Web-Derived Resources for Web Information Retrieval: From Conceptual Hierarchies to Attribute Hierarchies

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

Enrique Alfonseca
Google Inc.
110 Brandschenkestrasse
Zurich, Switzerland 8002
ealfonseca@google.com

ABSTRACT

A weakly-supervised extraction method identifies concepts within conceptual hierarchies, at the appropriate level of specificity (e.g., *Bank* vs. *Institution*), to which attributes (e.g., *routing number*) extracted from unstructured text best apply. The extraction exploits labeled classes of instances acquired from a combination of Web documents and query logs, and inserted into existing conceptual hierarchies. The correct concept is identified within the top three positions on average over gold-standard attributes, which corresponds to higher accuracy than in alternative experiments.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing Abstracting methods; I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms

Algorithms, Experimentation

Keywords

Knowledge acquisition, class attributes, named entities, conceptual hierarchies, Web search, unstructured text

1. INTRODUCTION

Motivation: Current methods for large-scale class attribute extraction produce ranked lists of attributes at encouraging accuracy levels for a variety of input classes. For example, the top attributes extracted from query logs in [8] for the classes *Actor* and *Painter* are:

• *Actor*: [awards, height, age, date of birth, weight, birthdate, birthplace, cause of death, real name];

• *Painter*: [paintings, biography, bibliography, autobiography, artwork, self portraits, quotations, bio, quotes, life history].

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA

While technically correct, many of the extracted attributes are in fact descriptive of a more general concept (in this case, *Person*), and thus fail to capture the properties that best differentiate the given classes (*Actor* and *Painter*) from other classes. Indeed, attributes such as *height*, *age* and *date of birth* are relevant for the class *Actor*, but they are also relevant for many other classes such as *Painter*, *Musician*, *Physicist* or *Professor*. Thus, even the most accurate of the current extraction methods fail to estimate how specific the extracted attributes are, relative to their respective classes.

Contributions: This paper integrates open-domain class attributes extracted automatically from a combination of Web documents and query logs, into existing conceptual hierarchies constructed manually by experts. For this purpose, a set of more than 9,000 open-domain classes containing a total of around 200,000 instances are acquired along with their ranked lists of attributes from unstructured text. The classes of instances are linked into conceptual hierarchies available in WordNet [3]. The analysis of the ranked lists of attributes extracted for individual classes, and the co-occurrence of attributes with class instances within query logs, allows for the computation of ranked lists of potential concepts to which the attributes are most likely to apply. Although the amount of supervision during extraction is strictly limited to a few widely-used extraction patterns, and as few as five seed attributes, the evaluation of the lists of potential concepts illustrates that the correct level of specificity can be identified accurately for a variety of open-domain class attributes. The resulting accuracy is significant both in absolute value (corresponding to the correct concept being returned within the top three positions on average over gold-standard attributes), and relative to experiments with manually-compiled or other automatically-acquired classes of instances (with accuracy improving by 80% and higher).

Applications: The identification of the appropriate level of specificity of an extracted attribute has immediate impact on the usefulness and coverage of the lists of attributes acquired from text for arbitrary classes. The usefulness of the attributes increases, as attributes descriptive of a more general concept (e.g., *date of birth* for *Person*) are identified and demoted within the lists of attributes produced for more specific classes (e.g., *Actor*). Conversely, the coverage of the attributes increases, as attributes descriptive of a more general concept (e.g., *Person*) are inherited and used to augment the lists of attributes produced for more specific classes (e.g., *Painter*, *Musician*, *Physicist* or *Professor*) from which the otherwise relevant attributes may be absent due to data sparseness. Outside of the task of attribute extraction, lexical and encyclopedical resources created by experts [3] or through volunteer contributions [3] tend to be organized

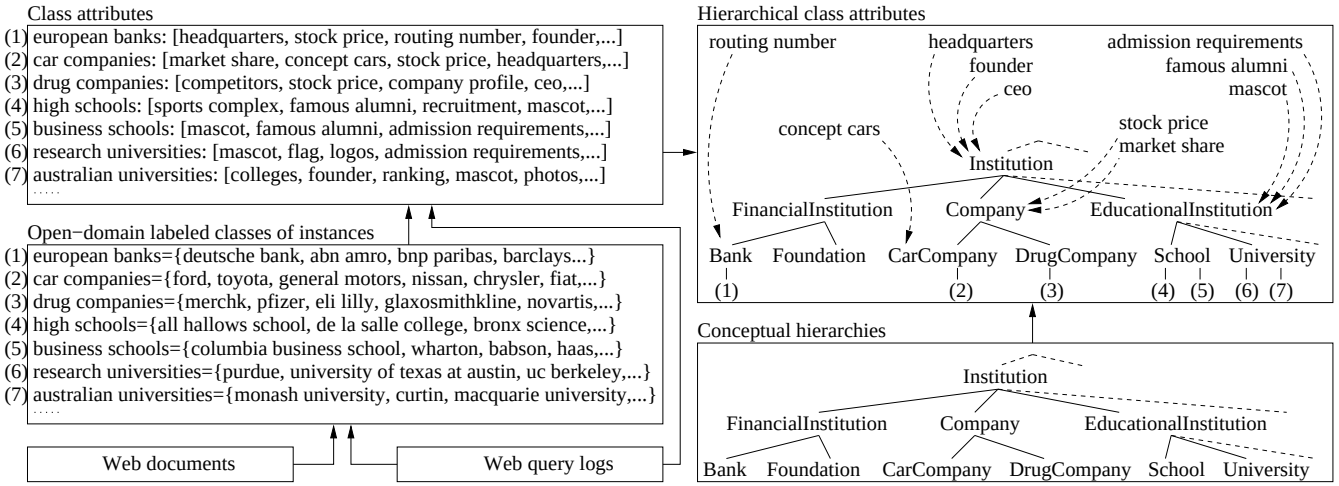

Figure 1: Overview of weakly supervised identification of concepts to which various attributes apply

