Lightweight Lexical and Semantic Evidence for Detecting Classes Among Wikipedia Articles

Marius Paşca
Google
Mountain View, California
mars@google.com

Travis Wolfe
Google
Mountain View, California
twolfe@google.com

ABSTRACT

A supervised method relies on simple, lightweight features in order to distinguish Wikipedia articles that are classes (*"Shield volcano"*) from other articles (*"Kilauea"*). The features are lexical or semantic in nature. Experimental results in multiple languages over multiple evaluation sets demonstrate the superiority of the proposed method over previous work.

CCS CONCEPTS

• **Information systems** → Content analysis and feature selection; • **Computing methodologies** → Information extraction; Lexical semantics;

KEYWORDS

Knowledge acquisition, open-domain information extraction, topic classification, classes, semantics

ACM Reference Format:
Marius Paşca and Travis Wolfe. 2019. Lightweight Lexical and Semantic Evidence for Detecting Classes Among Wikipedia Articles. In *The Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19), February 11–15, 2019, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3289600.3291020

1 INTRODUCTION

Motivation: Enjoying continuous growth through collaborative contributions by human editors in hundreds of languages, Wikipedia is a key resource in efforts to organize the world's knowledge into large, open-domain knowledge repositories. A variety of knowledge repositories [1, 17, 26], including Freebase [3], Wikidata [43], Knowledge Graph [39] and Concept Graph [46], derive at least their initial, core knowledge from semi-structured textual content available in Wikipedia articles. Wikipedia and knowledge repositories derived from it are useful in a variety of tasks pertaining to knowledge acquisition from text [17, 25, 45–47], text analysis [24, 34, 35] and information retrieval [4, 8, 18, 22, 38, 41] including commercial Web search, helping to potentially transform search results from sets of hyperlinks to relevant documents into sets of concepts directly relevant to users' queries [39].

Most Wikipedia articles correspond to concepts that are instances (*"Kilauea"*) as opposed to classes (*"Volcano"*), in part due to the encyclopedic nature of Wikipedia. But many language editions of Wikipedia contain hundreds of thousands of articles each. This makes the subset of Wikipedia articles that are likely classes significant in size, even when compared to resources such as WordNet [12], which focus specifically on representing not instances but classes. Neither Wikipedia nor other larger knowledge repositories derived from it distinguish articles that are classes.

Contributions: The method proposed in this paper relies on simple, lightweight lexical features collected from the text of Wikipedia articles, as well as semantic features from outside of Wikipedia, as evidence towards detecting a subset of articles that are classes. The features are applicable to English and other languages. They are inexpensive to collect. The features do not require linguistic preprocessing tools such as part of speech taggers, named entity recognizers, syntactic or semantic parsers. They are either lexical, expected to apply horizontally, widely across articles independently of their domains; or semantic (knowledge-based), expected to apply vertically, narrowly only within limited domains. Over various combinations of existing evaluation sets being used as training vs. test data, the method acquires classes at better levels of trade-off between precision and recall than those achieved by a recently-introduced method.

2 DETECTION OF CLASSES

2.1 Task

Classes: Classes are placeholders for sets of instances that share common properties. A class such as *"Shield volcano"* is a placeholder for a set of instances such as *"Kilauea"* and *"Hofsjökull"*. In contrast, *"Kilauea"* is an instance and not a class, since it cannot act as a placeholder for any other set of its own instances. Classes (*"Shield volcano"*) may be specializations of other classes (*"Volcano"*), if the instances of the former share more properties, in addition to the properties shared by the instances of the latter. Through specialization, classes effectively organize the set of all possible instances into a hypothetical conceptual hierarchy, whose leaf concepts at the bottom are instances, whereas intermediate concepts would be iteratively more general classes.

Task: As a consequence of its encyclopedic nature, the very large majority of articles in Wikipedia correspond to concepts that are instances (*"Kilauea"*, *"Hofsjökull"*) as opposed to classes (*"Shield volcano"*). In random samples, as many as 97 out of 100 Wikipedia articles may be instances [27]. But even if instances dominate, that still leaves what we estimate to be low hundreds of thousands of articles in Wikipedia that might be classes. That is a much larger

WSDM '19, February 11–15, 2019, Melbourne, VIC, Australia

ACM ISBN 978-1-4503-5940-5/19/02…$15.00
https://doi.org/10.1145/3289600.3291020

and finer-grained potential set of classes than, e.g., the small number of types available in the type hierarchy in Freebase [3]. Unfortunately, Wikipedia does not distinguish articles that are classes from those that are not. Neither do large knowledge graphs such as [39, 46], which otherwise rely heavily on creating and maintaining internal concepts for most if not all Wikipedia articles. The goal of the method being proposed here is the selection of as many Wikipedia articles that are classes as accurately as possible, out of all Wikipedia articles.

2.2 Applications

Enriching Knowledge Repositories: The detection of Wikipedia articles that are classes can be applied immediately and transitively to concepts in large knowledge repositories [3, 39, 46] that were created from and correspond to those Wikipedia articles.

Expansion of Lexical Dictionaries: Due to the high cost of manual maintenance and expansion, valid open-domain concepts may be missing from expert-created lexical resources like WordNet. Wikipedia articles extracted as classes represent an inexpensive source of high-quality candidate concepts (e.g., *"Wooden roller coaster"*, *"Polar filament"*) for manual insertion into future, expanded versions of such lexical resources.

Topic Decomposition: By decomposing potentially compositional Wikipedia articles (e.g., *"Photochemical logic gate"*) into one or more constituent articles (*"Photochemistry"*, *"Logic gate"*), the meaning of a compositional article can be approximately defined in aggregate from the meaning of the constituent articles [6]. Existing methods [28] for decomposing Wikipedia articles lack "additional signals to better distinguishing between fully compositional and non-compositional" articles (cf. [28]). Wikipedia articles that are classes contribute towards such a signal. Articles that are classes are more likely to be compositional, whereas articles that are not classes (*"Joseph Lovering House"*) are more likely to be non-compositional.

Wikipedia Hierarchies: Edges in hierarchies constructed over Wikipedia articles [13, 16] are hypernymy relations from more specific articles to more general categories or articles. However, recent hierarchies do not contain edges from *"Mahmoud Hashemi Shahroudi"* and *"Jaldapara National Park"*, on one hand, to *"Chief Justice of Iran"* and *"Reserve forest"*, on the other hand. In fact, the articles *"Chief Justice of Iran"* and *"Reserve forest"* are not among the intermediate nodes in the hierarchies. But they are extracted as classes by the method being proposed here (e.g., based on the presence of the plural-form category *"Category:Chief Justices of Iran"* for the article *"Chief Justice of Iran"*), thus pointing to potential gaps in hierarchies extracted from Wikipedia.

