Lightweight Lexical and Semantic Evidence for Detecting Classes Among Wikipedia Articles

Marius Paşca, Travis Wolfe, Google, 2019

Keyu Han (keyuhan2@illinois.edu)

ECE 594, Feb. 03, 2022

Overview

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.

7 Parts

Introduction——Detection of Classes——Method——Experimental Settings——Evaluation Results——Related Work——Conclusion

1. Introduction

Basic Info

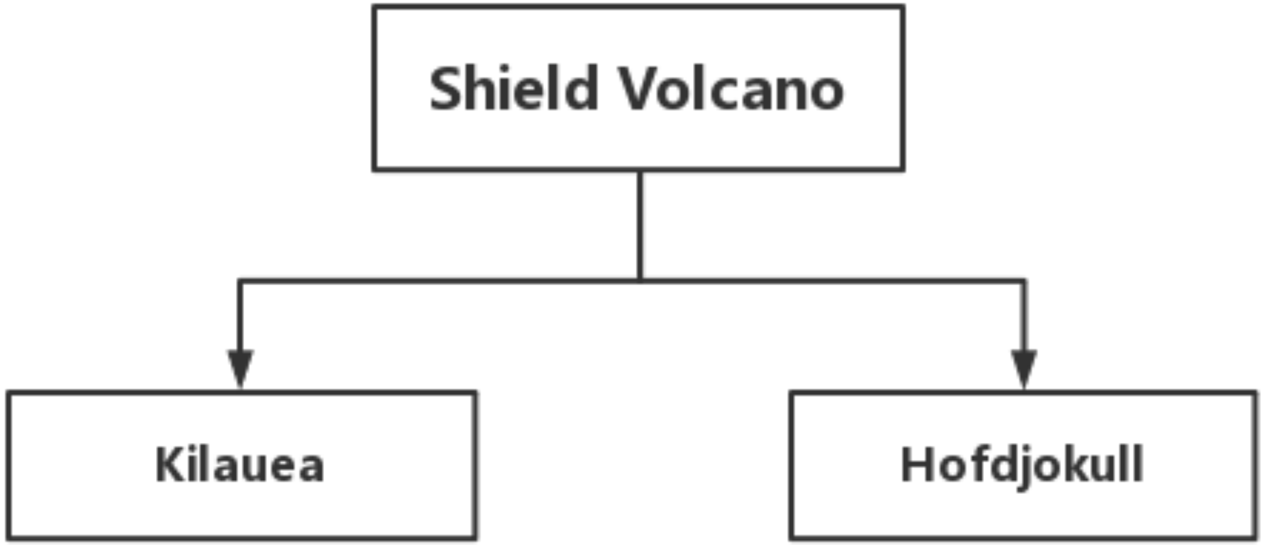
- **dataset**: Wikipedia articles
- **dataset features**: features are lexical or semantic in nature
- **aim**: distinguish Wikipedia articles that are classesfrom other articles

  (Shield volcano —— Kilauea)
- **test/evaluation methods**: using multiple languages over multiple evaluation sets
- **result**: demonstrate the superiority of the proposed method over previous work

2. Detection of Class

Classes

- **Definition**: placeholders for sets of instances that share common properties.
- e.g. Shield volcano” is a placeholder for a set of instances such as “Kilauea” and “Hofsjökull”



2. Detection of Class

Problem

- 97 out of 100 Wikipedia articles may be instances.
- Wikipedia does not distinguish articles that are classes from those that are not.
- 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”).
- large knowledge graphs also rely heavily on creating and maintaining internal concepts for most if not all Wikipedia articles.

Goals

the selection of as many Wikipedia articles that are classes as accurately as possible, out of all Wikipedia articles.