into hierarchies of classes of instances, and would therefore benefit from the systematic identification of the most general concepts to which attributes best apply, as a necessary step towards the automatic inclusion of attributes to augment the existing hierarchical resources. In Web search, the results returned to a query that refers to a named entity (e.g., *Claude Monet*) can be augmented with a compilation of relevant facts, based on previously-identified attributes that are specific to the class to which the named entity belongs. Moreover, the original query can be refined into semantically-justified query suggestions, by concatenating it with one of the specific attributes for the corresponding class (e.g., *Claude Monet paintings* for *Claude Monet*).

2. EXTRACTION OF HIERARCHICAL ATTRIBUTES

2.1 Overview

Figure 1 shows how Web textual data is used to acquire open-domain class attributes over hierarchies, through the sequential extraction of: 1) open-domain, labeled classes of instances, by applying a few extraction patterns to unstructured text within documents, while guiding the extraction based on the contents of query logs (bottom-left in Figure 1); 2) class attributes that capture quantifiable properties of those classes, by mining query logs while guiding the extraction based on a few attributes provided as seed examples (top-left in the figure); and 3) hierarchical class attributes, by computing the concepts within existing conceptual hierarchies, to which extracted attributes best apply, after automatically linking labeled class instances under hierarchy concepts (top-right in the figure).

2.2 Extraction of Flat Classes and Attributes

Labeled Classes of Instances: Similarly to [9], the extraction of labeled classes of instances relies on hand-written patterns, widely used in literature on extracting conceptual hierarchies from text [5, 14]:

$\langle$[..] $\mathcal{C}$ [such as|including] $\mathcal{I}$ [and|,|.]$\rangle$,

where $\mathcal{I}$ is a potential instance (e.g., *BNP Paribas*) and $\mathcal{C}$ is a potential class label for the instance (e.g., *European banks*), for example in the sentence: *"Investors will also keep an eye on results from European banks such as BNP Paribas [..]"*.

In the patterns, the boundaries of potential class labels $\mathcal{C}$ are simply approximated from the part-of-speech tags of the sentence words, as a base (i.e., non-recursive) noun phrase identified as a sequence of adjectives or nouns ending in a plural-form noun. In the example sentence from above, the class label is *European banks*, which consists of a plural-form noun and a preceding modifier. If no such phrase is found, the pattern match is discarded. In comparison, to detect the boundaries of potential instances $\mathcal{I}$, we hypothesize that relevant instances of any kind must occur as search queries containing an instance and nothing else. In practice, the right boundaries of the instances $\mathcal{I}$ in the extraction patterns are identified by simply checking that the sequence of words within the pattern that corresponds to the potential instance $\mathcal{I}$ can be found as an entire query in query logs. During matching, all string comparisons are case-insensitive. If no such query is found, the pattern match is discarded [9].

Since most queries are typed in lower case by their users, the collected data is uniformly converted to lower case. The quality of the collected pairs of a class and an instance is further refined with an inexpensive heuristic, which identifies the head noun occurring most frequently across the potential class labels $\mathcal{C}$ of an instance $\mathcal{I}$, then discards the labels whose head nouns are not the most frequent head noun. For example, since the most frequent head of the labels associated with *bnp paribas* is *banks*, class labels such as *european banks* and *largest french commercial banks* are retained, whereas *members*, *hong kong customers* or *financial institutions* are discarded, thus promoting precision of the class labels at the expense of lower recall. After filtering, the resulting pairs of an instance and a label are arranged into instance sets (e.g., {*deutsche bank*, *abn amro*, *bnp paribas*,...}), each associated with a class label (e.g., *european banks*).

Attributes of Labeled Classes of Instances: The labeled classes of instances collected automatically from Web documents are passed as input to the second extraction phase (top-left in Figure 1), which acquires class attributes by mining a collection of Web search queries. The attributes capture properties that are relevant to the class. The extraction of attributes exploits the set of class instances rather than the associated class label, and has four stages as described in [8]:

1) identification of a noisy pool of candidate attributes, as remainders of queries that also contain a class instance. In the case of the class *companies*, whose instances include *delphi* and *apple computer*, the queries *"installing delphi"* and *"apple computer headquarters"* produce the candidate attributes *installing* and *headquarters*;

2) construction of internal search-signature vector representations for each candidate attribute obtained in the first stage, based on a second pass over queries (e.g., *"coca cola company one year stock price target"*) that contain a candidate attribute (*stock price*) and a class instance (*coca cola*). These vectors consist of counts tied to the frequency with which an attribute occurs with "templatized" queries. The latter are automatically derived from the original queries, by replacing specific attributes and instances with common placeholders, e.g., *"X for Y"*. The query *"coca cola company one year stock price target"* results in a new entry being added to the search-signature vector of *stock price* with respect to the class *companies*, corresponding to the query template $[\ ]_{prefix}$ $[company\ one\ year]_{infix}$ $[target]_{postfix}$;

3) construction of a reference internal search-signature vector representation for a small set of seed attributes (e.g., *headquarters* and *stock price* for *companies*) provided as input. A reference vector is the normalized sum of the individual vectors corresponding to the seed attributes;

4) ranking of candidate attributes with respect to each class (e.g., *companies*), by computing similarity scores between their individual vector representations and the reference vector of the seed attributes.