3 METHOD

3.1 Lexical Features Within Wikipedia

Clues in Wikipedia Articles: By design, the proposed method for detecting Wikipedia articles that are classes relies on simple, shallow analysis of occurrences of the article title (*"Shield volcano"*) within the article text. The analysis consists simply in searching, among such occurrences, for three types of clues: 1) contexts surrounding the occurrences in article text, which match one of a few, simple contextual patterns; 2) morphological variation among different occurrences; and 3) presence of lowercase occurrences. The clues apply to English as well as to other, though not all, languages.

Lexical Clue 1: Pre-defined contextual patterns: The context around an occurrence of the article title appears in text is sometimes strongly suggestive of the article title being a class. Specifically, an occurrence (*"[..] shield volcano [..]"*) preceded by an indefinite article (*"[..] a shield volcano [..]"*) may indicate that the occurrence is a countable noun, especially if the sequence is part of what might be a definition of the concept described in the article (*"[..] A shield volcano is a type of [..]"*). It is not unreasonable to expect such definitions to appear at least in some Wikipedia articles, either towards the beginning of the article, to quickly define the concept before diving into its details; or later in the article. Concretely, if the case-insensitive occurrence O of the article title in a sentence from the article is such that it matches one of the following language-specific patterns, then the occurrence is taken as evidence that the article (*"Shield volcano"*) might be a class:

En: [a|an] O [was|is] || *Fr*: [un|une] O [était|est] || *Es*: [un|una] O [fue|es]

Lexical Clue 2: Morphological variation: Countable nouns often have different singular vs. plural forms. The article title may occur in the text within the article in plural (*"[..] autres volcans boucliers sont présents [..]"*) or singular form (*"[..] Le plus grand volcan bouclier connu [..]"*) or both. Such morphological variation is taken as evidence that the article (*"Volcan bouclier"*) might be a class. Referring to a concept in plural implies that there can be more than one instance of the concept, which suggests that the concept is a class. Conversely, the absence of plural forms (*"Kilaueas"*) suggests that the concept (*"Kilauea"*) is not a class.

Lexical Clue 3: Capitalization: Lowercase occurrences of the article title within sentences (*"[..] es un volcán en escudo que [..]"*) suggest that the article (*"Volcán en escudo"*) might be a class. In contrast, mixed-case occurrences (*"[..] Con el tiempo, Kilauea se edificó [..]"*) suggest that the article (*"Kilauea"*) might not be a class.

Features from Wikipedia Articles: From the three types of clues, several counts are computed as features for each Wikipedia article, over the occurrences of the article title within the text of the article: a) C_1(contextual pattern match) is the count of case--insensitive occurrences that match a pre-defined contextual pattern; b) C_2(identity), C_3(plural) are the counts of case-insensitive occurrences in identical vs. plural form; c) C_4(mixedcase), C_5(lowercase) are the counts of case-sensitive occurrences in mixed case vs. lowercase; d) C_6(mixedcase plural), C_7(lowercase plural) are the counts of case-sensitive occurrences of plural forms in mixed case vs. lowercase; and e) C_8(plural category) is the count of case-insensitive parent Wikipedia categories [32] of the article (*"Shield volcano"*) that are plural forms (*"Category:Shield volcanoes"*) of the article title. A few binary features are also derived from existing count features that are expected to provide relatively stronger (i.e., reliable) clues that the article is a class: f) is-positive(C_1(contextual pattern match)) and is-positive(C_7(lowercase plural)) are set to 1 or 0, depending on whether the respective counts are positive or 0. Since articles have at most one similarly-titled parent category in plural form in Wikipedia, the C_8(plural category) already acts as a binary feature. The features are inspired by rules and heuristics proposed in [26, 27].

In order to compute the features described above from Wikipedia articles in a given target language, language-specific prerequisites consist in the creation of contextual patterns, possibly using existing patterns in other languages as inspiration; and the identification of frequent, though not necessarily complete, rules for plural noun formation in the target language. The pre-requisites are relatively inexpensive and easy to satisfy for many target languages other than English. Once the pre-requisites have been satisfied, the actual automatic computation of the set of simple, numerical features from each Wikipedia article in the target language is lightweight, inexpensive and fast. It does not require any linguistic processing tools to be available in the target language, whether for part of speech tagging, parsing etc.

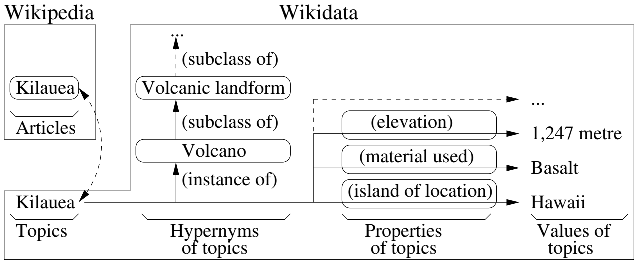

Figure 1: Examples of hypernym topics and properties available in Wikidata for Wikipedia articles

3.2 Semantic Features Outside Wikipedia

Lexical vs. Semantic Features: Intuitions presented and features collected so far are lexical. They apply horizontally, across all Wikipedia articles. They are independent from the underlying domains (e.g., geology vs. entertainment vs. finance) or categories (e.g., volcanoes vs. actors vs. index funds) under which the various topics might belong. In contrast, semantic features do not generalize across domains or categories. Instead, they are expected to apply only to possibly-narrow, vertical slices through the space of all topics. With an eye towards existing large knowledge repositories and the kind of knowledge they contain, there are at least two types of semantic (knowledge-based) clues to consider.

Semantic Clue 1: Hypernyms: If hypernym topics were available for a Wikipedia article, they would not necessarily be expected to help much in deciding whether the topic is an instance or class. Indeed, classes and instances may easily share hypernyms, such as *"Kilauea"* and *"Shield Volcano"* sharing the hypernym *"Volcano"*, making the presence of a particular hypernym unreliable evidence, at best, on whether the topic is or is not a class. However, if an article (*"Kilauea"*) were known to be an InstanceOf of a hypernym (*"Volcano"*), then the presence of the hypernym could be useful in deciding whether the article is an instance or a class. Such hypernyms, if available for a Wikipedia article, are taken as evidence towards determining whether the article might be a class.

Semantic Clue 2: Properties: Similarly to hypernyms, the presence of certain properties known to apply to a Wikipedia article could be relevant, even if only in a narrow domain rather than across domains. Topics are likely to be instances and not classes, if they are known to have properties such as being *located* in a particular location such as *"Hawaii"*; or to have a certain *date of birth* such as *"1936"*; or be associated with a certain *record label* such as *"Armada Music"*. Properties, if available for a Wikipedia article, are taken as evidence towards determining whether the article might be a class.