2. Detection of Class

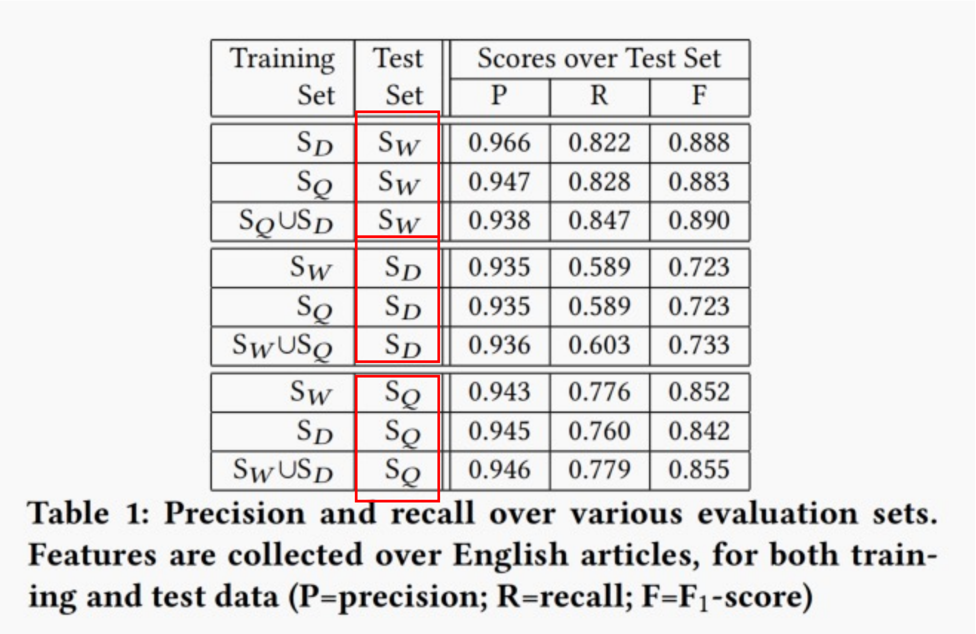
Applications

- **large knowledge repositories** - Enriching Knowledge Repositories.
- **Expansion of Lexical Dictionaries**: Wikipedia articles extracted as classes represent an inexpensive source of high-quality candidate concepts - the high cost of manual maintenance and expansion, valid open-domain concepts may be missing from expert-created lexical resources like WordNet(WordNet labels the semantic relations among words, whereas the groupings of words have the meaning similarity.)
- **Topic Decomposition** - Existing methods for decomposing Wikipedia articles lack “additional signals to better distinguishing between fully compositional and noncompositional” articles (cf. [28]).
- **Wikipedia Hierarchies** - Wikipedia articles are written by teams of independent volunteers in the absence of formal hierarchical organizational structures

3. Envaluation Results

3.1 Results with Lexical Features

Extraction over English Articles



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)

combining both of the other evaluation sets into a single training set, this study brings only a small improvement in F1-scores, relative to using only one of other evaluation sets.

3. Envaluation Results

3.1 Results with Lexical Features

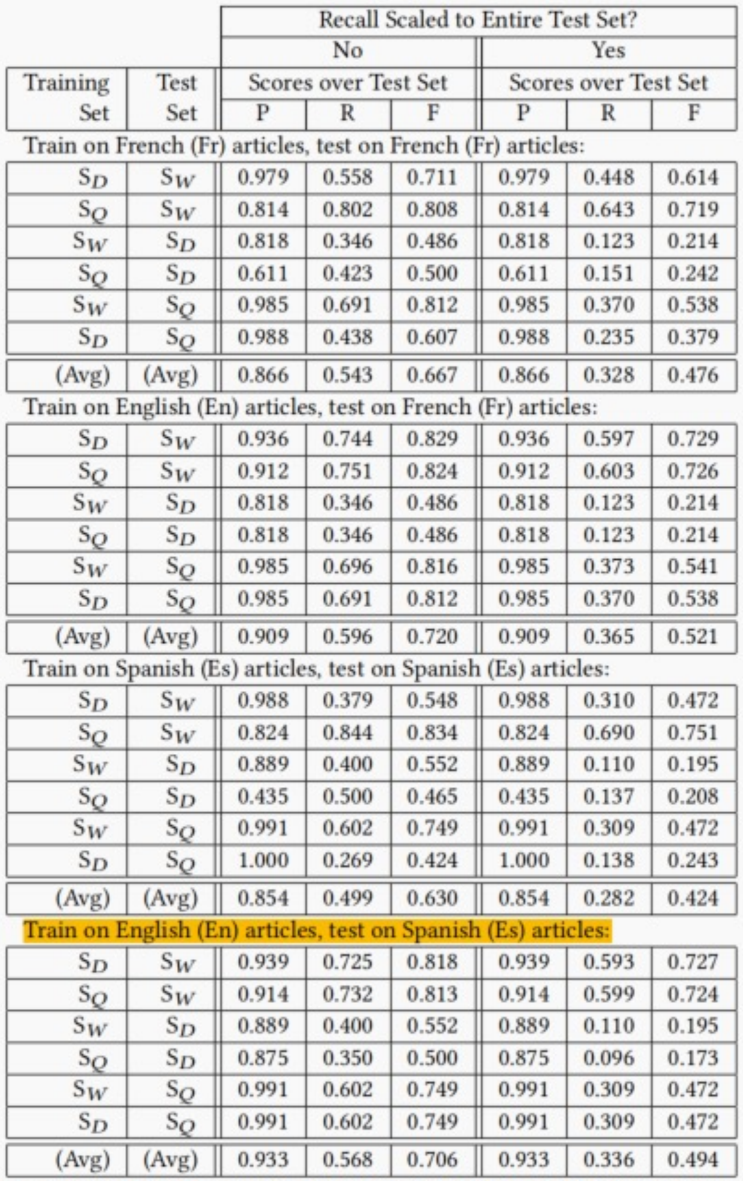
Extraction over Articles in Other Languages

- **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).
- **Training & test dataset**: English/ French/Spanish
- **Conclusion**

1. a large fraction of the evaluation sets used as training sets is lost, when training on SD ; but when training on SW, little is lost

2. SD’s recall can be reduced the most precisely when testing on SW or SQ in a cross - language training.

3. changes recall from 0.744 to 0.558 when training SD and testing on SW. In the bottom table.