The result of the four stages is a ranked list of attributes (e.g., [*headquarters*, *mission statement*, *stock price*,...]) for each class (e.g., *companies*).

The instances of each input class are automatically generated as described earlier, rather than manually assembled. Furthermore, the amount of supervision is limited to seed attributes being provided for only one of the classes, whereas [8] requires seed attributes for each class. To this effect, the extraction includes modifications such that only one reference vector is constructed internally from the seed attributes during the third stage, rather one such vector for each class in [8]; and similarity scores are computed cross-class by comparing vector representations of individual candidate attributes against the only reference vector available during the fourth stage, rather than with respect to the reference vector of each class in [8].

2.3 Linking Labeled Classes into Hierarchies

Conceptual Hierarchies: Manually-constructed language resources such as WordNet provide reliable, wide-coverage upper-level conceptual hierarchies, by grouping phrases with the same meaning (e.g., {*analgesic*, *painkiller*, *pain pill*}) into sets of synonyms (or synsets, in WordNet terminology) associated with the same definition (e.g., "a medicine used to relieve pain"). Synsets are organized into conceptual hierarchies (e.g., *painkillers* are a subconcept, or a hyponym, of *drugs*) [3].

To determine the points of insertion of automatically-extracted labeled classes under hand-built WordNet hierarchies, the class labels are looked up in WordNet using built-in morphological normalization routines. When a class label (e.g., *age-related diseases*) is not found in WordNet, it is looked up again after iteratively removing its leading words (e.g., *related diseases*, and *diseases*) until a potential point of insertion is found where one or more senses exist in WordNet for the class label. As explained below, one of the available senses is chosen as the point of insertion of the class label and its associated instances, thus extending the conceptual hierarchies with instances acquired from text.

First-Sense Selection: An efficient heuristic for sense selection is to uniformly choose the first (that is, most frequent) sense of the label in WordNet, as point of insertion. Due to its simplicity, the heuristic is bound to make errors whenever the correct sense is not the first one, thus incorrectly linking *academic journals* under the sense of *journals* as personal diaries rather than periodicals, and *active volcanoes* under the sense of *volcanoes* as fissures in the earth, rather than mountains formed by volcanic material. Nevertheless, previous experimental results on linking Wikipedia categories [13] to WordNet concepts suggest that first-sense selection may be more effective in practice than other techniques [15]. Thus, a class label and its associated instances are inserted under the first WordNet sense available for the class label.

Similar-Sense Selection: Rather than always choosing the first of the senses available in WordNet for a class label, a more intuitively appropriate heuristic is to select the sense to which the set of instances associated to the class label is the most similar semantically. The semantic similarity between the set of instances, on one hand, and each sense from WordNet, on the other hand, is approximated through the distributional similarities [6] collected from Web documents between individual instances, on one hand, and sense-descriptive phrases collected from WordNet, on the other hand. The phrases considered to be descriptive of a given WordNet sense (e.g., the second sense of *journals*) are: a) synonyms, b) siblings, or coordinate terms (e.g., *review*, *digest*, *issue*), c) immediate superconcepts, or hypernyms (e.g., *periodical*) and d) immediate subconcepts, or hyponyms (e.g., *annals*) available in the WordNet hierarchies around that sense. For each of these four types of sense-descriptive phrases, a scoring function aggregates the individual distributional similarity scores $DistSim$ between each instance $\mathcal{I}$ and each descriptive phrase $\mathcal{P}_T$ of type T, normalized by the counts of instances and descriptive phrases. A linear combination of these four scores constitutes the score between a class label and one of the available WordNet senses:

$$SimScore = \sum_{T} (w_T \times \frac{\sum_{\mathcal{I},\mathcal{P}_T} DistSim(\mathcal{I}, \mathcal{P}_T)}{\log(1 + |\{\mathcal{I}\}| \times |\{\mathcal{P}_T\}|)})$$

where T is one of the four types of descriptive phrases (e.g., synonyms) and w_T is the relative weight assigned to that type. Thus, a class label and its associated senses are inserted under the WordNet sense with the highest similarity score $SimScore$. In case of ties, the insertion falls back to choosing the sense, out of the tied senses, that is most frequent in WordNet.

2.4 Estimation of Attribute Specificity

Hierarchical Propagation: As mentioned earlier, a labeled class is accompanied by a ranked list of attributes extracted from query logs. With the labeled classes and their associated sets of instances linked under various concepts from the conceptual hierarchies, either based on first or on similar-sense selection, it is possible to propagate the attributes upwards over the conceptual hierarchies. The set of possible hierarchy concepts to which an attribute may apply is restricted to all concepts that are superconcepts, up to the hierarchy roots, of the class labels for which the attribute was extracted from query logs. For example, *headquarters* is

among the attributes extracted for the class labels *european banks* and *car companies* (top-left in Figure 1). Since the two class labels are inserted under the hierarchy concepts *Bank* and *CarCompany* respectively (top-right in Figure 1), any of the superconcepts *Bank*, *FinancialInstitution*, *Institution*, *CarCompany* or *Company* may be the correct level of specificity of the attribute *headquarters* relative to the hierarchy.

