Interpreting Compound Noun Phrases Using Web Search Queries

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

Abstract

A weakly-supervised method is applied to anonymized queries to extract lexical interpretations of compound noun phrases (e.g., *"fortune 500 companies"*). The interpretations explain the subsuming role (*"listed in"*) that modifiers (*fortune 500*) play relative to heads (*companies*) within the noun phrases. Experimental results over evaluation sets of noun phrases from multiple sources demonstrate that interpretations extracted from queries have encouraging coverage and precision. The top interpretation extracted is deemed relevant for more than 70% of the noun phrases.

1 Introduction

Motivation: Semantic classes of interest to Web users are often expressed as lexical class labels (e.g., *"fortune 500 companies"*, *"italian composers"*, *"victorinox knives"*). Each class label hints at the implicit properties shared among its instances (e.g., *general electric*, *gaetano donizetti*, *swiss army jetsetter* respectively). Class labels allow for the organization of instances into hierarchies, which in turn allows for the systematic development of knowledge repositories. This motivates research efforts to acquire as many relevant class labels of instances as possible, which have received particular emphasis (Wang and Cohen, 2009; Dalvi et al., 2012; Flati et al., 2014). The efforts are part of the larger area of extracting open-domain facts and relations (Banko et al., 2007; Hoffart et al., 2013; Yao and Van Durme, 2014), ultimately delivering richer results in Web search.

Different methods can associate instances (*general electric*) with both class labels (*"fortune 500 companies"*) and facts (<*general electric*, *founded in*, *1892*>) extracted from text. But the class labels tend to be extracted, maintained and used separately from facts. Beyond organizing the class labels hierarchically (Kozareva and Hovy, 2010), the meaning of a class label is rarely explored (Nastase and Strube, 2008), nor is it made available downstream to applications using the extracted data.

Contributions: The method introduced in this paper is the first to exploit Web search queries to uncover the semantics of open-domain class labels in particular; and of compound noun phrases in general. The method extracts candidate, lexical interpretations of compound noun phrases from queries. The interpretations turn implicit properties or subsuming roles (*"listed in"*, *"from"*, *"made by"*) that modifiers (*fortune 500*, *italian*, *victorinox*) play within longer noun phrases (*"fortune 500 companies"*, *"italian composers"*, *"victorinox knives"*) into explicit strings. The roles of modifiers relative to heads of noun phrase compounds cannot be characterized in terms of a finite list of possible compounding relationships (Downing, 1977). Hence, the interpretations are not restricted to a closed, pre-defined set. Experimental results over evaluation sets of noun phrases from multiple sources demonstrate that interpretations can be extracted from queries for a significant fraction of the input noun phrases. Without relying on syntactic analysis, extracted interpretations induce implicit bracketings over the interpreted noun phrases. The bracketings reveal the multiple senses, some of which are more rare but still plausible, in which the same noun phrase can be sometimes explained. The quality of interpretations is encouraging, with at least one interpretation deemed relevant among the top 3 retrieved for 77% of the

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 335–344,
Denver, Colorado, May 31 – June 5, 2015.

noun phrases with extracted interpretations. The top interpretation is deemed relevant for more than 70% of the noun phrases.

Applications: The extracted interpretations can serve as a bridge connecting class labels and facts. Relevant interpretations allow one to potentially derive missing facts (<*general electric*, *listed in*, *fortune 500*>) from existing class labels (<*general electric*, *fortune 500 companies*>) and vice versa. In addition, relevant interpretations of class labels are themselves class labels inferred for the same instances. Examples are <*general electric*, *companies listed in fortune 500*>, or <*general electric*, *companies in fortune 500*>, based on <*general electric*, *fortune 500 companies*>. If the input class labels are organized hierarchically (<*fortune 500 companies*, *companies*>), interpretations explain why more specific class labels (*"fortune 500 companies"*, *"german companies"*, *"dow jones industrial average companies"*, *"french companies"*) do not merely belong under more general ones (*"companies"*), but do so along shared interpretations (*companies*→*listed in*→{*fortune 500*, *dow jones industrial average companies*}; vs. {*companies*→*from*→{*germany*, *france*}); and, more generally, aid in the better understanding of noun phrases.

