Low-Complexity Heuristics for Deriving Fine-Grained Classes of Named Entities from Web Textual Data

Marius Paşca

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

Abstract

We introduce a low-complexity method for acquiring fine-grained classes of named entities from the Web. The method exploits the large amounts of textual data available on the Web, while avoiding the use of any expensive text processing techniques or tools. The quality of the extracted classes is encouraging with respect to both the precision of the sets of named entities acquired within various classes, and the labels assigned to the sets of named entities.

1. Introduction

Class instances of various types constitute a large fraction of the search queries submitted most frequently by Web users. Class instances also occur often in Web documents, confirming the special role that they play in natural language, as they are used to refer to objects and concepts of common interest. Although work on named entity recognition traditionally focuses on the acquisition and identification of instances within a small set of coarse-grained classes, the distribution of instances within query logs indicate that Web search users are interested in a finer-grained set of classes. Depending on prior knowledge, personal interests and immediate needs, users may submit queries in the medical domain, inquiring about the symptoms of *leptospirosis* or the treatment of *monkeypox*, both of which are instances of *zoonotic diseases*, or the risks and benefits of *surgical procedures* such as *prk* and *angioplasty*. Other users may be more interested in geography, through queries referring to *uganda* and *angola*, which are *african countries*, or *active volcanoes* like *etna* and *kilauea*. The wide variation of the domains of interest to Web users illustrates the potential impact that the availability of a large set of fine-grained classes of instances may have in Web search. A variety of text processing tasks, including coreference resolution (McCarthy and Lehnert, 1995), named entity recognition (Stevenson and Gaizauskas, 2000) and seed-based information extraction (Riloff and Jones, 1999), can also directly take advantage of the extracted sets of classes of instances.

Starting from a few Is-A extraction patterns widely used in information extraction literature (Hearst, 1992), this paper introduces a few precision-enhancing heuristics that take advantage of textual data available on the Web, by mining a collection of Web search queries and a collection of Web documents to acquire a large number of open-domain classes in the form of instance sets (e.g., {*leptospirosis*, *brucellosis*, *lyme disease*, *monkeypox*, *psittacosis*,...}) associated with class labels (e.g., *zoonotic diseases*). By exploiting the contents of query logs during the extraction of labeled classes of instances from Web documents, we acquire thousands of classes covering a wide range of topics and domains. The extraction of classes requires a small amount of supervision, in the form of a few Is-A extraction patterns.

2. Extraction of Fine-Grained Classes

2.1. Document Pre-Processing

The contents of the Web documents, from which the labeled classes of instances are extracted, is converted to text by filtering out Html tags. The documents are split into sentences, tokenized and part-of-speech tagged using the TnT tagger (Brants, 2000).

2.2. Pattern-Based Extraction

In order to acquire pairs of an instance and an associated class label from text, we apply a small set of manually-created extraction patterns. The patterns were introduced in (Hearst, 1992) and successfully used in a large body of previous work on extracting Is-A pairs from text. For simplicity and robustness when applied to large amounts of Web text, the number of extraction patterns is limited to a very small set, namely $\langle\ \mathcal{C}$ [such as] $\mathcal{I}\ \rangle$ and $\langle\ \mathcal{C}$ [including] $\mathcal{I}\ \rangle$. As such, the patterns represent a low-complexity solution to the problem of extracting candidate pairs of an instance $\mathcal{I}$ (e.g., *brucellosis*) and an associated class label $\mathcal{C}$ (e.g., *zoonotic diseases*) from noisy text.

For each match within a sentence, the patterns determine the right boundary of the class label, and the left boundary of the class instance respectively. The left boundary of the class label is identified through shallow analysis of the part-of-speech tags of the sentence words situated immediately to the left of the pattern match. If the sequence of tags corresponds to a base (i.e., non-recursive) noun phrase whose last component is a plural-form noun, then the beginning of the noun phrase is the left boundary of the class label. Otherwise, the pattern match is discarded. In comparison, the part-of-speech tags of the sentence words cannot be used reliably to identify the right boundary of the class instance. Indeed, instances of arbitrary classes exhibit significant variation in form, from simpler-to-identify sequences of capitalized, proper nouns (e.g., *Mauna Loa* and