Coverage-Based Attribute Specificity: To assess the level of specificity to which an attribute corresponds in the hierarchy, a score is computed for the attribute with each of the hierarchy concepts to which the attribute may belong. The formula that computes the score of a hierarchy concept $\mathcal{H}$, for some attribute $\mathcal{A}$, promotes concepts $\mathcal{H}$ for which more of the instances, from their inherited subconcepts $\mathcal{C} \subset \mathcal{H}$, co-occur with the attribute $\mathcal{A}$ in some query $\mathcal{Q}$ from query logs:

$$CvgScore(\mathcal{A}, \mathcal{H}) = \frac{|\{\mathcal{I} : \mathcal{I} \in \mathcal{C}, \mathcal{C} \subset \mathcal{H}, \mathcal{Q}(\mathcal{A}, \mathcal{I})\}|}{\log(1 + |\{\mathcal{I} : \mathcal{I} \in \mathcal{C}, \mathcal{C} \subset \mathcal{H}\}|)}$$

The computed scores define the relative ranking of hierarchy concepts for an attribute, such that the attribute can be placed within the hierarchy under the concept with the highest score. As illustrated earlier in the top-right part of Figure 1, attributes such as *headquarters* are thus associated with the concepts to which they best apply, in this case *Institution*.

3. EXPERIMENTAL SETTING

Textual Data Sources: The acquisition of open-domain knowledge 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 queries is a random sample of fully-anonymized queries in English submitted by Web users in 2006. The sample contains about 50 million unique queries. Each query is accompanied by its frequency of occurrence in the logs. Other sources of similar data are available publicly for research purposes [4]. The document collection consists of around 100 million documents in English, as available in a Web repository snapshot from 2006. The textual portion of the documents is cleaned of HTML, tokenized, split into sentences and part-of-speech tagged using the TnT tagger [2].

Parameters for Extracting Labeled Classes: The extraction method collects labeled classes of instances from the input documents. During pattern matching, the instance boundaries are approximated by checking that the collected instances occur among the top five million queries with the highest frequency within the input query logs. The extracted data is further filtered by discarding classes with fewer than 25 instances, and retaining the top 100 instances in each class. The labeled classes are linked under conceptual hierarchies available within WordNet 3.0, which contains a total of 117,798 English noun phrases grouped in 82,115 concepts (or synsets). The extracted set of labeled classes consists of 9,519 class labels associated to a total of 199,571 unique instances, all of which are linked under WordNet concepts.

Parameters for Extracting Class Attributes: The degree of supervision for extracting attributes of labeled classes is limited to 5 seed attributes (*population*, *area*, *president*, *flag* and *climate*) provided for only one of the extracted labeled classes, namely *european countries*. The top 50 attributes extracted for each class are retained for the upward propagation towards higher-level WordNet concepts under which the class labels are linked.

Parameters for Sense Selection: During the linking of labeled classes to conceptual hierarchies from WordNet via similar-sense selection, the scoring weights w_T are set to 0.5 for synonyms, 0.2 for superconcepts (hypernyms) and subconcepts (hyponyms), and 0.1 for coordinate terms. The distributional similarities that support the selection of senses are collected from the input collection of Web documents based on [6].

4. EVALUATION

4.1 Flat-Class Attributes

Accuracy of Extracted Attributes: Table 1 shows class labels, instances and attributes extracted from text, before propagation in WordNet, for a random sample of 200 classes out of the larger set of 9,519 labeled classes acquired from text. The accuracy of a ranked list of attributes acquired for a class is computed in accordance with methodology previously introduced in [8], by manually assigning a correctness label to each attribute of the ranked lists, such that an attribute is *vital* if it must be present in an ideal list of attributes of the class; *okay* if it provides useful but non-essential information; and *wrong* if it is incorrect. To compute the overall precision score over a ranked list of extracted attributes, the correctness labels are converted to numeric values (1.0 for *vital*, 0.5 for *okay*, 0.0 for *wrong*), and precision at some rank N in the list is thus measured as the sum of the assigned values of the first N candidate attributes, divided by N. As shown in Table 1, some of the extracted class labels (e.g., *central cities*) are relatively ambiguous, and their class attributes may be incorrect (e.g., *state flag*, *natural history museum* and *sun newspaper* for the class *central cities*). The precision of the extracted attributes, as an average over the sample of 200 classes, is 0.71 at rank 10, and 0.65 at rank 20.

4.2 Hierarchical Attributes

Experimental Runs: The experiments consist of six different runs, which correspond to different choices for the source of conceptual hierarchies and class instances linked to those hierarchies, as illustrated in Table 2. In the first run, denoted N, the class instances are those available within the latest version of WordNet (3.0) itself via HasInstance relations. In the second and third runs from Table 2, K_1 and K_2, the class instances are those added to an earlier version of WordNet (2.1) as part of previous work [14]. K_1 does not include the HasInstance instances already available in WordNet, whereas K_2 includes them. The fourth run from Table 2, Y, corresponds to an extension of WordNet based on the manually-assembled classes of instances from categories in Wikipedia, as available in the 2007-w50-5 version of Yago [15]. Note that runs N, K_1, K_2 exploit resources of class instances created as part of previous work, without making any changes to those resources. The last two runs from Table 2, E_f and E_s, correspond to the fully-fledged extraction from unstructured text described in this paper. In E_f, class labels are linked to the first sense available at the point of insertion in WordNet, whereas in E_s the class labels are linked to the most similar sense.