Source of Semantic Features: Considering Wikipedia and other knowledge repositories derived from Wikipedia, a prominent candidate source of hypernyms and properties of Wikipedia articles is Wikidata, for several reasons. First, Wikidata marks any Wikipedia articles that are equivalent to Wikidata topics explicitly, making it trivial to find Wikidata topics equivalent to Wikipedia articles and vice versa. Second, just like Wikipedia, Wikidata is an actively developed, growing resource that benefits from continuous human editing and curation. Third, Wikidata not only contains hypernyms explicitly, but distinguishes between InstanceOf and SubclassOf relations. Since its early stages, InstanceOf relations have been well represented among other types of relations in Wikidata [43]. In comparison, in Wikipedia, parent categories and transitively ancestor (parents of parents) categories may (*"Category:Active volcanoes"*, *"Category:Volcanoes"*) or may not (*"Category:Cenozoic Hawaii"*, *"Category:Earth"*) be hypernyms of a child Wikipedia article (*"Kilauea"*) [13, 16, 31, 32]. Since hypernym categories are not distinguished in any way from other categories in Wikipedia, Wikipedia categories are not an adequate source of hypernyms, let alone of InstanceOf hypernyms, of a Wikipedia article. Fourth, properties are explicitly available in Wikidata, in the form of predicates (*record label*) of relations connecting topics (*"Armin van Buuren"*) to other topics (*"Armada Music"*) in Wikidata. For these reasons, semantic features are derived from Wikidata.

Semantic Features from Wikidata: As shown in Figure 1, given a Wikidata topic (e.g., *Q188698* corresponding to *"Kilauea"*), the set of its InstanceOf hypernyms initially consists of Wikidata topics (*"Volcano"*) that are right arguments of InstanceOf relations, if any, whose left argument is the given topic. The set is then expanded, by transitively collecting other Wikidata topics (*"Volcanic landform"*, *"Landform"*) that are right arguments of SubclassOf relations in Wikidata, whose left arguments are hypernyms (*"Volcano"*) collected so far for the given topic. Thus, the path that connects a given Wikidata topic up to one of its InstanceOf ancestor topics consists of a first edge that must be an InstanceOf relation; and optionally additional edges that must be SubclassOf relations.

The set of properties of a Wikidata topic is the set of predicates of relations, if any, whose left argument is the given topic in Wikidata. The predicates InstanceOf and SubclassOf are excluded. It is only the predicates (*record label*), and not also the right arguments (*"Armada Music"*) of the relations, that are collected as properties. As long as a property applies and there is some value for it as the right argument, what the actual value is is deemed irrelevant.

The properties and InstanceOf hypernyms of a given Wikidata topic (*Q188698*) are transferred to the Wikipedia article (*"Kilauea"*) marked as equivalent to the Wikidata topic in Wikidata.

The set of all properties or InstanceOf hypernyms, collected for one or more Wikipedia articles, is converted into a set of Wikidata-based binary features computed for each Wikipedia article. Each of the binary features indicates whether the respective property or InstanceOf hypernym was or was not collected for the respective Wikipedia article. The set of binary features represent semantic features, which can be added to the lexical features.

4 EXPERIMENTAL SETTING

Supervised Learning: The sets of features associated with each Wikipedia article are the input to a linear classification algorithm with hinge loss as the choice of loss function, as implemented in Vowpal Wabbit [20]. Other loss functions or non-linear algorithms might be used. The features collected for various sets of Wikipedia articles are used as training or test data, as described later. An entry from the test data is classified as being a class vs. not a class, if the score computed by the classifier is positive vs. non-positive.

Data Sources: The experiments operate over the Wikipedia snapshot used in [27]. Disambiguation and redirect pages are discarded. Semantic features are extracted for each Wikipedia article from this snapshot, based on data from a snapshot of Wikidata from June 2018. For Wikipedia articles with no equivalent Wikidata topics in the Wikidata snapshot, their sets of semantic features are assumed to be empty.

Evaluation Sets: Three evaluation sets introduced in [27] serve as the source data for training and testing the proposed method. Each evaluation set consists in pairs of a Wikipedia article in English and a gold label indicating whether the article is a class or not.

The first evaluation set, $\mathbf{S}_W$, is derived from *Instance* relations available in WordNet [12]. The set collects the left (more specific) arguments (*"Mauna Loa"*) in such relations as gold non-classes (i.e., gold instances), and the right (more general) arguments (*"Volcano"*) as gold classes; and maps those arguments to equivalent Wikipedia articles, if any, based on a pre-existing, manually-created set of such mappings [27]. The second and third evaluation sets are random samples of Wikipedia articles annotated manually. The random samples are documents (articles) from Wikipedia drawn either uniformly ($\mathbf{S}_D$) or after query-based automatic filtering meant to reduce the effect of instances being much more numerous than classes in Wikipedia ($\mathbf{S}_Q$). The three sets contain 5,735 (S_W), 2,000 (S_D) and 1,000 (S_Q) entries, divided into 547 and 5,188 (S_W), 73 and 1,927 (S_D) and 362 and 628 (S_Q) gold classes and gold non-classes respectively (cf. [27] for more details on the evaluation sets).

Training and Test Sets: The evaluation sets are employed as training data or test data, in various possible combinations. For example, one possible combination is to employ S_W as training data and S_Q as test data. Individual entries in the data serve as positive examples, if their gold label is class; or negative examples, if their gold label is instance. In any combination, entries that may be shared among the training set and the test set are removed from the training set but retained in the test set. There are only a few such shared entries, namely 3, 7 and 8 entries, for the combinations of S_W and S_D, S_W and S_Q, and S_D and S_Q respectively. Thus, regardless of which evaluation sets are selected as training vs. test sets, no entries appear in both the training and test sets. At the same time, an evaluation set selected as a test set is always used in its entirety without changes, ensuring that any results computed over the evaluation set are directly comparable to results reported in previous work over the same evaluation set.

Extraction Parameters: As its occurrences are identified in the article text, the article title is first normalized, to remove portions within parentheses, thus converting *"Circuit (administrative division)"* into *"Circuit"* for that purpose. Such portions are not consistently present vs. absent in titles of Wikipedia articles; they are similarly removed in previous methods operating over Wikipedia data [13, 27, 32]. Depending on the feature being computed, the occurrences may be case-insensitive or case-sensitive. Features that involve plural forms are computed based on a few approximate, not necessarily complete, simple rules for plural formation in the respective target languages. For example, plural forms are often formed by adding the suffix *"-s"* in French, English and Spanish, or by adding the suffix *"-es"* in the latter two languages. In English, the rules are complemented by lemmatization data from WordNet [12], thus accommodating irregular plural forms like *"corpora"* for *"corpus"* in the article title *"Text corpus"*.

Only for semantic features, separately for each combination of a training set and a test set, features activated for fewer than five of the gold classes from the training set are discarded.