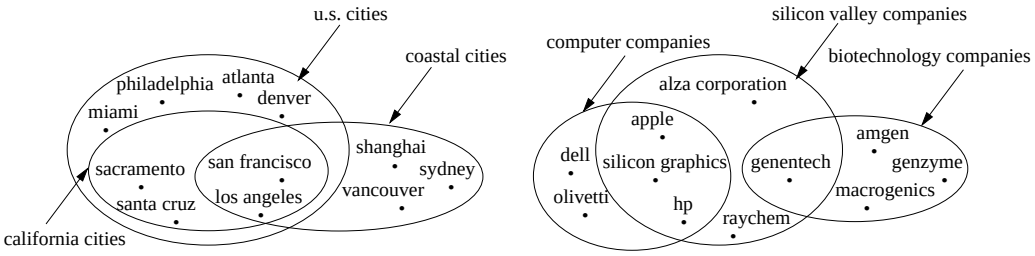

Figure 1: Examples of sets of instances associated with various class labels

British Airways) to sequences that are more difficult to locate, such as movie titles, expressions and sayings, species, chemical substances etc. Therefore, the right boundary of the class instances is determined simply by the earliest occurrence of a delimiter (that is, a comma or a full stop) after the pattern match within the sentence.

2.3. Precision-Oriented Filters

The pairs of a class instance and an associated class label extracted from text via pattern matching are further refined through three precision-oriented heuristics. The first heuristic aims at discarding pairs containing spurious class instances that were extracted due to an undesirable pattern match:

Heuristic 1: Discard pairs of a class label and a class instance, if the class instance is not frequently submitted as a full-length Web search query.

The rationale behind the first heuristic is that, sooner or later, users interested in a particular class instance will inquire about that instance. The inquiries will take many forms, including submissions to a Web search engine of full-length queries containing only the class instance. Whenever a class instance does not occur as an entire, case-insensitive query in query logs, the class instance and its associated class labels are discarded from the pairs extracted via pattern matching. To further trade off recall for higher precision, a second heuristic is applied:

Heuristic 2: Discard pairs of a class label and a class instance, if the head of the class label is not the one that is the most frequently associated with the class instance in the extracted pairs.

The second heuristic analyzes the head nouns of all class labels $\mathcal{C}$ collected via pattern matching for a given instance $\mathcal{I}$. The heuristic identifies which head noun occurs most frequently across the potential class labels of the instance, then discards the labels whose head nouns are not the most frequent head noun. For example, the most frequent head of the labels associated with *brucellosis* is *diseases*. Therefore, class labels such as *zoonotic diseases* and *communicable diseases* are retained, whereas *dangerous bacteria* and *tests* are discarded, thus promoting precision of the class labels at the expense of lower recall.

After filtering, the resulting pairs are arranged into sets of class instances, as shown in Figure 1. After discarding classes with fewer than 25 instances, the top 100 instances are retained for each class.

3. Evaluation

3.1. Experimental Setting

The acquisition of labeled classes of instances relies on unstructured text available within a combination of Web documents maintained by, and search queries submitted to the Google search engine.

The collection of Web search queries is a random sample of fully-anonymized queries in English submitted by Web users in 2006. The sample contains approximately 50 million unique queries. Each query is accompanied by its frequency of occurrence in the logs.

The document collection consists of approximately 100 million Web documents in English, as available in a Web repository snapshot from 2006.

3.2. Quantitative Results

The extracted data consists of pairs of an instance and a class label, such that each class label is associated with 25 to 100 instances. Table 1 illustrates the extracted classes ranked according to their popularity within query logs, measured by the frequencies of the class labels as full queries (e.g., *games* or *arcade games*) within query logs.

Thanks to the open-domain nature of the precision-oriented filters, the extracted classes are not restricted to any single domain of interest. Instead, the classes cover a wide range of topics and domains, including medicine (e.g., *genetic disorders* at rank 175 and *personality disorders* at rank 255 in Table 1), finance (e.g., *mutual funds* at rank 191), geology (*sedimentary rocks* at rank 245) and entertainment (e.g., *games* at rank 1).

3.3. Qualitative Results

The coverage of the extracted instances is measured against one of the popular lexical resources in natural language applications, namely the WordNet lexical database (Fellbaum, 1998). WordNet encodes English concepts in the form of sets of synonyms, or synsets (e.g., {*port of entry, point of entry*}), associated with a common definition (e.g., *"a port in the United States where customs officials are stationed to oversee the entry and exit of people and merchandise"*). WordNet synsets are organized hierarchically, such that more specific concepts, or hyponyms, are located under more general concepts, or hypernyms. Recent versions of WordNet also provide explicit Has-Instance relations, which correspond to Instance-Of relations between a class (e.g., *painter*) and one of its instances (e.g., *Amedeo Modigliano*).