2 Interpreting Noun Phrases

Hypothesis: Let N be a compound noun phrase, containing a head H preceded by modifiers M. Each of H and M may contain one or multiple tokens. Being a compound, the sequence of modifiers and head in N act as a single noun (Downing, 1977; Hendrickx et al., 2013). If N is relevant and of interest to Web users, then in a sufficiently large corpus it will eventually be referred to in relatively more verbose search queries, which explain the implicit role that modifiers M play relative to the head H.

Acquisition from Queries: To illustrate the intuition above, consider the noun phrases *"water animals"* and *"zone 7 plants"*. If enough Web users are interested in the concepts represented by these noun phrases, then the phrases are likely to be submitted as search queries. In addition, some Web users seeking similar information are likely to submit queries that make the role of the modifiers *water* and *zone 7* explicit, such as *"animals living in water"* or *"plants that grow in zone 7"*.

As illustrated in Figure 1, the extraction method

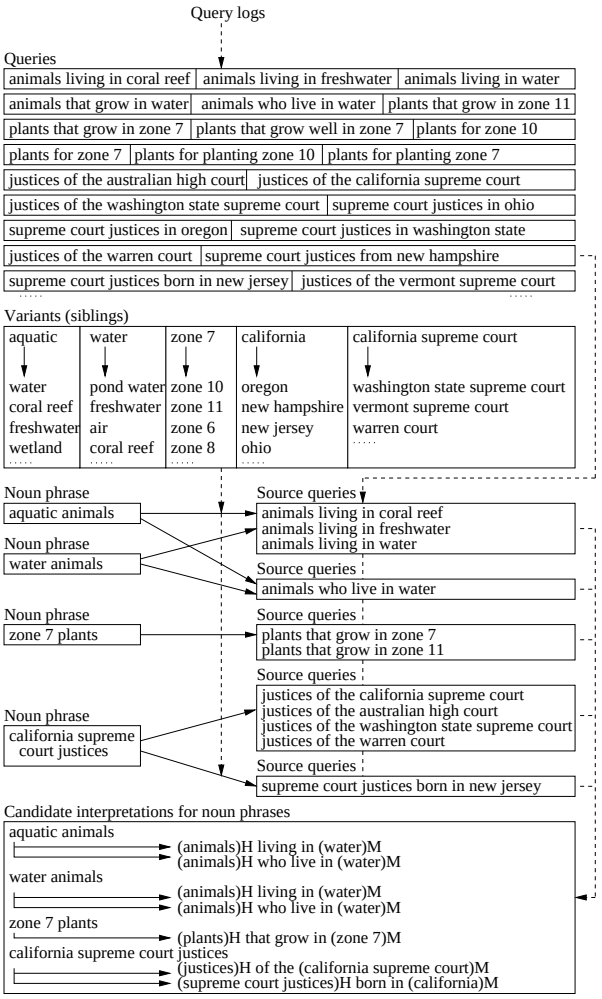

Figure 1: Overview of extraction of interpretations of noun phrases from Web search queries

proposed in this paper takes as input a vocabulary of noun phrases, as well as a set of anonymized queries from which possible interpretations for the noun phrases must be extracted. The extraction consists of several steps: (1) the selection of a subset of queries that may be candidate interpretations of some yet-to-be-specified noun phrases; (2) the matching of the selected queries to the noun phrases to interpret; and (3) the aggregation of matched queries into candidate interpretations extracted for a noun phrase.