Training	Test	Scores over Test Set		
Set	Set	P	R	F
S_D	S_W	0.966	0.822	0.888
S_Q	S_W	0.947	0.828	0.883
$S_Q \cup S_D$	S_W	0.938	0.847	0.890
S_W	S_D	0.935	0.589	0.723
S_Q	S_D	0.935	0.589	0.723
$S_W \cup S_Q$	S_D	0.936	0.603	0.733
S_W	S_Q	0.943	0.776	0.852
S_D	S_Q	0.945	0.760	0.842
$S_W \cup S_D$	S_Q	0.946	0.779	0.855

Table 1: Precision and recall over various evaluation sets. Features are collected over English articles, for both training and test data (P=precision; R=recall; F=F$_1$-score)

5 EVALUATION RESULTS

5.1 Results with Lexical Features

Extraction over English Articles: Table 1 summarizes the performance of the proposed method, when trained and tested over features collected over Wikipedia articles in English. In the upper vs. middle vs. lower portions of the table, the test sets are the S_W, S_D and S_Q evaluation sets. The training sets are the other two evaluation sets or their union. Precision scores are above 0.9 across the various test sets. Recall scores over S_D are the lowest at around 0.6, followed by S_Q, which are lower than over S_W. The F$_1$-scores exhibit the same trend, exceeding 0.7 for S_D and reaching almost 0.9 for S_W. For each test set, combining both of the other evaluation sets into a single training set brings only a small improvement in F$_1$-scores, relative to using only one of other evaluation sets.

Extraction over Articles in Other Languages: Table 2 gives detailed scores when the proposed method is tested on target languages other than English, namely French (in the upper portion of the table) or Spanish (in the lower portion). Test data uses features collected over Wikipedia articles in the target language rather than in English. In contrast, for each target language, training data is collected either from English articles; or from articles in the same target language, namely French or Spanish. The former choice corresponds to what one may refer to effectively as cross-language training and testing, whereas the latter involves same-language

		Recall Scaled to Entire Test Set?					
		No			Yes		
Training	Test	Scores over Test Set			Scores over Test Set		
Set	Set	P	R	F	P	R	F
Train on French (Fr) articles, test on French (Fr) articles:							
S_D	S_W	0.979	0.558	0.711	0.979	0.448	0.614
S_Q	S_W	0.814	0.802	0.808	0.814	0.643	0.719
S_W	S_D	0.818	0.346	0.486	0.818	0.123	0.214
S_Q	S_D	0.611	0.423	0.500	0.611	0.151	0.242
S_W	S_Q	0.985	0.691	0.812	0.985	0.370	0.538
S_D	S_Q	0.988	0.438	0.607	0.988	0.235	0.379
(Avg)	(Avg)	0.866	0.543	0.667	0.866	0.328	0.476
Train on English (En) articles, test on French (Fr) articles:							
S_D	S_W	0.936	0.744	0.829	0.936	0.597	0.729
S_Q	S_W	0.912	0.751	0.824	0.912	0.603	0.726
S_W	S_D	0.818	0.346	0.486	0.818	0.123	0.214
S_Q	S_D	0.818	0.346	0.486	0.818	0.123	0.214
S_W	S_Q	0.985	0.696	0.816	0.985	0.373	0.541
S_D	S_Q	0.985	0.691	0.812	0.985	0.370	0.538
(Avg)	(Avg)	0.909	0.596	0.720	0.909	0.365	0.521
Train on Spanish (Es) articles, test on Spanish (Es) articles:							
S_D	S_W	0.988	0.379	0.548	0.988	0.310	0.472
S_Q	S_W	0.824	0.844	0.834	0.824	0.690	0.751
S_W	S_D	0.889	0.400	0.552	0.889	0.110	0.195
S_Q	S_D	0.435	0.500	0.465	0.435	0.137	0.208
S_W	S_Q	0.991	0.602	0.749	0.991	0.309	0.472
S_D	S_Q	1.000	0.269	0.424	1.000	0.138	0.243
(Avg)	(Avg)	0.854	0.499	0.630	0.854	0.282	0.424
Train on English (En) articles, test on Spanish (Es) articles:							
S_D	S_W	0.939	0.725	0.818	0.939	0.593	0.727
S_Q	S_W	0.914	0.732	0.813	0.914	0.599	0.724
S_W	S_D	0.889	0.400	0.552	0.889	0.110	0.195
S_Q	S_D	0.875	0.350	0.500	0.875	0.096	0.173
S_W	S_Q	0.991	0.602	0.749	0.991	0.309	0.472
S_D	S_Q	0.991	0.602	0.749	0.991	0.309	0.472
(Avg)	(Avg)	0.933	0.568	0.706	0.933	0.336	0.494

Table 2: Precision and recall over various evaluation sets. Features are collected over French or Spanish articles, for test data; and over either English or same-language (French or Spanish) articles, for training data (P=precision; R=recall; F=F$_1$-score; Avg=average over evaluation sets)

training and testing. The left portion of the table shows scores computed over (variable) subsets of the English articles in the evaluation sets used as test sets, which have equivalent articles in the target language, and therefore for which features could be indeed collected in the other language. Comparatively, in the right portion of Table 2, the denominator in computing recall scores is scaled up (that is, increased) to the entire evaluation set used as a test set. In other words, the set of false negatives ignores articles absent from the test set, in the left portion of the table; but does include them, in the right portion of the table. This scales recall scores down, and makes them directly comparable to recall scores obtained with other methods over the same evaluation sets as test sets.

Scores in Table 2 are generally lower for same-language than for cross-language training and testing. At first sight, this might be interpreted as an indication that the method does not perform as well in other languages as it does in English. However, closer inspection reveals that the lower same-language scores are likely due simply to less training data being available in that setting. Indeed, when collecting training data from French rather than English articles, the number of English articles in the training sets for which features can be collected via the target language is lower than the total number of entries in the respective training sets by 6%, for S_W; by almost 3 times, for S_Q; and by more than 4 times, for S_D. In other words, when collecting training data in the target language rather than in English, a large fraction of the evaluation sets used as training sets is lost, when training on S_D; but relatively little is lost, when training on S_W. Not surprisingly, in Table 2, switching from cross-language training and testing to same-language training and testing reduces recall the most precisely when training on S_D (and testing on S_W or S_Q). For example, in the case of testing in French, changing training data from English to French, while consistently testing in French, changes recall from 0.744 to 0.558, when training on S_D and testing on S_W.