Rank	Class Label	Rank	Class Label	Rank	Class Label	Rank	Class Label
1	games	75	plants	151	gadgets	225	franchises
5	poems	81	bacteria	155	enzymes	231	civil rights
11	cars	85	spiders	161	magazines	235	exotic cars
15	airlines	91	whole foods	165	universities	241	explorers
21	holidays	95	java games	171	batteries	245	sedimentary rocks
25	horses	101	castles	175	genetic disorders	251	religions
31	arcade games	105	spells	181	tests	255	personality disorders
35	cartoons	111	addresses	185	mammals	261	coins
41	books	115	classic cars	191	mutual funds	265	weapons
45	fun games	121	video games	195	shapes	271	vegetables
51	flags	125	party games	201	satellites	275	airports
55	careers	131	famous people	205	symptoms	281	kitchen appliances
61	cards	135	kids	211	surnames	285	war games
65	fairy tales	141	candles	215	codes	291	architects
71	watches	145	robots	221	paintings	295	controversial topics

Table 1: Popularity of the extracted labeled classes, measured by the frequency of occurrence of the class labels as full, case-insensitive queries in query logs

Hypernym		Instances		Cvg
Synset	Definition	Examples	Count	
Australian state	one of the several states constituting Australia	New South Wales, Queensland, South Australia, Tasmania	6	1.00
existentialist	a philosopher who emphasizes freedom of choice and personal responsibility [..]	Albert Camus, Beauvoir, Camus, Heidegger, Jean-Paul Sartre	8	1.00
government building	a building that houses a branch of government	Capitol, Capitol Building, Pentagon, White House	4	1.00
search engine	a computer program that retrieves documents or files or data from a database [..]	Ask Jeeves, Google, Yahoo	3	1.00
port of entry	a port in the United States where customs officials are stationed to oversee [..]	Aberdeen, Bellingham, Brownsville, Greater New York	25	0.88
possession	a territory that is controlled by a ruling state	American Virgin Islands, Faroes, Faeroe Islands, Faeroes, Macau	13	0.77
university	establishment where a seat of higher learning is housed [..]	Brown, Brown University, Carnegie Mellon University	44	0.75
couturier	someone who designs clothing	Balenciaga, Calvin Klein, Calvin Richard Klein, Dior	13	0.69
fictional animal	animals that exist only in fiction (usually in children's stories)	Donald Duck, Easter bunny, Mickey Mouse, Mighty Mouse	6	0.67
memorial	a structure erected to commemorate persons or events	Great Pyramid, Lincoln Memorial, Pyramids of Egypt	6	0.67
painter	an artist who paints	Amedeo Modigliano, Andy Warhol, Anna Mary Robertson Moses	218	0.66
continent	one of the large landmasses of the earth	Africa, Antarctic continent, Europe, Eurasia, Gondwanaland, Laurasia	13	0.62
educator	someone who educates young people	Abbott Lawrence Lowell, Bethune, Booker T. Washington, Carl Orff	54	0.48
eon	the longest division of geological time	Archaeozoic, Archaeozoic aeon, Archean aeon, Archean eon	24	0.08
anarchist	an advocate of anarchism	Bakunin, Bartolomeo Vanzetti, Prince Peter Kropotkin	14	0.00
microscopist	a scientist who specializes in research with the use of microscopes	Anton van Leeuwenhoek, Anton van Leuwenhoek, Swammerdam	6	0.00
national anthem	a song formally adopted as the anthem [..]	The Star-Spangled Banner	2	0.00
rebellion	organized opposition to authority; a conflict in which one faction tries to wrest control [..]	Great Revolt, Indian Mutiny, Peasant's Revolt, Sepoy Mutiny	4	0.00
soil horizon	a layer in a soil profile	A horizon, A-horizon, B horizon, B-horizon, C horizon, C-horizon	6	0.00
Average (over 945 hypernyms)		-	18.71	0.39

Table 2: Coverage of extracted instances, measured by the percentage of instances encoded under various WordNet hypernyms via Has-Instance relations that occur among the extracted instances (Cvg=coverage)