Queries as Candidate Interpretations: The input queries are matched against the extraction patterns from Table 1. The use of targeted patterns in information extraction has been suggested before (Hearst, 1992; Fader et al., 2011). In our case, the patterns match queries that start with an arbitrary ngram H, followed by what is likely a

Extraction Pattern → Examples of Matched Queries
Passive constructs:
H [VBN|VBD|VBG] [<anything>] M [<anything>] → (plants)$_H$ grown in (zone 7)$_M$ → (supreme court justices)$_H$ born in (new jersey)$_M$ → (medicinal plants)$_H$ used as (ayurvedic)$_M$ drugs → (manipulatives)$_H$ used in (elementary math)$_M$
Prepositional constructs:
H [IN|TO] [<anything>] M [<anything>] → (plants)$_H$ for (zone 7)$_M$ → (justices)$_H$ of the (california supreme court)$_M$ → (medicinal plants)$_H$ in (ayurvedic)$_M$ products → (math manipulatives)$_H$ for (elementary)$_M$ level
Relative pronoun constructs:
H [that|who|which] [<anything>] M [<anything>] → (plants)$_H$ that grow in (zone 7)$_M$ → (animals)$_H$ who live in (water)$_M$ → (medicinal plants)$_H$ that are used in (ayurveda)$_M$ → (math manipulatives)$_H$ that are taught in the (elementary)$_M$ classroom

Table 1: Extraction patterns matched against queries to identify candidate interpretations (H, M=head and modifier of a hypothetical noun phrase)

passive, prepositional or relative-pronoun construct, followed by another ngram M, and optionally followed by other tokens. The ngrams H and M contain one or more tokens. The patterns effectively split matching queries into four consecutive sequences of tokens Q=[Q_1 Q_2 Q_3 Q_4], where H and M correspond to Q_1 and Q_3, and Q_4 may be empty. For example, the pattern in the lower portion of Table 1 matches the query *"(plants)$_H$ that grow in (zone 7)$_M$"*, which is one of the queries shown in the upper portion of Figure 1.

Mapping Noun Phrases to Interpretations: Each noun phrase to interpret is split into all possible decompositions of two consecutive sequences of tokens N=[N_1 N_2], where the two sequences correspond to a hypothetical modifier and a hypothetical head of the noun phrase. For example, the noun phrase *"zone 7 plants"* is split into [*"zone"*, *"7 plants"*] and separately into [*"zone 7"*, *"plants"*]. If N_1 and Q_3, and N_2 and Q_1 respectively, match, then the matching query Q (e.g., *"(plants)$_H$ that grow in (zone 7)$_M$"*) is retained as a candidate interpretation of the noun phrase N (*"(zone 7)$_M$ (plants)$_H$"*), as shown in the middle portion of Figure 1.

Mapping via Modifier Variants: At its simplest, the matching of the hypothetical modifier relies on strict string matching. Alternatively, original modifiers in the noun phrases to interpret may be matched to queries via expansion variants. Variants are phrases that likely play the same role, and therefore share interpretations, as modifiers relative to the head in a noun phrase. Variants allow for the extraction of candidate interpretations that may otherwise not be available in the input data. For example, in Figure 1, the variant *new jersey* available for *california* allows for the matching of *california* in the noun phrase *"(california)$_M$ (supreme court justices)$_H$"*, with *new jersey* in the query *"(supreme court justices)$_H$ born in (new jersey)$_M$"*. The candidate interpretation *"(supreme court justices)$_H$ born in (california)$_M$"* is extracted for the noun phrase *"(california)$_M$ (supreme court justices)$_H$"*, even though the query *"supreme court justices born in california"* is not present among the input queries.

Possible sources of variants include distributionally similar phrases (Lin and Wu, 2009), where the phrases most similar to a modifier would act as its variants. Mappings from adjectival modifiers in noun phrases (e.g., *aquatic* in *"aquatic animals"* in Figure 1) into the nominal counterparts (e.g., *water*) that are likely to occur in interpretations (e.g., *"(animals)$_H$ who live in (water)$_M$"*) are also useful. Concretely, as described later in Section 3, variants are generated using WordNet (Fellbaum, 1998), distributional similarities and Wikipedia.