Gold Standard: As shown in Table 3, the estimation of the level of specificity of attributes over conceptual hierarchies is evaluated over a gold standard of open-domain attributes

Class		Precision of Attributes			Top Ten Extracted Attributes
#	Class Label={Set of Instances}	@5	@10	@20	
1	acids={hydrochloric acid, nitric acid, phosphoric acid, acetic acid, lactic acid,...}	1.00	0.95	0.92	molecular formula, melting point, titration curve, food sources, molar mass, molecular weight, pka, chemical structure, material safety data sheet, extinction coefficient
10	athletes={bodybuilders, runners, michael jordan,...}	1.00	0.90	0.92	autobiography, biography, nickname, bibliography, childhood, life story, charities, timeline, screensaver, bio
30	central cities={chicago, austin, philadelphia, brasilia, milan, tokyo, nanjing, denver,...}	0.60	0.70	0.65	climate, population, state flag, natural history museum, geography, skyscrapers, subway system, sun newspaper, streetmap, best restaurants
50	designers={hanae mori, donna karan, giorgio armani, oscar de la renta, issey miyake,...}	0.80	0.70	0.85	fashion shows, logo, fashion designer, bridal dresses, stephanie, little black dress, wedding gowns, sarah jessica parker, fashions, bridal collection
70	firms={merrill lynch, lehman brothers, ernst & young,...}	1.00	0.95	0.80	mission statement, competitors, logo, company profile, swot analysis, history, meaning, ceo, webmail, official website
90	investors={venture capital firms, warren buffett,...}	0.80	0.85	0.90	biography, bio, quotes, investment strategies, definition, profile, family, photos, directory, ranking
110	motor vehicles={trucks, trailers, motorcycles, cars, buses, boats, tractors,...}	0.30	0.25	0.37	clipart, blue book value, coloring pages, flash games, screensavers, bluebook, kelly blue book, history, graphics, compass
130	populations={uninsured, developmentally disabled,...}	0.40	0.35	0.52	definition, stress, meaning, psychology, clip art, poems, magazine, origin, cartoon, symptoms
150	sciences={botany, cryobiology, physics, ethology,...}	0.80	0.70	0.75	timeline, glossary, history, dictionary, encyclopedia, definition, animations, exercises, careers, picture
170	stories={three little pigs, fairy tales, ballet shoe,...}	0.90	0.75	0.72	book review, theme, parody, author, book summary, clipart, symbolism, coloring page, synopsis, chapter summaries
190	tumors={neuroblastoma, meningiomas, gliomas, osteosarcoma, melanomas,...}	1.00	1.00	0.92	histopathology, pathophysiology, definition, epidemiology, immunohistochemistry, differential diagnosis, genetics, cytology, cytogenetics, staging
200	zoonotic diseases={rabies, west nile virus, leptospirosis, brucellosis, lyme disease,...}	1.00	1.00	1.00	scientific name, causative agent, mode of transmission, life cycle, pathology, meaning, prognosis, incubation period, symptoms, phylum
	Average-Class (200 classes)	**0.75**	**0.71**	**0.65**	

Table 1: Open-domain flat classes and associated attributes extracted from unstructured text

Description	Source of Hierarchy and Instances					
	N	K_1	K_2	Y	E_f	E_s
WordNet version	3.0	2.1	2.1	3.0	3.0	3.0
Include instances from WordNet?	√	-	√	√	-	-
Include instances from elsewhere?	-	√	√	√	√	√
Total instances ($\times 10^3$)	17.4	110.2	127.6	1,296.5	199.5	199.5
Total classes	945	2,465	3,078	30,338	9,519	9,519

Table 2: Source of conceptual hierarchy and class instances for various experimental runs

and WordNet concepts to which the attributes ideally apply. In the gold standard, the attributes are randomly selected from among the relevant (*vital* or *okay*) attributes acquired by various runs, whereas their WordNet gold concepts (one or more per attribute) are identified manually. For instance, the attribute *national anthem* is associated to a synset situated at the internal offset 08544813 in WordNet 3.0, which groups together the synonymous phrases {*country*, *state*, *land*} and provides them with the definition *"the territory occupied by a nation"*. Since some of the experimental runs rely on WordNet 2.1 rather than WordNet 3.0, the gold attributes are also manually mapped to WordNet 2.1 synsets, whose internal offsets are different but are semantically equivalent to their WordNet 3.0 counterparts, with respect to component phrases, associated definitions, and localization within the conceptual hierarchy.

The concepts from the gold standard exhibit variation with respect to their depths within WordNet conceptual hierarchies, ranging from a minimum of 3 for {*object*, *physical object*}, which is manually specified for the gold attribute *colors*, to a maximum of 13 for {*airliner*}, which is one of the two gold concepts specified for the gold attribute *seating chart*. Besides possibly being associated to more than one concept, attributes from gold standard may in fact be associated to multiple concepts corresponding to different meanings of the attribute (e.g., *volume* of an *academic journal* vs. *volume* of a *planet*. Overall, the gold standard contains 150 unique attributes linked to 78 unique WordNet concepts, thus seeking varied experimentation on several dimensions, while taking into account the time intensive nature of manual accuracy judgments often required in the evaluation of information extraction systems [1].