		Recall Scaled to Entire Test Set?					
		No			Yes		
Training Set	Test Set	Scores over Test Set			Scores over Test 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)

3. Envaluation Results

3.2 Results with Semantic Features

Impact of Features from Wikidata

- **element**: with semantic features limited to only hypernyms(**Fshp**) or only properties (**Fspr**)
- **result**

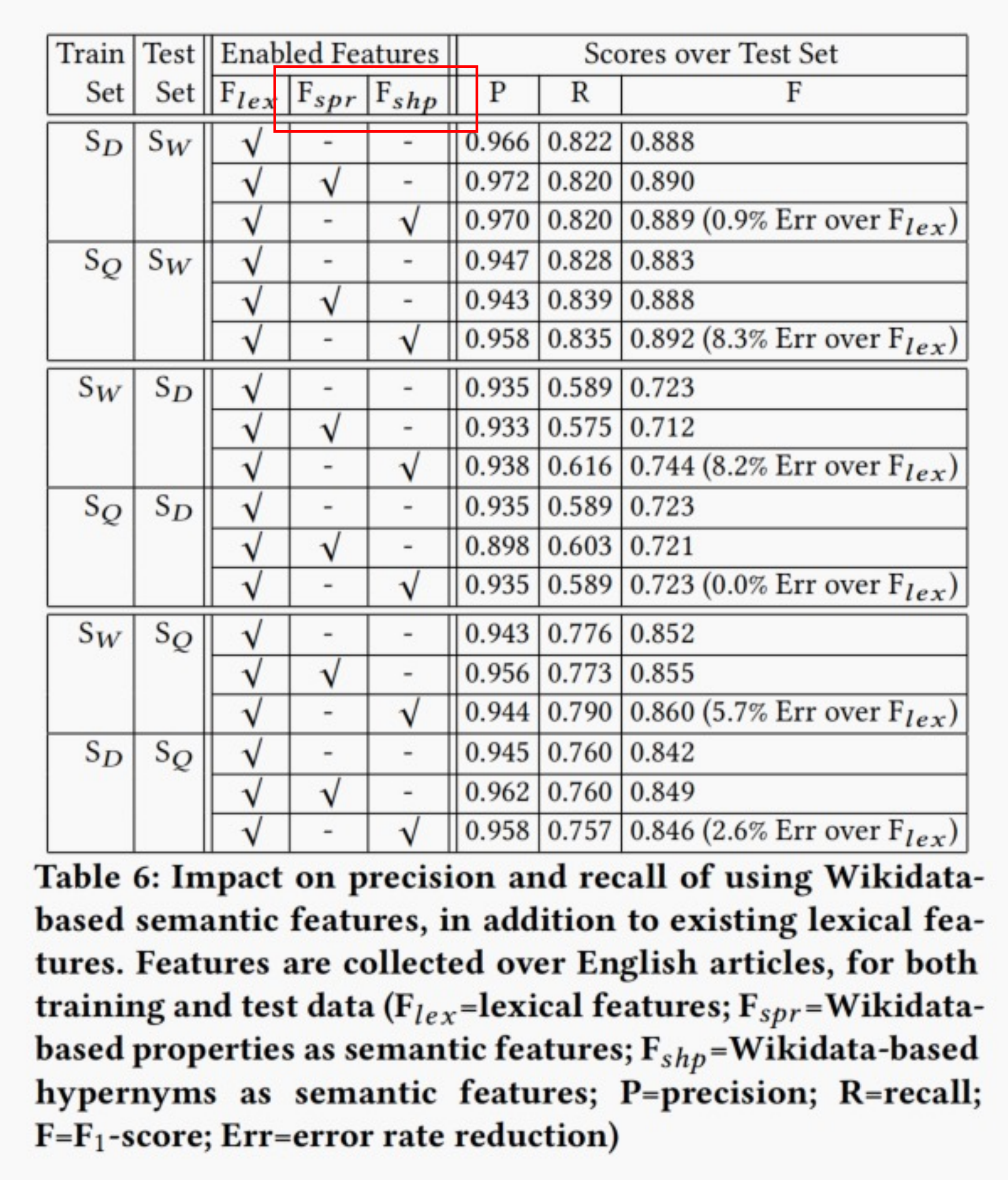
1. adding semantic features from Wikidata properties (Fspr) causes inconsistent changes to scores.

