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

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.

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.

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