted Annotations		
Reference Set	Run	
	B_{wcn}	R_{prp}
All Wiki	0.553	0.765
Gold Wiki	0.498	0.722

Table 3: Fraction of Wikipedia categories for which various runs extract some annotations for at least one modifier constituent. Computed as a fraction of the reference sets of all Wikipedia categories (All Wiki) and also of all Wikipedia categories from the gold evaluation set (Gold Wiki)

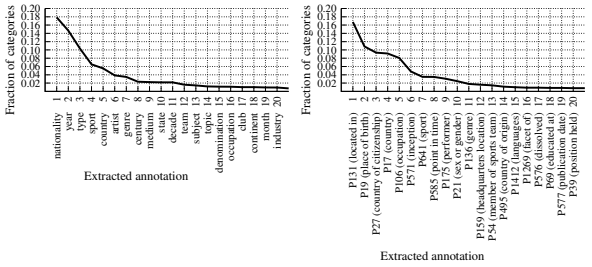

Figure 2: Most frequent annotations extracted over all categories from Wikipedia by run B_{wcn} (left graph) and by run R_{prp} (right graph). Computed as the fraction of Wikipedia categories for which a particular annotation is extracted for one of its modifier constituents

are merged and sorted alphabetically, before being presented to two human annotators. The annotators manually assign correctness labels to the entries in the evaluation set. The agreement is 84%, when computed as a percentage of entries annotated by both annotators being assigned identical correctness labels; and 0.525, when computed as Cohen's Kappa coefficient.

4 Evaluation Results

Coverage: When coverage is measured as the fraction of Wikipedia categories for which some annotations are extracted, the proposed method outperforms the baseline run B_{wcn} in Table 3. Figure 2 shows the annotations extracted most frequently by run B_{wcn} and by the proposed method. The horizontal axis represents the extracted annotations, sorted from most to least frequently extracted. There are 1,612 unique annotations extracted by B_{wcn} but virtually all of them are really type annotations rather than capturing any property annotations. In comparison, the proposed method extracts as many as 519 unique property annotations.

Evaluation Metrics: To automatically assess the annotations extracted by an experimental run, over the target categories in the evaluation set, their cor-

Run	Scores					
	Macro-Averaged			Micro-Averaged		
	P	R	F	P	R	F
B_{wcn}	0.526	0.007	0.013	0.526	0.008	0.015
R_{prp}	0.925	0.517	0.663	0.919	0.514	0.659

Table 4: Precision (P) and recall (R) (F=F_1-score)

rectness labels are retrieved from the evaluation set. The correctness labels are converted to individual correctness scores shown in the earlier Table 1. Micro- and macro-averaged precision and recall scores are computed out of individual correctness scores. Micro-averaged scores are computed as an average over all annotations extracted by an experimental run. Macro-averaged scores are first computed separately for each target category, then averaged over all target categories.

Precision and Recall: Table 4 compares the performance of the extraction methods. The baseline run B_{wcn} has low macro-averaged recall and limited precision. In comparison, the properties extracted by the proposed method are more numerous and more accurate than the properties (really, types) extracted by the baseline. The proposed method gives uniformly higher F_1-scores than the baseline.

Table 5 gives examples of annotations extracted by the baseline run B_{wcn} vs. the proposed method R_{prp}. Although not shown in the table, annotations are ambiguous strings for run B_{wcn}, vs. disambiguated properties or topics from Wikidata or Wikipedia for run R_{prp}. Annotations are extracted for modifier constituents that are ambiguous strings for run B_{wcn}, e.g., *"Indian"*, *"Jain"*; vs. disambiguated topics for run R_{prp}, e.g., the Wikipedia articles *"art:India"*, *"art:Jainism"*.