Aggregation of Candidate Interpretations: Candidate interpretations of a noun phrase are aggregated from source queries that matched the noun phrase. The frequency score of a candidate interpretation is the weighted sum of the frequencies of source queries from which the candidate interpretation is collected, possibly via variants of modifiers. In the weighted sum, the weights are similarity scores between the original modifier from the noun phrase, on one hand, and the variant from the source query into which the modifier was mapped, on the other hand. For example, in Figure 1, the frequency score of the candidate interpretation *"(plants)$_H$ that grow in (zone 7)$_M$"* for the noun phrase *"(zone 7)$_M$ (plants)$_H$"* is the weighted sum of the frequencies of the source queries *"plants that grow in zone 7"* and *"plants that grow in zone 11"*. The weights for the variants *zone 7* and *zone 11* relative to the original modifier *zone 7* may be 1.0 (identity) and 0.8 (distributional similarity), whereas the weights of adjectival modifiers such as *water* for *aquatic* may be 1.0. Separately from the frequency score, a penalty score is computed that penalizes interpretations containing extraneous tokens. Specifically, the penalty counts

the number of nouns or adjectives located outside the modifier and head. Candidate interpretations extracted for a noun phrase are ranked in increasing order of their penalty scores or, in case of ties, in decreasing order of their frequency scores.

3 Experimental Setting

Sources of Textual Data: The experiments rely on a random sample of around 1 billion fully-anonymized Web search queries in English. The sample is drawn from queries submitted to a general-purpose Web search engine. Each query is available independently from other queries, and is accompanied by its frequency of occurrence in the query logs.

Sources of Variants: The original form of the modifiers is denoted as **orig-phrase**. Three types of variant phrases are collected for the purpose of matching modifiers within noun phrases to interpret, with phrases from queries. Relations encoded as Value-Of, Related-Noun and Derivationally-Related relations in WordNet (Fellbaum, 1998) are the source of **adj-noun** variants. They map around 6,000 adjectives into one or more nouns (e.g., (*french*→*france*), (*electric*→*electricity*), (*aquatic*→*water*)). A repository of distributionally similar phrases, collected in advance (Lin and Wu, 2009) from a sample of around 200 million Web documents in English, is the source of **dist-sim** variants. For each of around 1 million phrases, the variants consist of their 50 most similar phrases (e.g., *art garfunkel*→{*carly simon*, *melissa manchester*, *aaron neville*, ..}).

A snapshot of all Wikipedia articles in English, as available in June 2014, is the source of **wiki-templ** variants. For each of around 50,000 phrases, their wiki-templ variants are collected from Wikipedia categories sharing a common parent Wikipedia category (e.g., *"albums by artist"*) and having a common head (*"art garfunkel albums"*, *"black sabbath albums"*, *"metallica albums"*). The different modifiers (*art garfunkel*, *black sabbath*, *metallica*) that accompany the shared head are collected as variants of one another. Among the four types of variants, wiki-templ variants are applied only when the noun phrase to interpret, and the source Wikipedia category names from which the variants were collected, have the same head. For example, *X=art garfunkel*→{*black sabbath*, *metallica*, *50 cent*, ..} is applied only in the context of the noun phrase *"X albums"*.

Vocabularies of Noun Phrases: The extraction method acquires interpretations from queries, for noun phrases from three vocabularies. **ListQ** is a set of phrases *X* (e.g., *"aramaic words"*) from queries in the form [*list of X*], where the frequency of the query [*X*] is at most 100 times higher than the frequency of the query [*list of X*], and the frequency of the latter is at least 5. **IsA** is a set of class labels (e.g., *"academy award nominees"*), originally extracted from Web documents via Hearst patterns (Hearst, 1992), and associated with at least 25 instances each (e.g., *zero dark thirty*). **WikiC** is a set of Wikipedia categories that contain some tokens in lowercase beyond prepositions and determiners, and whose heads are plural-form nouns (e.g., *"french fiction writers"*). Only phrases that are one of the full-length queries from the input set of Web search queries are retained in the respective sets, as vocabularies of noun phrases to interpret; other phrases are discarded.