To assess the reliability of the gold standard, a second, temporary gold standard is created independently by another human annotator for the same set of 150 gold attributes. A comparison of the synsets manually selected as part of the gold standard by the two annotators indicates an inter-annotator agreement of 0.68, which we find to be acceptable given that WordNet senses are often too fine-grained [10].

Evaluation Procedure: For each experimental run, the output consists of pairs of an attribute and a ranked list of concepts to which the attribute is most likely to belong, according to the coverage-based score described in Sec-

Gold-Standard Attribute	Gold-Standard Concept		
	Synset	Offset	Definition
circulatory system	{organism, being}	00004475	a living thing that has (or can develop) the ability to act or function independently
civilian casualties	{military action, action}	00952963	a military engagement
	{operation, military operation}	00955060	activity by a military or naval force (as a maneuver or campaign)
famous alumni	{educational institution}	08276342	an institution dedicated to education
market share	{company}	08058098	an institution created to conduct business
national anthem	{country, state, land}	08544813	the territory occupied by a nation
paintings	{painter}	10391653	an artist who paints
subway system	{city, metropolis, urban center}	08524735	a large and densely populated urban area

Table 3: Sample from gold-standard of attributes and manually-selected WordNet concepts to which the attributes ideally apply

tion 2.4. Only the top 10 concepts returned in each ranked list are considered. Furthermore, the evaluation is restricted to the subset of output attributes that are gold-standard attributes. Note that the subsets of attributes may be different among experimental runs that rely on distinct sets of labeled classes, which is the case for all runs except E_f and E_s (see Table 2). The accuracy of a ranked list of concepts returned for a gold-standard attribute is measured by two scoring metrics that correspond to the mean reciprocal rank score (MRR) [17] and a modification of it (DRR):

$$MRR = \max \frac{1}{rank}, \quad DRR = \max \frac{1}{rank \times (1 + PathToGold)}$$

where *rank* is the rank (from 1 up to 10) of a concept in the returned list, and PathToGold is the length of the minimum path (along IsA edges) in the conceptual hierarchies between the concept, on one hand, and any of the gold-standard concepts specified in the gold standard for the attribute, on the other hand. The length PathToGold is minimum, that is, 0, if the returned concept is the same as the gold-standard concept. Conversely, a gold-standard attribute receives no credit (that is, DRR is 0) if no path is found in the hierarchies between the top 10 returned concepts and any of the gold-standard concepts, or if the ranked list of concepts returned for the attribute is empty. The accuracy of an experimental run is the average of DRRs of individual attributes, computed over the gold standard of attributes. As an illustration, the first concept returned for the attribute *band members* in run E_s is the synset {*dance band*, *band*, *dance orchestra*}. Since the latter is a direct subconcept of the gold-standard concept {*musical organization*, *musical organisation*, *musical group*} in WordNet, the distance PathToGold is 1, and therefore the computed accuracy score for the attribute *band members* is 0.5.

Qualitative Results: Table 4 is a view on the accuracy of the ranked lists of concepts extracted for various attributes by the experimental runs. The first row shows the number of gold-standard attributes, which is fixed across the runs. The second and third rows show the (variable) number of gold-standard attributes for which some concepts are returned in each individual run, and the number of gold-standard attributes for which the correct concept is returned at rank 1, thus receiving maximum credit. For example, for run N, non-empty ranked lists of concepts are returned for 48 of the 150 gold-standard attributes, and 11 of those 48 attributes have the gold-standard concept returned at rank 1.

The fourth row from Table 4 illustrates the fraction of gold-standard attributes for which the gold-standard concept is returned at rank 1, thus receiving maximum credit, computed over the subsets of attributes for which some concepts were returned in each individual run. Thus, the scores shown in the fourth row from Table 4 are equivalent to strict precision@1 scores. The fifth and sixth rows show the corresponding MRR and DRR scores. The scores in the fourth through sixth rows focus on the precision of the returned ranked lists of concepts, thus rewarding runs that provide high accuracy even if they do so over few of the gold-standard concepts. In contrast, the scores shown in the seventh and ninth rows take into consideration the entire set of 150 gold-standard attributes, thus capturing both precision and recall. In the case of run N from the table, the DRR scores are 0.438 over the 48 gold-standard attributes for which some concepts are returned, and 0.140 over the entire set of 150 gold-standard attributes.

Scoring Metric	Experimental Run					
	N	K_1	K_2	Y	E_f	E_s
|A|	150	150	150	150	150	150
|R|	48	44	74	94	**143**	**143**
|F|	11	10	21	9	**26**	**28**
|F|/|R|	**0.229**	0.227	**0.283**	0.095	0.181	0.195
MRR_R	**0.332**	0.278	**0.376**	0.137	0.302	0.310
DRR_R	**0.438**	0.355	**0.452**	0.283	0.402	0.423
|F|/|A|	0.073	0.067	0.140	0.060	**0.173**	**0.186**
MRR_A	0.106	0.081	0.185	0.086	**0.288**	**0.295**
DRR_A	0.140	0.104	0.223	0.177	**0.379**	**0.403**