In Table 5, the baseline run B_{wcn} can only annotate modifier constituents such as *"Indian"*, *"Jain"* in *"ctg:20th-century Indian Jain politicians"*. In contrast, the proposed method may also annotate head constituents, such as *"politicians"*. An additional experiment quantifies the role of annotations extracted for head constituents in increasing the recall of the proposed method. For each of a subset of 200 of the target categories from the evaluation set, the annotations extracted by R_{prp} are manually inspected in order to identify and temporarily discard annotations of head (rather than of modifier) constituents. For example, annotations extracted by R_{prp} for *"platform games"* in *"ctg:3D platform games"* or for *"Artists"* in *"ctg:Artists from Liverpool"* are temporarily discarded. Temporarily

Run: Extracted Annotations
Category: 1872 ballet premieres:
B: *1872*→[year]
R: (none)
Category: 1873 ships:
B: (none)
R: *1873*→[prp:P729 (service entry)]
Category: 20th-century Indian Jain politicians:
B: *20th-century*→[century]; *Indian*→[nationality]
R: *20th-century*→[prp:P569 (date of birth)]; *Indian*→[prp:P27 (country of citizenship)]; *Jain*→[prp:P140 (religion)]; *politicians*→[prp:P106 (occupation)]
Category: Orange Democratic Movement politicians:
B: *Orange Democratic Movement*→[party]
R: *Orange Democratic Movement*→[prp:P102 (member of political party)]; *politicians*→[prp:P106 (occupation)]
Category: Orange Goblin albums:
B: *Orange Goblin*→[artist]
R: *Orange Goblin*→[prp:P175 (performer)]
Category: Orange Is the New Black characters:
B: (none)
R: *Orange Is the New Black*→[prp:P1441 (present in work)]
Category: Orange liqueurs:
B: (none)
R: *Orange*→[prp:P186 (material used)]
Category: Oral Roberts Golden Eagles women's basketball seasons:
B: *basketball*→[sport]; *Oral Roberts Golden Eagles*→[school]; *women's*→[membership]
R: *basketball*→[prp:P641 (sport)]; *Oral Roberts Golden Eagles*→[prp:P5138 (season of club or team)]; *women's basketball*→[prp:P2094 (competition class)]
Category: Zombie novels:
B: (none)
R: *Zombie*→[prp:P180 (depicts)]; *novels*→[prp:P136 (genre)]

Table 5: Examples of annotations extracted by runs B_{wcn} vs. R_{prp} for a sample of target categories (B=run B_{wcn}; R=run R_{prp}; prp=Property)

discarding the annotations extracted for head constituents causes recall scores of R_{prp} over the subset of 200 target categories to decrease by 12.9%. Therefore, the ability of the proposed method to also annotate head constituents plays only a limited part in its superior recall relative to the baseline B_{wcn} in Table 4.

Classes of Errors: Among the errors affecting the quality of extracted properties, the most frequent is the non-optimal selection of a property, out of several available candidate properties. Since many of the descendant articles of the category *"ctg:1890s comics"* are topics introduced in that decade, the property *"prp:P571 (inception)"* is extracted, which is acceptable but may not be ideal. Similarly, the property *"prp:P20 (place of death)"* is extracted for *"Mongol"* in *"ctg:Mongol khans"*, because of evidence in Wikidata that individual khans not only led but also often died in that ter-

ritory. For *"ctg:1990s Serbian television series endings"* and *"ctg:Thai historical films"*, most if not all individual descendant articles are about works (series or movies) in that language. Yet the two Wikipedia categories are primarily about works from that territory and not about works in that language, which means that the property extracted for *"Serbian"* and *"Thai"* should ideally be *"prp:P495 (country of origin)"* or similar; and not *"prp:P364 (original language of work)"*, which is actually extracted. This is also an infrequent case where annotating the same modifier constituent with more than one, instead of at most one, property might be useful. Among the small number of descendant articles available in Wikipedia for the category *"ctg:Carolina Panthers broadcasters"*, Wikidata properties and values do not mention *"Carolina Panthers"*, for some of them (*"art:Tim Brando"*, *"art:Roman Gabriel"*); and mention it occasionally (for *"art:Eugene Robinson"*) but with the property *"prp:P54 (member of sports team)"*. While it is not unusual for retired players to subsequently provide news coverage of their former teams, the property is strictly incorrect. A similar phenomenon causes the annotation *"prp:P19 (place of birth)"* to be extracted for *"Yozgat"* in *"ctg:People from Yozgat Province"*. Such errors are arguably more serious, since not only is the annotation incorrect but the modifier constituent (*"Yozgat"*) is also incorrectly selected. The occurrence of errors does not preclude correct annotations from being extracted for other modifier constituents: *"Yozgat Province"* is simultaneously and correctly annotated as *"prp:P19 (place of birth)"*.