Test Set	Method	Scores over Test Set		
		P	R	F
Extract/predict classes in En:				
S_W	B_{wd}	0.891	0.728	0.802
	B_{rb}	0.946	0.762	0.844
	L_{rn}(Train=S_D)	**0.966**	**0.822**	**0.888** (+5.2% over B_{rb})
S_D	B_{wd}	0.871	0.110	0.195
	B_{rb}	**0.939**	0.425	0.585
	L_{rn}(Train=S_W)	0.935	**0.589**	**0.723** (+23.5% over B_{rb})
S_Q	B_{wd}	0.969	0.434	0.599
	B_{rb}	**0.982**	0.610	0.753
	L_{rn}(Train=S_D)	0.966	**0.822**	**0.888** (+17.9% over B_{rb})
Extract/predict classes in En∪Fr∪Es:				
S_W	B_{wd}	0.891	0.728	0.802
	B_{rb}	**0.944**	0.802	0.867
	L_{rn}(Train=S_D)	0.912	**0.868**	**0.890** (+2.6% over B_{rb})
S_D	B_{wd}	0.871	0.110	0.195
	B_{rb}	**0.912**	0.425	0.579
	L_{rn}(Train=S_W)	0.900	**0.616**	**0.732** (+26.4% over B_{rb})
S_Q	B_{wd}	0.969	0.434	0.599
	B_{rb}	**0.980**	0.666	0.793
	L_{rn}(Train=S_D)	0.942	**0.812**	**0.872** (+9.9% over B_{rb})

Table 3: Precision and recall over various evaluation sets, for baseline methods (B_{wd}, B_{rb}) and for the proposed method (L_{rn}). For the proposed method, features are collected over English articles, for training data; and over English or French or Spanish articles, for test data (P=precision; R=recall; F=F$_1$-score; En=English; En∪Fr∪Es=union over multiple languages)

Comparison to Baseline Methods: The proposed, supervised-learning method (denoted $\mathbf{L}_{rn}$) is compared in Table 3 against several baselines.

The first baseline (denoted $\mathbf{B}_{wd}$) extracts a set of Wikipedia articles as classes, based solely on data available in Wikidata. In Wikidata, more specific topics are connected to more general topics via InstanceOf or SubclassOf relations. The baseline decides whether

Method	Source Language	Percentage of All English Articles	Source Language	Percentage of All English Articles
B_{wd}	En	0.75%	En∪Fr∪Es	0.75%
B_{rb}		1.50%		1.62%
L_{rn}		2.09% (B_{rb}+39%)		2.34% (B_{rb}+44%)

Table 4: Percentage (Pct) of all Wikipedia articles in English that are extracted as classes by baseline methods (B_{wd}, B_{rb}) and proposed method (L_{rn}). The extraction is based on evidence collected either from Wikipedia articles in English only (En), or from Wikipedia articles in multiple languages (En∪Fr∪Es). The proposed method is trained on S_W∪S_Q

a Wikipedia article is a class or not, based on whether the Wikipedia article is equivalent (in Wikidata) to a Wikidata topic that is the right (more general) argument of some InstanceOf or SubclassOf relation in Wikidata. For example, one of the InstanceOf relations in Wikidata connects the topic *Q308801* (*"Achtung Baby"*) to *Q482994* (*"Album"*). The baseline collects the Wikipedia article marked in Wikidata as equivalent to the right argument *Q482994* of the relation, namely the article *"Album"*, as being a class. A possible variant would be to additionally require the collected Wikipedia article to appear as the right argument in a minimum number of distinct InstanceOf relations, with higher thresholds expected to give higher precision at the expense of lower recall.

The second baseline method (denoted $\mathbf{B}_{rb}$), introduced in [27], is a rule-based method that, based on occurrences of the title of a Wikipedia article in the article text, decides whether the article is a class or not. Note that, by comparing against this baseline, the proposed method is also transitively compared to a series of other baselines against which the baseline itself was compared in previous work; cf. [27] for descriptions of those other baselines (e.g., [26, 50]) and their scores on the same evaluation sets.

In the upper portion of Table 3, the methods rely on evidence collected from English articles. In the case of the proposed method, this means training and testing over features collected from English articles. Between the two baselines, B_{wd} gives lower scores than B_{rb}. The gap in recall scores is narrower over the S_W evaluation set but wider over S_Q and especially S_D. It suggests that B_{wd} can more easily extract relatively more general classes (*"University"*) but has limited utility in extracting relatively more specific classes (*"Neutron research facility"*). Between the baselines and the proposed method, although the proposed method gives generally lower precision scores than the B_{rb} baseline, it compensates with disproportionately higher recall scores. It produces consistently higher F_1-scores across the evaluation sets.

In the lower portion of Table 3, the methods take advantage of evidence collected simultaneously from English, French and Spanish (En∪Fr∪Es), in order to extract (or predict, in the case of the proposed method) articles in English that are classes. For this purpose, articles extracted as classes in each of the other target languages are first mapped to their equivalent articles in English, if any (or discarded, otherwise); then merged (via union) together with the articles extracted in English. For example, if a method extracted the articles *"Bridge"* in English, *"Volcan bouclier"* in French and {*"Escudo de red"*, *"Furgoneta"*} in Spanish as classes, then the articles it would extract in English based on all languages simultaneously would be {*"Bridge"*, *"Shield volcano"*, *"Van"*}. For the B_{wd} baseline, extraction based on multiple languages is in fact no different than extraction from English alone. Indeed, Wikidata marks a Wikipedia article in English as equivalent to a topic in Wikidata, regardless of whether or how many Wikipedia articles in other languages are also equivalent to the same Wikidata topic. For the B_{rb} baseline as well as the proposed method, extraction based on multiple languages slightly increases the scores in Table 3. The proposed method gives higher F_1-scores than the baselines B_{rb} and B_{wd}.

Wikipedia Articles in English Extracted as Classes
Category:Astronomical objects (6 child articles):
B_{rb}: (none)
L_{rn}: {Deep-sky object}
Category:Healthcare occupations (122 child articles):
B_{rb}: {Polysomnographic technologist}
L_{rn}: {Assistive technology service provider, Barefoot doctor, Community health agent, Community health worker, Feldsher, Health visitor, Licensed practical nurse, Medical radiation scientist, Medical scribe, Mid-level practitioner, Music therapy, Nursing, Online doctor, Orderly}
Category:Journalism occupations (27 child articles):
B_{rb}: (none)
L_{rn}: {Community correspondent, Food critic, News presenter, Rewrite man}
Category:Towers (55 child articles):
B_{rb}: (none)
L_{rn}: {Flak tower, Lantern tower, Moonlight tower}
Category:Types of restaurants (103 child articles):
B_{rb}: (none)
L_{rn}: {Automated restaurant, Chifa, Concession stand, Cosplay restaurant, Dinner theater, Greasy spoon, Hofbrau, Juke joint, Kopi tiam, Paladar, Snack bar, Soda shop, Underground restaurant}

Table 5: Complete sets of English articles listed in Wikipedia as child articles of various parent categories, which are extracted as classes only by the baseline method B_{rb} or only by the proposed method L_{rn} trained on S_W∪S_Q

Absolute Recall: Going beyond the evaluation sets used in experiments so far, Table 4 evaluates the impact of the proposed method (and the baseline methods) going back to the goal stated early on, namely identifying as many Wikipedia articles that are classes as accurately as possible. The table ignores accuracy and focuses instead on absolute recall. It shows the fraction of all Wikipedia articles in English that are identified as classes. The proposed method has a significant advantage over the baselines in extracting more articles as classes, both when using evidence only in English and when using evidence in multiple languages. The table confirms that the proposed method has better recall than the baselines B_{wd} and B_{rb}, while other experimental results presented so far separately show that the B_{rb} baseline has no significant precision advantage over the proposed method.