Class Label	Examples of Instances		Prec
	Judged to be Correct	Judged to be Incorrect	
Domain: Medicine			
zoonotic diseases	rabies, brucellosis, leptospirosis, salmonellosis, plague, west nile virus, lyme disease, monkeypox, psittacosis	tuberculosis, foodborne, rinderpest, bacterial, hog cholera, enteric	0.84
surgical procedures	prk, lasik, rk, liposuction, joint replacement, chemical peels, angioplasty, hysterectomy, gastric bypass surgery	orthopedic, laparoscopic, botox injections, endoscopic, arthroscopic	0.87
Domain: Geography			
african countries	nigeria, kenya, south africa, zimbabwe, uganda, ghana, tanzania, zambia, botswana, angola	afghanistan, iraq, india, indonesia, countries, pakistan, saudi arabia	0.66
active volcanoes	etna, mauna loa, kilauea, stromboli, ruapehu, aso, mount vesuvius, shishaldin, mount apo, popocatepetl	vulcano	0.96
Domain: Entertainment			
movies	star wars, matrix, titanic, finding nemo, armageddon, independence day, scream, american pie, austin powers	silent stars, family fare	0.97
british actors	kenneth branagh, maggie smith, tim curry, rupert everett, ewan mcgregor, helen mirren, colin firth	janet suzman, cyril cusack	0.97
Domain: Travel			
airlines	british airways, air france, delta, lufthansa, qantas, continental, singapore airlines, cathay pacific, klm	american airline, cathy pacific, cathay	0.97
car rental companies	hertz, alamo, avis, thrifty, europcar, auto europe, budget car rental, holiday autos, national car rental	sports car hire, enterprize, us, hertz car hire, american express	0.84
Average (over 8 classes)	-	-	0.88

Table 3: Precision of instances associated with a sample of extracted classes (Prec=precision)

The instances available within WordNet via Has-Instance relations constitute a benchmark against which the coverage of the instances extracted from text can be automatically computed. To this effect, each component phrase of a synset encoded via a Has-Instance relation under a hypernym synset in WordNet is collected as a benchmark instance of that synset. For instance, the synset corresponding to *soil horizon*, defined as *"a layer in soil profile"*, has three Has-Instance synsets in WordNet, each of which contains two synonyms: {*A-horizon*, *A horizon*}, {*B-horizon*, *B horizon*} and {*C-horizon*, *C horizon*}. Therefore the synset *soil horizon* has 6 instances in the benchmark. As shown in the first four columns of Table 2, the resulting benchmark consists of a total of 945 hypernym synsets, with an average of 18 instances per synset in WordNet.

In order to receive full credit for a WordNet synset in terms of coverage, all its WordNet instances of the synset must occur among the extracted instances as full-length, case-insensitive matches. Any variations due to alternative spelling (e.g., *Faeroes* vs. *faroes*) or level of specificity (e.g., *Abbott Lawrence Lowell* vs. *abbott lowell*) result in failed comparisons, and therefore a lack of any credit towards the computed coverage scores. The last column in Table 2 shows that the coverage varies significantly across the WordNet synsets. For some of the synsets (e.g., *existentialist* and *search engine*), all WordNet instances occur among the instances extracted from Web text. At the bottom end of the coverage score spectrum, none of the instances available in WordNet for *anarchist* and *soil horizon* are found among the extracted instances.

Table 3 summarizes the precision of the extracted data, which is computed by manually inspecting the instances extracted for a sample of eight labeled classes. The lowest precision score, 0.66, is obtained for the class label *african countries*. Most of the errors within this class are due to the incorrect extraction of non-African countries as part of the same instance set. The precision-oriented heuristics contribute to a precision score of 0.88, as an average over the instance sets associated to the eight sample classes.

4. Conclusion

This paper introduces a few simple, lightweight precision-oriented heuristics for compiling sets of class instances from unstructured text, as an alternative to iterative extraction starting from a few seeds (Riloff and Jones, 1999). When applied to a large repository of Web documents, the heuristics contribute to the acquisition of a large number of accurate sets of labeled classes.

5. References

T. Brants. 2000. TnT - a statistical part of speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP-00)*, pages 224–231, Seattle, Washington.

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

M. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 539–545, Nantes, France.

K. McCarthy and W. Lehnert. 1995. Using decision trees for coreference resolution. In *Proceedings of the 14th International Joint Conference on Artifi cial Intelligence (IJCAI-95)*, pages 1050–1055, Montreal, Quebec.

E. Riloff and R. Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 16th National Conference on Artifi cial Intelligence (AAAI-99)*, pages 474–479, Orlando, Florida.

M. Stevenson and R. Gaizauskas. 2000. Using corpus-derived name lists for named entity recognition. In *Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP-00)*, Seattle, Washington.