Vocabulary	Relative Coverage			
	R	Q	I	I / Q
ListQ	406,249	406,249	277,193	0.682
IsA	613,148	405,262	282,927	0.698
WikiC	248,615	87,878	63,518	0.723

Table 2: Relative coverage of noun phrase interpretation, over noun phrases from various vocabularies (R=number of raw noun phrases; Q=subset of noun phrases from R that are queries; I=subset of noun phrases from Q with some extracted interpretation(s); I/Q=fraction of noun phrases from Q that are present in I)

Parameter Settings: The noun phrases to interpret and queries are both part-of-speech tagged (Brants, 2000). From among candidate interpretations extracted for a noun phrase, interpretations whose penalty score is higher than 1 are discarded. When computing the frequency score of a candidate interpretation as the weighted sum of the frequencies of source queries, the weights assigned to various variants are 1.0, for orig-phrase, adj-noun and wiki-templ variants; and the available distributional similarity scores within [0.0, 1.0], for dist-sim variants.

4 Evaluation Results

Relative Coverage: Because it is not feasible to manually compile the exhaustive sets of all string forms of valid interpretations of all (or many) noun phrases, we compute relative instead of absolute coverage. As illustrated in Table 2, some interpretations are extracted from queries for more than 500,000 of the noun phrases from all input vocabu-

Gold Set: Sample of Noun Phrases
ListQ: 1911 pistols, 2009 movies, alabama sororities, alaskan towns, american holidays, aramaic words, argumentative essays, arm loans, army ranks, .., yugioh movies
IsA: academy award nominees, addicting games, advanced weapons systems, android tablet, application layer protocols, astrological signs, automotive parts, .., zip code
WikiC: 2k sports games, aaliyah songs, advertising slogans, airline tickets, alan jackson songs, ancient romans, andrea bocelli albums, athletic shoes, .., wii accessories

Table 3: Gold sets of 100 noun phrases per vocabulary

laries, or around 70% of all input noun phrases.

Precision of Interpretations: From an input vocabulary, an initial weighted sample of 150 noun phrases with some extracted interpretations is manually inspected. The sampling weight is the frequency of the noun phrases as queries. A noun phrase from the selected sample is either retained, or discarded if deemed to be a non-interpretable phrase. A noun phrase is not interpretable if it is in fact an instance (*"new york"*, *"alicia keys"*) rather than a class; or it is not a properly formed noun phrase (*"watch movies"*); or does not refer to a meaningful class (*"3 significant figures"*). The manual inspection ends, once a sample of 100 noun phrases has been retained. The procedure gives weighted random samples of 100 noun phrases, drawn from each of the ListQ, IsA and WikiC vocabularies. The samples, shown in Table 3, constitute the gold sets of phrases ListQ, IsA and WikiC, over which precision of interpretations is computed. Note that, since the samples are random, Wikipedia categories that contribute to the automatic construction of wiki-templ variants may be selected as gold phrases in WikiC. This is the case for three of the gold phrases in WikiC.

The top 20 interpretations extracted for each gold phrase are manually annotated with correctness labels. As shown in Table 4, an interpretation is annotated as: correct and generic, or correct and specific, if relevant; okay, if useful but containing non-essential information; or wrong. To compute the precision score over a gold set of phrases, the correctness labels are converted to numeric values. Precision of a ranked list of extracted interpretations is the average of the correctness values of the interpretations in the list.

Table 5 provides a comparison of precision scores at various ranks in the extracted lists of interpretations, as an average over all phrases from a gold set.