Per-Category Comparative Recall: In another practical comparison beyond the existing evaluation sets, Table 5 shows the complete sets of articles extracted as classes only by the B_{rb} baseline or only by the proposed method, out of all child articles listed directly under a few categories in Wikipedia. For example, out of the

6 child articles of the parent category *"Category:Astronomical objects"*, only the proposed method extracts the article *"Deep-sky object"* as a class. Note that the selected parent categories are not guaranteed to have only classes as child articles. For example, one of the child articles of the category *"Category:Towers"* is *"Tower of Elahbel"*, which is not a class. In fact, another category *"Category:21st-century actresses"* is selected precisely because most, if not all, of its child articles are expected to not be classes, e.g., *"Jelena Jovanova"*. Extracting any of them as classes would likely be incorrect. In a positive sign for either method relative to the other, none of them extracts any additional (likely incorrect) child articles of the category *"Category:21st-century actresses"* as classes. For other categories in Table 5, the proposed method often manages to extract classes that the B_{rb} baseline cannot. The classes extracted only by the proposed method look encouragingly accurate.

Relative Feature Contribution: Separate ablation experiments temporarily disable subsets of features pertaining to a) contextual patterns (C_1, is-positive(C_1)); b) morphological variation (C_3, C_6, C_7, C_8, is-positive(C_7)); and c) capitalization (C_5, C_7). When compared to enabling all features, ablation reduces F_1-scores in English by 15.5% (a), 21.4% (b) or 21.9% (c) respectively on average, when using either S_Q or S_W as training set, and S_D as test set. The results show that the different types of clues and associated features all contribute towards overall performance.

Discussion: As it computes features based on occurrences of the article title in the text of the article, the proposed method assumes that such occurrences exist and are indeed mentions of the concept being described in the article. The method is likely to perform worse when either of these assumptions does not hold. First, for shorter articles, too many features collected from the articles may be zero, for them to be useful in determining whether the articles are classes or not. If the articles are equivalent to other, better fleshed-out articles in other languages, they may still be extracted as classes based on evidence in those other languages. Although the English article *"Vereda"* is not extracted as a class based on evidence in English, it is still extracted as a class based on evidence in Spanish, since its equivalent article *"Vereda"* is extracted as a class in Spanish. Conversely, although *"Abbesse"* is not extracted as a class in French based on evidence in that language, it is still extracted as a class based on evidence in English, via its equivalent English article *"Abbess"*. But for articles (*"Trade item"*) with no equivalent articles in other languages, the lack of enough evidence towards the computed features remains an issue. Second, occurrences of the article title in article text are sometimes not really mentions of that concept but rather of some other concepts. The occurrences may be incorrectly interpreted as presence of any of the three types of lexical clues, namely contextual patterns, in *"[..] A Funky Situation is the 21st studio album [..]"* (for *"Funky Situation"*); morphological variation, in *"[..] The UFOs can survive for far longer [..]"* (for *"UFO (TV series)"*); or capitalization, in *"[..] from scratch in a simple editor that is part of Scratch [..]"* (for *"Scratch (programming language)"*).

5.2 Results with Semantic Features

Impact of Features from Wikidata: Table 6 summarizes the impact of semantic features derived from Wikidata, on top of lexical features from Wikipedia already used in earlier experiments.

Train Set	Test Set	Enabled Features			Scores over Test Set		
		F_{lex}	F_{spr}	F_{shp}	P	R	F
S_D	S_W	√	-	-	0.966	0.822	0.888
		√	√	-	0.972	0.820	0.890
		√	-	√	0.970	0.820	0.889 (0.9% Err over F_{lex})
S_Q	S_W	√	-	-	0.947	0.828	0.883
		√	√	-	0.943	0.839	0.888
		√	-	√	0.958	0.835	0.892 (8.3% Err over F_{lex})
S_W	S_D	√	-	-	0.935	0.589	0.723
		√	√	-	0.933	0.575	0.712
		√	-	√	0.938	0.616	0.744 (8.2% Err over F_{lex})
S_Q	S_D	√	-	-	0.935	0.589	0.723
		√	√	-	0.898	0.603	0.721
		√	-	√	0.935	0.589	0.723 (0.0% Err over F_{lex})
S_W	S_Q	√	-	-	0.943	0.776	0.852
		√	√	-	0.956	0.773	0.855
		√	-	√	0.944	0.790	0.860 (5.7% Err over F_{lex})
S_D	S_Q	√	-	-	0.945	0.760	0.842
		√	√	-	0.962	0.760	0.849
		√	-	√	0.958	0.757	0.846 (2.6% Err over F_{lex})

Table 6: Impact on precision and recall of using Wikidata-based semantic features, in addition to existing lexical features. Features are collected over English articles, for both training and test data (F_{lex}=lexical features; F_{spr}=Wikidata-based properties as semantic features; F_{shp}=Wikidata-based hypernyms as semantic features; P=precision; R=recall; F=F_1-score; Err=error rate reduction)

Not shown in the table, scores with semantic features F_{shp} or F_{spr} alone, while disabling lexical features F_{lex}, are lower than with F_{lex} alone. Without lexical features and with semantic features limited to only hypernyms (F_{shp}) or only properties (F_{spr}), F_1-scores range between 0.015 and 0.413 (for F_{shp}) and between 0.026 and 0.552 (for F_{spr}). Adding semantic features from Wikidata properties (F_{spr}) causes inconsistent changes to scores, with the overall effect appearing to be negative rather than positive. Comparatively, adding semantic features derived from Wikidata hypernyms (F_{shp}) consistently gives identical or improved F_1-scores, with commensurately small reduction in error rates, depending on the combination of evaluation sets employed as training data or test data. While improvements are small, they build on top of results of the proposed method with lexical features (F_{lex}) alone, which are already superior to results reported earlier for the B_{rb} baseline.