2. adding semantic features from hypernyms (Fshp) gives improved F1-scores, with small reduction in error rates.

- **Conclusion**

1. Semantic features may still be useful if more training data became available.

2. the results given by lexical features alone is encouraging.



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)

4. Method

4.1 Lexical Features Within Wikipedia

Clues in Wikipedia Articles

- 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
- **Lexical Clue 1**: Pre-defifined contextual patterns
- **Lexical Clue 2**: Morphological variation
- **Lexical Clue 3**: Capitalization

4. Method

4.1 Lexical Features Within Wikipedia

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

- **8 clues:contextual pattern match, identity, plural, mixedcase, lowercase, mixedcase plural, lowercase plural, plural category**

- a) C1(**contextual pattern match**) is the count of case--insensitive occurrences that match a pre-defifined contextual pattern; b) C2(**identity**), C3(**plural**) are the counts of case-insensitive occurrences in identical vs. plural form; c) C4(**mixedcase**), C5(**lowercase**) are the counts of case sensitive occurrences in mixed case vs. lowercase; d) C6(**mixedcase plural**), C7(**lowercase plural**) are the counts of case-sensitive occurrences of plural forms in mixed case vs. lowercase; and e) C8(**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

4. Method

4.2 Semantic Features Outside Wikipedia

Lexical vs. Semantic Features

- Intuitions presented and features collected so far are lexical. They apply horizontally, across all Wikipedia articles
- 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
- Semantic Clue 1: **Hypernyms**: classes and instances may easily share hypernyms, such as “Kilauea” and “Shield Volcano” sharing the hypernym “Volcano”
- Semantic Clue 2: **Properties**: 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” .

4. Method

4.2 Semantic Features Outside Wikipedia

Semantic Features from Wikidata

- The set of properties of a Wikidata topic is the set of predicates of relations
- The properties and InstanceOf hypernyms of a given Wikidata topic 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

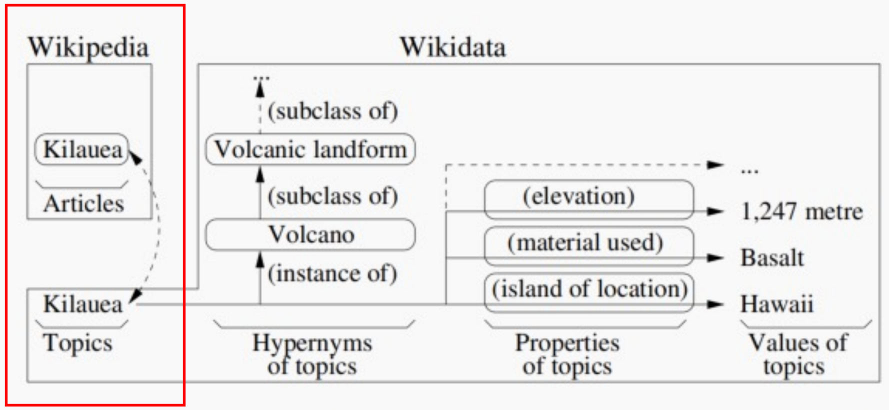


Figure 1: Examples of hypernym topics and properties available in Wikidata for Wikipedia articles

5. Experimental Setting

- **Supervise 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. Other loss functions or non-linear algorithms might be used.
- **Data Sources**: Semantic features are extracted for each Wikipedia article from this snapshot, based on data from a snapshot of Wikidata from June 2018
- **Evaluation Sets**: Three evaluation sets introduced in other paper serve as the source data for training and testing the proposed method. The first evaluation set, SW , is derived from Instance relations available in WordNet. The second and third evaluation sets are random samples of Wikipedia articles annotated manually.
- **Training an Test Sets**: The evaluation sets are employed as training data or test data, in various possible combinations. one possible combination is to employ SW as training data and SQ as test data
- **Extraction Parameters**: to normalize the article title by removing portions within parentheses (eg.“Circuit (administrative division)” ->“Circuit”)

6. Conclusion

- Current work investigates the role of n-grams and syntactic dependencies as low-level features collected from article text in Wikipedia.
- low-level features: of n-grams and syntactic dependencies - article text in Wiki
- the role of evidence are:
  - within **the article**: around occurrences of the article title (“Shield volcano”) ,
  - within **other Wikipedia articles** : disambiguated occurrences (“[..] Paka is a shield volcano located in [..]”)
  - within **other Web documents**