Table 6 shows modifier constituents from the gold evaluation set annotated only by run B_{wcn}, in the upper portion; or only by R_{prp}, in the lower portion. More modifier constituents are annotated by run R_{prp} alone than by run B_{wcn} alone. The most common cause of R_{prp} failing to extract any annotations are missing properties and values in Wikidata, particularly when the categories have only a small number of descendant articles in Wikipedia. For the category *"ctg:Probinsya Muna Development Initiative politicians"*, none of the Wikidata properties of the few descendant articles in Wikipedia (e.g., *"art:Antonio Cuenco"*) refer to the relevant political party, namely to *"Probinsya Muna Development Initiative"*.

Impact of More Supporting Articles: The article *"art:Gary Oldman"* is a descendant of the category *"20th-century English male actors"*. Since the article has the property-value pair *"prp:21 (sex or gender)"* and *"tpc:Male"* in Wikidata, it is in the article support set for assigning the property *"prp:21 (sex or gender)"* as an annotation of the modifier *"male"* within the category. When multiple candidate properties are available for a modifier constituent of a category, the candidate property with the largest article support set is selected as the property annotation. Intuitively, the selection of the property annotation is expected to be more vs. less reliable, depending on whether the counts of Wikipedia articles supporting the various candidate properties are larger or smaller.

Figure 3 investigates the phenomenon, by requiring the count of supporting Wikipedia articles of a candidate property to be larger than a minimum count, in order for the candidate property to be considered for extraction. In the figure, increasing the minimum count leads to more reliable candidate property annotations and higher precision in the left graph. As expected, increasing the minimum count also causes significant loss in re-

Run (Cnt): Examples of Modifier Constituents
B_{wcn} (175): 1672 treaties; 3 ft 6 in gauge railways in Sierra Leone; Agriculture companies of Spain; Charleston Alley-Cats players; Earl Scruggs songs; Fossil fuel power stations in Pakistan; Hittite dictionaries; Probinsya Muna Development Initiative politicians; South Sudanese people in sports; Verve Records remix albums
R_{prp} (401): Oregon elections, 1882; Arab architects; People from Bangalore Urban district; Rivers of Cascade County, Montana; Crawley Down Gatwick F.C. players; Songs written by Irving Gordon; Later Yan people; Melodic death metal albums; Mercyhurst Lakers women's ice hockey; Philadelphia Police Department officers; Renaissance Revival architecture in Indiana

Table 6: Examples of modifier constituents (underlined) within categories from the gold evaluation set, for which some annotation(s) are extracted only by B_{wcn} vs. only by R_{prp} (Cnt=total count of unique such modifier constituents from the gold evaluation set, for the respective run)

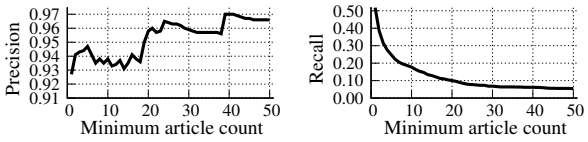

Figure 3: Macro-averaged precision (left graph) and recall (right graph) of run R_{prp}, as a function of the minimum count of supporting Wikipedia articles. An annotation is extracted for a Wikipedia category only when the number of supporting Wikipedia articles exceeds a given minimum on the horizontal axis.

call in the right graph. As more data is gradually added to Wikidata over time, the proposed method is likely to select from among candidate properties with gradually more supporting Wikipedia articles, which could lead to gradually higher precision of extracted annotations, according to Figure 3.