Discussion: Despite its construction through manual editing, data in Wikidata can still cause confusion. The Wikidata topics *Q1482076* and *Q1632287* are equivalent in Wikidata to the Wikipedia articles *"Propaganda model"* and *"Caster board"*. But the topics are simultaneously the left arguments of InstanceOf relations in Wikidata, connecting them to *Q571* (*"book"*) and *Q5398426* (*"television series"*) respectively. Most likely, either the associated Wikipedia articles or the InstanceOf relations or both are incorrect in Wikidata. In another example, the Wikidata topic *Q1026252* is an InstanceOf *"school"* and has the property *inception* (with the value *"1979"*). But its equivalent Wikipedia article *"Calandreta"* is not about one particular instance but about a class (group) of schools of a certain kind. The property *named after* might seem like a useful

Examples of Wikipedia Articles
No longer decomposed: Advances in Applied Mathematics, Four Seasons Hotel Moscow, Joseph Lovering House, Lockport Industrial District, Tokyo Seitoku University, Treaty of San Francisco
Still decomposed: Bishop of Hereford, Chromatic button accordion, Photochemical logic gate, Photographic lens design, Treadle pump, United Nations General Assembly resolution, Vacuum flange

Table 7: Examples of Wikipedia articles that are either no longer or continue to be decomposed by the method from [28], based on the absence vs. presence of the decomposed articles among the classes extracted by the proposed method when trained on $\mathbf{S}_W \cup \mathbf{S}_Q$

clue that the topic to which it applies is an instance. But the Wikidata topic *Q318541* has this property (with the value *"Daniel Kirkwood"*), although its equivalent Wikipedia article *"Kirkwood gap"* is clearly about a class and not an instance. In a somewhat similar phenomenon, this time affecting hypernyms instead of properties, both topics *Q832799* (*"Yellow Submarine"*) and *Q922853* (*"Campaign song"*) have the InstanceOf hypernym *Q7366* (*"Song"*) in Wikidata, although the Wikipedia article equivalent to *Q922853* is clearly about a class of songs and not about a particular instance.

Lexical and semantic features are inexpensive to compute. The former capture phenomena such as morphology and case. The latter are available thanks to the significant amount of decentralized human labor driving the growth and maintenance of Wikidata. Although semantic features alone are not as useful, their addition to lexical features does add value, albeit small, on top of lexical features. Semantic features may still be useful if more training data became available. With or without semantic features, the results given by lexical features alone, which are superior to baselines such as B_{rb}, are encouraging.

5.3 Impact on Downstream Applications

Topic Decomposition: By restricting the decomposition method from [28] to only decompose Wikipedia articles that are extracted as classes by our proposed method, the precision of decomposed articles that are indeed compositional increases. Table 7 shows examples that the method from [28] still or no longer decomposes, based on their presence among extracted classes.

Class Categories and Article Hierarchies: Applications may benefit from Wikipedia categories [16, 22] in addition to articles. Articles extracted as classes by the proposed method can be used to extract Wikipedia categories that are classes. A simple procedure is as follows. First, the subset of all categories (*"Category:Albums"*) whose names are identical to, or plural forms of, one of their child Wikipedia articles (*"Album"*) that are extracted as classes are themselves extracted as classes. Second, their descendant categories (recursively children of child categories) such as *"Category: Armada Music albums"*, whose name chunks (*"albums"*) are identical to the case-insensitive ancestor category name, are also extracted as classes. The procedure produces almost twice as many categories as classes, as there are unique hypernym categories in a Wikipedia hierarchy from articles to categories constructed in [16]. Examples of categories extracted as classes here, but not extracted as hypernyms in [16], are *"Category: Greek trance musicians"* and *"Category: 20th-century Argentine painters"*.

6 RELATED WORK

WordNet [12] distinguishes classes from other concepts, by manually connecting instance concepts (synsets) up to their more general concepts through an (*Instance*) relation rather than a general-purpose hypernymy relation. Implicitly, WordNet concepts linked up through hypernymy rather than *Instance* relations are classes, just as concepts that are hypernyms of some other concepts are also classes. Cyc [21] makes distinctions between *SetOrCollections*, or classes, and *Individuals*. But only thousands of Cyc concepts are equivalent to any Wikipedia articles [50].

Wikipedia [36] does not distinguish articles that are classes. Articles are organized into fine-grained categories, which in turn are organized into iteratively coarser-grained categories. Collecting Wikipedia categories as classes would not be an adequate solution to identifying classes in Wikipedia. Wikipedia categories often do not correspond to classes. An intermediate step in [30] aims at distinguishing such Wikipedia categories, based on the coherence of the coarse-grained types available in DBpedia [1] for their descendant Wikipedia articles. The method from [50] also identifies classes among Wikipedia categories rather than articles. Since many Wikipedia articles do not have a corresponding, similarly-titled Wikipedia parent category, evidence towards which categories are classes has limited utility towards distinguishing which articles are classes. Existing work where Wikipedia serves as the reference resource in some task tends to rely on articles rather than categories [2, 5, 14, 29, 35], with few exceptions [22].

Most methods [26, 50] that distinguish classes in Wikipedia require access to a part of speech tagger, a syntactic parser or a named entity recognizer and apply to English data only. In contrast, and similarly to the rule-based method introduced recently in [27], our method does not need access to linguistic processing tools and applies to multiple languages. Its results are superior to [27] and transitively to other baselines evaluated in [27].

Previous work in open-domain information extraction [9–11, 23, 33, 40, 48] often uses Wikipedia data [15, 27, 42, 44, 45]. Wikipedia plays a role in information retrieval [4, 7, 8, 18, 22, 37, 38, 41, 49], knowledge acquisition [19, 25, 45, 47] and the construction of structured knowledge repositories [1, 17, 43, 46].

7 CONCLUSION

Playing an increasing role in enhancing Web search results, the largest knowledge repositories available rely on data available in and associated with Wikipedia articles. The method proposed in this paper associates Wikipedia articles with a type of information that is missing from existing knowledge repositories, namely distinguishing articles that are classes. Current work investigates the role of ngrams and syntactic dependencies as low-level features collected from article text in Wikipedia; and the role of evidence not just around occurrences of the article title (*"Shield volcano"*) within the article, but also around disambiguated occurrences within other Wikipedia articles (*"[..] Paka is a shield volcano located in [..]"*) and, more generally, within other Web documents.

REFERENCES

[1] 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 (2009), 154–165.

[2] R. Blanco, G. Ottaviano, and E. Meij. 2015. Fast and Space-Efficient Entity Linking in Queries. In *Proceedings of the 8th ACM Conference on Web Search and Data Mining (WSDM-15)*. Shanghai, China, 179–188.

[3] 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)*. Vancouver, Canada, 1247–1250.

[4] 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)*. Vancouver, Canada, 1870–1879.

[5] A. Chisholm and B. Hachey. 2015. Entity disambiguation with Web links. *Transactions of the Association for Computational Linguistics* 3 (2015), 145–156.

[6] P. Downing. 1977. On the Creation and Use of English Compound Nouns. *Language* 53 (1977), 810–842.

[7] X. Du and C. Cardie. 2018. Harvesting Paragraph-level Question-Answer Pairs from Wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL-18)*. Melbourne, Australia, 1907–1917.