Table 4: Accuracy of ranked lists of WordNet concepts extracted for various runs. A=entire set of 150 gold-standard attributes; R=(variable) subsets of the 150 gold-standard attributes for which some concepts are returned in each individual run; F=(variable) subsets of the gold-standard 150 attributes for which the returned ranked list of WordNet concepts has a DRR score of 1.0; DRR_R=average DRR score over R subsets only; DRR_A=average DRR score over A

In the fourth, fifth and sixth rows from the table, the runs with the highest accuracy over the subsets of attributes for which some concepts were returned are K_2 and N. Both runs take advantage heavily (run K_2) or exclusively (run N) of manually-compiled instances already available within WordNet, as indicated earlier in Table 2. In other words, if the available classes of instances are compiled manually, then the placement of relevant attributes over conceptual hierarchies can exceed DRR scores of 0.45. When considering the accuracy over the entire set of 150 gold-standard attributes, in the seventh, eight and ninth rows, the relative performance of the various runs changes. Since WordNet is

Gold-Standard Attribute	Gold-Standard Concept Synset (Offset)	Returned Concept Being Evaluated			
		Rank	Synset (Offset)	PathToGold	DRR
circulatory system	{organism, being} (00004475)	3	{marine animal, sea animal,...} (01319467)	2	0.167
civilian casualties	{military action, action} (00952963)	-	no concept returned	-	0.000
famous alumni	{educational institution} (08276342)	1	{educational institution} (08276342)	0	1.000
national anthem	{country, state, land} (08544813)	2	{country, state, land} (08544813)	0	0.500
paintings	{painter} (10391653)	1	{physical entity} (00001930)	7	0.142
subway system	{city, metropolis, urban center} (08524735)	3	{city, metropolis, urban center} (08524735)	0	0.333

Table 5: Evaluation scores computed for various concepts from ranked lists of concepts returned in run E_s

not meant to be an encyclopedic resource, it contains a limited number of instances. Therefore, run N relies on fewer instances and class labels than other runs (see Table 2), a disadvantage that is apparent in run N obtaining low scores in the seventh through ninth rows from Table 4.

The significant advantage of run Y, which has access to the largest number of instances and class labels, does not result in a more accurate placement of the attributes over conceptual hierarchies. In fact, run Y returns the correct gold-standard concept at rank 1 for the fewest (that is, 9 in the third row) gold-standard attributes among all runs, and produces relatively low DRR scores (e.g., 0.177 in the seventh row of Table 4).

The placement of attributes over conceptual hierarchies is more accurate, when the labeled classes acquired from text are inserted under the most similar, rather than first sense available for them in WordNet. In comparison, previous work notes that WordNet senses are often too fine-grained, making the task of choosing the correct sense difficult even for humans [10], and shows that choosing the first sense from WordNet is sometimes better than more intelligent disambiguation techniques [12].

The runs using our automatically-extracted labeled classes (E_f and E_s) clearly outperform not only runs using manually-compiled labeled classes (N and Y), but also other runs using automatically-extracted labeled classes (K_1 and K_2). Concretely, the DRR scores over the entire set of gold-standard attributes are 0.384 for E_f and 0.403 for E_s, which correspond to improvements of 72% and 80% respectively over the next best run, namely K_2 with a DRR score of 0.223. Table 5 illustrates some of the scores assigned to various gold-standard concepts, based on ranked lists of concepts returned for them in run E_s.

Impact of Parameters for Sense Selection: As specified earlier in Section 3, the results for run E_s are obtained after setting the scoring weights in the *SimScore* formula introduced in Section 2.3 to 0.5 for synonyms, 0.2 for superconcepts (hypernyms) and subconcepts (hyponyms), and 0.1 for coordinate terms. Two issues that require further investigation are the impact of the smoothing effect introduced by using the log, rather than raw frequencies, in the denominator of the *SimScore* formula from Section 2.3; and the impact of alternative settings of the scoring weights in the same formula. Table 6 summarizes results from separate experiments, as the log is either removed or retained in the formula, and the weights are set to either the default values indicated above, or to 0.005 for synonyms, 0.003 for superconcepts (hypernyms), 0.990 for subconcepts (hyponyms), and 0.002 for coordinate terms. The accuracy of the computed ranked lists of WordNet concepts is affected more by the use of the log than it is by the choice of scoring weights.

Scoring Parameters	DRR
w_T=[0.5, 0.2, 0.2, 0.1], log=on	0.403
w_T=[0.5, 0.2, 0.2, 0.1], log=off	0.368
w_T=[0.005, 0.003, 0.99, 0.002], log=on	0.403
w_T=[0.005, 0.003, 0.99, 0.002], log=off	0.357

Table 6: Impact of alternative scoring parameters in the *SimScore* formula from Section 2.3. The accuracy of the ranked lists of WordNet concepts computed with various settings is measured as the average DRR score over the entire set of 150 gold-standard attributes. w_T are the scoring weights corresponding to synonyms, superconcepts (hypernyms), subconcepts (hyponyms), and coordinate terms respectively

Comparison to Previous Results: A previous method [9] iteratively computes ranked lists of potential attributes for concepts within an hierarchy, from ranked lists of attributes retrieved for flat classes inserted under concepts situated lower in the hierarchies. In order to evaluate how accurately that method can determine the level of specificity of an attribute, its ranked lists of attributes computed for various concepts must first be converted into ranked lists of concepts for various attributes. This is done for each attribute, by sorting, in decreasing order, the inverse ranks of the attribute within the ranked list of attributes computed for that concept, with simple alphabetical ranking of classes in case of inverse-rank ties. Thus, the DRR scores for [9], over the same set of 150 gold-standard attributes, are 0.285, when computed over only the subset of 135 gold-standard attributes for which some concepts are returned in [9]; and 0.256, when computed over the entire set of 150 gold-standard attributes. In comparison, the method presented in the current paper obtains DRR scores of 0.423 and 0.403 respectively, when taking advantage of the same set of flat classes and the same WordNet hierarchies as in [9].