Extraction in Other Languages: Extending the proposed method to languages other than English depends on the availability of resources in those languages. First, alternative names of Wikidata topics in other languages would be useful, to align ngrams of modifiers and values, as described in Section 2. Possible sources of such alternative names are titles of non-English Wikipedia articles equivalent to the Wikidata topics; and non-English topic names and aliases, if any, already available in Wikidata. Second, flexible ngram matching in other languages would be useful, similarly to how lemmas or stems are useful in English, as described in Section 3. Stemming and lemmatization may be available in some languages. In others, flexible ngram matching pairs could be collected from hyperlinks internal to Wikipedia. For example, "canadienne" and "canadien" are the anchor text of hyperlinks within the French articles titled *"Deborah Ellis"* and *"Yann Martel"* respectively. Both hyperlinks point to the French article titled *"Canada"*. Being able to flexibly match the resulting ngram pairs *"canadienne"* vs. *"canada"*, or *"canadien"* vs. *"canada"*, would be useful in the annotation of categories such as *"ctg:Écrivain canadien"*.

5 Related Work

As it extracts semantic annotations over open-domain concepts (namely, over categories from Wikipedia), the proposed method falls under the area of open-domain information extraction (Ernst et al., 2018; Qu et al., 2018; Sun et al., 2018; Zhu et al., 2019; Zhan and Zhao, 2020; Dash et al., 2020; Cao et al., 2020). Previous work in that area often uses Wikipedia data (Tsurel et al., 2017; Konovalov et al., 2017; Korn et al., 2019; Bornemann et al., 2020).

In previous work, annotations for modifier constituents within compositional noun phrases may be extracted out of an unbound set of ambiguous strings, with no explicit semantics and possibly redundant (*"from"*, *"born in"*, *"born at"*) (Hendrickx et al., 2013; Nakov and Hearst, 2013). Alternatively, when annotations are selected out of a small, manually-created set of candidate annotations (Tratz and Hovy, 2010; Shwartz and Waterson, 2018), they are too coarse-grained to be equivalent to *born in* or *headquartered in* etc. The method introduced in (Paşca and Buisman, 2015) decomposes compositional Wikipedia articles into constituent Wikipedia articles. For example, it decomposes *"art:Swiss passport"* into *"art:Switzerland"*, *"art:Passport"*. It does not attempt to otherwise understand or annotate the semantics of the constituents. It is applicable only to Wikipedia articles, although many more Wikipedia categories are compositional.

The method in (Paşca, 2017) extracts annotations over child categories based on their parent categories in Wikipedia. The method produces superior annotations to previous efforts (Nastase and Strube, 2013) to annotate categories based on data within Wikipedia itself. It extracts annotations that are strings without any associated descriptions or disambiguation. In contrast, the method proposed here extracts annotations as properties (*"prp:P569 (date of birth)"*) with defined descriptions and semantic meaning in Wikidata. It also disambiguates modifier constituents to the corresponding Wikidata topics or, if available, to corresponding Wikipedia articles. Such annotations and disambiguation add a layer of semantic understanding to hierarchies of articles and categories extracted from Wikipedia (Flati et al., 2014; Gupta et al., 2018), wherein categories are otherwise represented only as strings.

6 Conclusions

This paper takes advantage of data from Wikidata, to extract annotations for understanding the role played by various constituents in determining the meaning of Wikipedia categories. Unlike in previous work, the annotations are semantically-anchored properties and values, rather than ambiguous strings. They offer a better trade-off between precision vs. recall. Current work explores the utility of alternative sources besides Wikidata, in increasing the coverage of the annotations; and the role of the annotations in generating plausible categories for Wikipedia articles.

Acknowledgments

The author would like to thank Erin Bennett and Travis Wolfe, for comments on an earlier version of the paper; and Erin Bennett, for assistance with the evaluation set.

References

C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. 2009. DBpedia - a crystallization point for the Web of data. *Journal of Web Semantics*, 7(3):154–165.

K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 International Conference on Management of Data (SIGMOD-08)*, pages 1247–1250, Vancouver, Canada.