[8] 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)*. Cambridge, United Kingdom, 181–190.

[9] P. Ernst, A. Siu, and G. Weikum. 2018. HighLife: Higher-Arity Fact Harvesting. In *Proceedings of the 2018 Web Conference (WWW-18)*. Lyon, France, 1013–1022.

[10] O. Etzioni, A. Fader, J. Christensen, S. Soderland, and Mausam. 2011. Open Information Extraction: The Second Generation. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI-11)*. Barcelona, Spain, 3–10.

[11] A. Fader, S. Soderland, and O. Etzioni. 2011. Identifying Relations for Open Information Extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP-11)*. Edinburgh, Scotland, 1535–1545.

[12] C. Fellbaum (Ed.). 1998. *WordNet: An Electronic Lexical Database and Some of its Applications*. MIT Press.

[13] 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)*. Baltimore, Maryland, 945–955.

[14] O. Ganea, M. Ganea, A. Lucchi, C. Eickhoff, and T. Hofmann. 2016. Probabilistic Bag-Of-Hyperlinks Model for Entity Linking. In *Proceedings of the 25th World Wide Web Conference (WWW-16)*. Montreal, Canada, 927–938.

[15] O. Ganea and T. Hofmann. 2017. Deep Joint Entity Disambiguation with Local Neural Attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP-17)*. Copenhagen, Denmark, 2619–2629.

[16] 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)*. New Orleans, Louisiana, 4824–4831.

[17] 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 (2013), 28–61.

[18] J. Hu, G. Wang, F. Lochovsky, J. Sun, and Z. Chen. 2009. Understanding User's Query Intent with Wikipedia. In *Proceedings of the 18th World Wide Web Conference (WWW-09)*. Madrid, Spain, 471–480.

[19] 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)*. Perth, Australia, 1007–1014.

[20] J. Langford, A. Strehl, and L. Li. 2007. Vowpal Wabbit. http://hunch.net/ vw.

[21] D. Lenat. 1995. CYC: a Large-Scale Investment in Knowledge Infrastructure. *Commun. ACM* 38, 11 (1995), 32–38.

[22] 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)*. Lyon, France, 1623–1632.

[23] Mausam, M. Schmitz, S. Soderland, R. Bart, and O. Etzioni. 2012. Open Language Learning for Information Extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-12)*. Jeju Island, Korea, 523–534.

[24] R. Mihalcea. 2007. Using Wikipedia for Automatic Word Sense Disambiguation. In *Proceedings of the 2007 Conference of the North American Association for Computational Linguistics (NAACL-HLT-07)*. Rochester, New York, 196–203.

[25] V. Nastase and M. Strube. 2008. Decoding Wikipedia Categories for Knowledge Acquisition. In *Proceedings of the 23rd National Conference on Artificial Intelligence (AAAI-08)*. Chicago, Illinois, 1219–1224.

[26] V. Nastase and M. Strube. 2013. Transforming Wikipedia into a Large Scale Multilingual Concept Network. *Artificial Intelligence* 194 (2013), 62–85.

[27] M. Paşca. 2018. Finding Needles in an Encyclopedic Haystack: Detecting Classes Among Wikipedia Articles. In *Proceedings of the 2018 Web Conference (WWW-18)*. Lyon, France, 1267–1276.

[28] 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)*. Buenos Aires, Argentina, 896–902.

[29] X. Pan, T. Cassidy, U. Hermjakob, H. Ji, and K. Knight. 2015. Unsupervised Entity Linking with Abstract Meaning Representation. In *Proceedings of the 2015 Conference of the North American Association for Computational Linguistics (NAACL-HLT-15)*. Denver, Colorado, 1130–1139.

[30] 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)*. Ann Arbor, Michigan, 665–674.

[31] S. Ponzetto and R. Navigli. 2009. Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-09)*. Pasadena, California, 2083–2088.

[32] 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)*. Vancouver, British Columbia, 1440–1447.

[33] 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)*. Lyon, France, 1257–1266.

[34] L. Ratinov and D. Roth. 2012. Learning-Based Multi-Sieve Co-Reference Resolution with Knowledge. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-12)*. Jeju Island, Korea, 1234–1244.

[35] 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)*. Portland, Oregon, 1375–1384.

[36] M. Remy. 2002. Wikipedia: The Free Encyclopedia. *Online Information Review* 26, 6 (2002), 434.

[37] Z. Bouraoui S. Jameel and S. Schockaert. 2017. MEmbER: Max-Margin Based Embeddings. In *Proceedings of the 40th International Conference on Research and Development in Information Retrieval (SIGIR-17)*. Tokyo, Japan, 783–792.

[38] U. Scaiella, P. Ferragina, A. Marino, and M. Ciaramita. 2012. Topical Clustering of Search Results. In *Proceedings of the 5th ACM Conference on Web Search and Data Mining (WSDM-12)*. Seattle, Washington, 223–232.

[39] A. Singhal. 2012. Introducing the Knowledge Graph: Things, not Strings. Corporate blog.

[40] 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)*. Marina del Rey, California, 556–564.

[41] C. Tan, F. Wei, P. Ren, W. Lv, and M. Zhou. 2017. Entity Linking for Queries by Searching Wikipedia Sentences. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP-17)*. Copenhagen, Denmark, 68–77.

[42] 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)*. Cambridge, United Kingdom, 345–354.

[43] D. Vrandečić and M. Krötzsch. 2014. Wikidata: A Free Collaborative Knowledge Base. *Commun. ACM* 57 (2014), 78–85.

[44] Z. Wang, Z. Li, J. Li, J. Tang, and J. Pan. 2013. Transfer Learning Based Cross-lingual Knowledge Extraction for Wikipedia. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-13)*. Sofia, Bulgaria, 641–650.

[45] F. Wu and D. Weld. 2010. Open Information Extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*. Uppsala, Sweden, 118–127.

[46] 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)*. Scottsdale, Arizona, 481–492.

[47] Y. Yan, N. Okazaki, Y. Matsuo, Z. Yang, and M. Ishizuka. 2009. Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP-09)*. Singapore, 1021–1029.

[48] X. Yao and B. Van Durme. 2014. Information Extraction over Structured Data: Question Answering with Freebase. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL-14)*. Baltimore, Maryland, 956–966.

[49] S. Zhang and K. Balog. 2018. Ad Hoc Table Retrieval Using Semantic Similarity. In *Proceedings of the 2018 Web Conference (WWW-18)*. Lyon, France, 1553–1562.

[50] C. Zirn, V. Nastase, and M. Strube. 2008. Distinguishing Between Instances and Classes in the Wikipedia Taxonomy. In *Proceedings of the 5th European Semantic Web Conference (ESWC-08)*. Tenerife, Spain, 376–387.