5. RELATED WORK

Previous work on extracting attributes from unstructured text takes advantage of existing classes of instances, in order to acquire attributes from Web documents [16] or query logs [8]. The input classes of instances are conveniently assumed to be independent from one another, and be part of flat sets of classes, rather than conceptual hierarchies. As

a result, it is common to extract attributes that, even if relevant, are in fact attributes of superconcepts from which they should be inherited. For instance, extracted attributes include *name*, *species*, *picture*, *evolution* and *characteristics* for *Plant* in [16], or *age* and *date of birth* for *Actor* in [8].

The role of conceptual hierarchies in the acquisition of class attributes is explored only very recently in [9], but strictly to iteratively compute ranked lists of potential attributes for concepts situated higher in the hierarchies, from ranked lists of attributes retrieved from text for classes inserted under concepts situated lower in the hierarchies. In contrast, we address the more difficult task of identifying the concepts within conceptual hierarchies, to which various extracted attributes best apply, rather than simply computing ranked lists of attributes for the hierarchy concepts. Furthermore, whereas the available classes of instances are uniformly linked in [9] under the first sense available in WordNet, our method also explores the insertion under the sense that is the most semantically similar to the overall set of instances being inserted. Unlike [15], who note that first-sense selection provides the best performance over more complex alternatives in a particular task-based evaluation, our experimental results show an improvement over first-sense selection, for the task of placing attributes over conceptual hierarchies. Our method is related to previous work on ontologizing relations acquired from text [11], and on identifying predominant senses of words within various text corpora based on distributional similarities [7].

6. CONCLUSION

This paper introduces an extraction framework for exploiting labeled classes of instances acquired from a combination of documents and search query logs, to extract attributes over conceptual hierarchies. The insertion of the classes under existing conceptual hierarchies allows for the placement of attributes over the hierarchies, without a-priori restrictions to specific domains of interest. The experiments indicate that the placement is better if the insertion of labeled classes of instances under hierarchy concepts exploits distributional similarities, rather than simply choosing the most frequent of the available senses. The placement of attributes over hierarchies is more accurate when using text-derived, rather than manually-compiled classes of instances available within other resources. Current work investigates the impact of the semantic distribution of the classes of instances on the overall accuracy of attribute placement.

7. REFERENCES

[1] M. Banko and O. Etzioni. The tradeoffs between open and traditional relation extraction. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-08)*, pages 28–36, Columbus, Ohio, 2008.

[2] T. Brants. 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, 2000.

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

[4] W. Gao, C. Niu, J. Nie, M. Zhou, J. Hu, K. Wong, and H. Hon. Cross-lingual query suggestion using query logs of different languages. In *Proceedings of the 30th ACM Conference on Research and Development in Information Retrieval (SIGIR-07)*, pages 463–470, Amsterdam, The Netherlands, 2007.

[5] M. Hearst. 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, 1992.

[6] D. Lin and P. Pantel. Concept discovery from text. In *Proceedings of the 19th International Conference on Computational linguistics (COLING-02)*, pages 1–7, Taipei, Taiwan, 2002.

[7] D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4):553–590, 2007.

[8] M. Paşca. Organizing and searching the World Wide Web of facts - step two: Harnessing the wisdom of the crowds. In *Proceedings of the 16th World Wide Web Conference (WWW-07)*, pages 101–110, Banff, Canada, 2007.

[9] M. Paşca. Turning Web text and search queries into factual knowledge: Hierarchical class attribute extraction. In *Proceedings of the 23rd National Conference on Artificial Intelligence (AAAI-08)*, pages 1225–1230, Chicago, Illinois, 2008.

[10] M. Palmer, H. Dang, and C. Fellbaum. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(2):137–163, 2007.

[11] M. Pennacchiotti and P. Pantel. Ontologizing semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL-06)*, pages 793–800, Sydney, Australia, 2006.

[12] S. Pradhan, E. Loper, D. Dligach, and M. Palmer. SemEval-2007 Task-17: English lexical sample, SRL and all words. In *Proceedings of the 4th Workshop on Semantic Evaluations (SemEval-07)*, pages 87–92, Prague, Czech Republic, 2007.

[13] M. Remy. Wikipedia: The free encyclopedia. *Online Information Review*, 26(6):434, 2002.

[14] R. Snow, D. Jurafsky, and A. Ng. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL-06)*, pages 801–808, Sydney, Australia, 2006.

[15] F. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge unifying WordNet and Wikipedia. In *Proceedings of the 16th World Wide Web Conference (WWW-07)*, pages 697–706, Banff, Canada, 2007.

[16] K. Tokunaga, J. Kazama, and K. Torisawa. Automatic discovery of attribute words from Web documents. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*, pages 106–118, Jeju Island, Korea, 2005.

[17] E. Voorhees and D. Tice. Building a question-answering test collection. In *Proceedings of the 23rd International Conference on Research and Development in Information Retrieval (SIGIR-00)*, pages 200–207, Athens, Greece, 2000.