L. Bornemann, T. Bleifuß, D. Kalashnikov, and F. Naumann an D. Srivastava. 2020. Natural key discovery in Wikipedia tables. In *Proceedings of the 2020 Web Conference (WWW-20)*, pages 2789–2795, Taipei, Taiwan.

E. Cao, D. Wang, J. Huang, and W. Hu. 2020. Open knowledge enrichment for long-tail entities. In *Proceedings of the 2020 Web Conference (WWW-20)*, pages 384–394, Taipei, Taiwan.

D. Chen, A. Fisch, J. Weston, and A. Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL-17)*, pages 1870–1879, Vancouver, Canada.

A. Chisholm, W. Radford, and B. Hachey. 2017. Learning to generate one-sentence biographies from Wikidata. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL-17)*, pages 633–642, Valencia, Spain.

S. Dash, F. Chowdhury, A. Gliozzo, N. Mihindukulasooriya, and N. Fauceglia. 2020. Hypernym detection using strict partial order networks. In *Proceedings of the 34th National Conference on Artificial Intelligence (AAAI-20)*, New York, New York.

F. Ensan and E. Bagheri. 2017. Document retrieval model through semantic linking. In *Proceedings of the 10th ACM Conference on Web Search and Data Mining (WSDM-17)*, pages 181–190, Cambridge, United Kingdom.

P. Ernst, A. Siu, and G. Weikum. 2018. HighLife: Higher-arity fact harvesting. In *Proceedings of the 2018 Web Conference (WWW-18)*, pages 1013–1022, Lyon, France.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database and Some of its Applications*. MIT Press.

T. Flati, D. Vannella, T. Pasini, and R. Navigli. 2014. Two is bigger (and better) than one: the Wikipedia Bitaxonomy project. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL-14)*, pages 945–955, Baltimore, Maryland.

A. Gupta, R. Lebret, H. Harkous, and K. Aberer. 2018. 280 birds with one stone: Inducing multilingual taxonomies from Wikipedia using character-level classification. In *Proceedings of the 32nd National Conference on Artificial Intelligence (AAAI-18)*, pages 4824–4831, New Orleans, Louisiana.

P. Gupta, S. Rajaram, H. Schütze, and T. Runkler. 2019. Neural relation extraction within and across sentence boundaries. In *Proceedings of the 33rd National Conference on Artificial Intelligence (AAAI-19)*, pages 6513–6520, Honolulu, Hawaii.

I. Hendrickx, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Szpakowicz, and T. Veale. 2013. SemEval-2013 task 4: Free paraphrases of noun compounds. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval-13)*, pages 138–143, Atlanta, Georgia.

J. Hoffart, F. Suchanek, K. Berberich, and G. Weikum. 2013. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence Journal. Special Issue on Artificial Intelligence, Wikipedia and Semi-Structured Resources*, 194:28–61.

A. Konovalov, B. Strauss, A. Ritter, and B. O'Connor. 2017. Learning to extract events from knowledge base revisions. In *Proceedings of the 26th World Wide Web Conference (WWW-17)*, pages 1007–1014, Perth, Australia.

F. Korn, X. Wang, Y. Wu, and C. Yu. 2019. Automatically generating interesting facts from Wikipedia tables. In *Proceedings of the 2019 International Conference on Management of Data (SIGMOD-19)*, pages 349–361, Amsterdam, Netherlands.

D. Ma, Y. Chen, K. Chang, and X. Du. 2018. Leveraging fine-grained Wikipedia categories for entity search. In *Proceedings of the 2018 Web Conference (WWW-18)*, pages 1623–1632, Lyon, France.

A. Moniruzzaman, R. Nayak, M. Tang, and T. Balasubramaniam. 2019. Fine-grained type inference in knowledge graphs via probabilistic and tensor factorization methods. In *Proceedings of the 2019 Web Conference (WWW-19)*, pages 3093–3100, San Francisco, California.

S. Murty, P. Verga, L. Vilnis, I. Radovanovic, and A. McCallum. 2018. Hierarchical losses and new resources for fine-grained entity typing and linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL-18)*, pages 97–109, Melbourne, Australia.

P. Nakov and M. Hearst. 2013. Semantic interpretation of noun compounds using verbal and other paraphrases. *ACM Transactions on Speech and Language Processing*, 10(3):1–51.

V. Nastase and M. Strube. 2013. Transforming Wikipedia into a large scale multilingual concept network. *Artificial Intelligence*, 194:62–85.

M. Paşca. 2017. German typographers vs. German grammar: Decomposition of Wikipedia category labels into attribute-value pairs. In *Proceedings of the 10th ACM Conference on Web Search and Data Mining (WSDM-17)*, pages 315–324, Cambridge, United Kingdom.

M. Paşca. 2018. Finding needles in an encyclopedic haystack: Detecting classes among Wikipedia articles. In *Proceedings of the 2018 Web Conference (WWW-18)*, pages 1267–1276, Lyon, France.

M. Paşca and H. Buisman. 2015. Dissecting German grammar and Swiss passports: Open-domain decomposition of compositional entries in large-scale knowledge repositories. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI-15)*, pages 896–902, Buenos Aires, Argentina.

T. Piccardi, M. Catasta, L. Zia, and R. West. 2018. Structuring Wikipedia articles with section recommendations. In *Proceedings of the 41st International Conference on Research and Development in Information Retrieval (SIGIR-18)*, pages 665–674, Ann Arbor, Michigan.

S. Ponzetto and M. Strube. 2007. Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI-07)*, pages 1440–1447, Vancouver, British Columbia.

M. Qu, X. Ren, Y. Zhang, and J. Han. 2018. Weakly-supervised relation extraction by pattern-enhanced embedding learning. In *Proceedings of the 2018 Web Conference (WWW-18)*, pages 1257–1266, Lyon, France.

L. Ratinov, D. Roth, D. Downey, and M. Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-11)*, pages 1375–1384, Portland, Oregon.

V. Shwartz and C. Waterson. 2018. Olive oil is made of olives, baby oil is made for babies: Interpreting noun compounds using paraphrases in a neural model. In *Proceedings of the 2018 Conference of the North American Association for Computational Linguistics (NAACL-HLT-18)*, pages 218–224, New Orleans, Louisiana.

M. Sun, X. Li, X. Wang, M. Fan, Y. Feng, and P. Li. 2018. Logician: A unified end-to-end neural approach for open-domain information extraction. In *Proceedings of the 11th ACM Conference on Web Search and Data Mining (WSDM-18)*, pages 556–564, Marina del Rey, California.

S. Tratz and E. Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, pages 678–687, Uppsala, Sweden.

D. Tsurel, D. Pelleg, I. Guy, and D. Shahaf. 2017. Fun facts: Automatic trivia fact extraction from Wikipedia. In *Proceedings of the 10th ACM Conference on Web Search and Data Mining (WSDM-17)*, pages 345–354, Cambridge, United Kingdom.

D. Vrandečić and M. Krötzsch. 2014. Wikidata: A free collaborative knowledge base. *Communications of the ACM*, 57:78–85.

W. Wu, H. Li, H. Wang, and K. Zhu. 2012. Probase: a probabilistic taxonomy for text understanding. In *Proceedings of the 2012 International Conference on Management of Data (SIGMOD-12)*, pages 481–492, Scottsdale, Arizona.

J. Zhan and H. Zhao. 2020. Span model for open information extraction on accurate corpus. In *Proceedings of the 34th National Conference on Artificial Intelligence (AAAI-20)*, New York, New York.

S. Zhang and K. Balog. 2018. Ad hoc table retrieval using semantic similarity. In *Proceedings of the 2018 Web Conference (WWW-18)*, pages 1553–1562, Lyon, France.

Q. Zhu, X. Ren, J. Shang, Y. Zhang, A. El-Kishky, and J. Han. 2019. Integrating local context and global cohesiveness for open information extraction. In *Proceedings of the 12th ACM Conference on Web Search and Data Mining (WSDM-19)*, pages 42–50, Melbourne, Australia.