Label (Score) → Examples of (Noun Phrase: Interpretation)
cg (1.0) → (good short stories: short stories that are good), (bay area counties: counties in the bay area), (fourth grade sight words: sight words in fourth grade), (army ranks: ranks from the army), (who wants to be a millionaire winners: winners of who wants to be a millionaire), (us visa: visa for us)
cs (1.0) → (brazilian dances: dances of the brazilian culture), (tsunami charities: charities that gave to the tsunami), (stephen king books: books published by stephen king), (florida insurance companies: insurance companies headquartered in florida), (florida insurance companies: insurance companies operating in florida), (us visa: visa required to enter us)
qq (0.5) → (super smash bros brawl characters: characters meant to be in super smash bros brawl), (carribean islands: islands by the carribean), (pain assessment tool: tool for recording pain assessment)
xw (0.0) → (periodic functions: functions of periodic distributions), (simpsons episodes: episodes left of simpsons), (atm card: card left in wachovia atm)

Table 4: Examples of interpretations manually annotated with each correctness label (cg=correct generic; cs=correct specific; qq=okay; xw=incorrect)

Gold Set	Precision@N				
	@1	@3	@5	@10	@20
ListQ	0.770	0.708	0.655	0.568	0.465
IsA	0.730	0.598	0.530	0.423	0.329
WikiC	0.780	0.647	0.561	0.455	0.357

Table 5: Average precision, at various ranks in the ranked lists of interpretations extracted for noun phrases from various sets of gold phrases

At ranks 1, 5 and 20, precision scores vary between 0.770, 0.655 and 0.465 respectively, for the ListQ gold set; and between 0.730, 0.530 and 0.329 respectively, for the IsA gold set.

Presence of Relevant Interpretations: Sometimes it is difficult to even manually enumerate as many as 20 distinct, relevant string forms of interpretations for a given noun phrase. Measuring precision at a particular rank (e.g., 20) in a ranked list of interpretations may be too conservative. Table 6 summarizes a different type of scoring metric, namely the presence of any relevant interpretation, among the interpretations extracted up to a particular rank. Relevance is flexibly defined, by requiring the interpretations to have been assigned a certain correctness label, then computing the average number of gold phrases for which such interpretations are present up to a particular rank. When considering only interpretations annotated as correct and generic or correct and specific, in the second row of each vertical

Gold Set	Selected Correctness Labels			Average presence of any interpretations with any of the selected correctness labels				
	cg	cs	qq	@1	@3	@5	@10	@20
ListQ	√	√	√	0.790	0.860	0.870	0.880	0.880
	√	√		0.750	0.810	0.830	0.840	0.840
	√			0.720	0.780	0.790	0.800	0.800
		√		0.030	0.160	0.360	0.450	0.460
IsA	√	√	√	0.750	0.800	0.810	0.830	0.830
	√	√		0.710	0.770	0.790	0.800	0.800
	√			0.650	0.690	0.710	0.710	0.720
		√		0.060	0.220	0.350	0.480	0.520
WikiC	√	√	√	0.810	0.900	0.920	0.930	0.930
	√	√		0.750	0.830	0.860	0.860	0.870
	√			0.640	0.730	0.750	0.750	0.760
		√		0.110	0.210	0.370	0.520	0.560

Table 6: Average of scores indicating the presence or absence of any interpretations annotated with a correctness label from a particular subset of correctness labels. Computed over interpretations extracted up to various ranks in the ranked lists of extracted interpretations (cg=correct generic; cs=correct specific; qq=okay)

Noun Phrase → Multiple-Bracketing Interpretations
african american women writers → (writers)$_H$ who wrote about (african american)$_M$ women, (women writers)$_H$ who are (african american)$_M$, (writers)$_H$ who cover (african american)$_M$ women struggles
chinese traditional instruments → (traditional instruments)$_H$ of (china)$_M$, (instruments)$_H$ used in (chinese traditional)$_M$ music
elementary math manipulatives → (manipulatives)$_H$ for (elementary math)$_M$, (math manipulatives)$_H$ in the (elementary)$_M$ classroom, (manipulatives)$_H$ used in (elementary math)$_M$, (math manipulatives)$_H$ for (elementary)$_M$ level
global corporate tax rates → (corporate tax rates)$_H$ around the (world)$_M$, (tax rates)$_H$ on (global corporate)$_M$ profits

Table 7: Sample of noun phrases from the ListQ gold set, whose top 10 extracted interpretations induce multiple pairs of a head and a modifier of the noun phrases (H=head; M=modifier)

portion in Table 6, the scores at rank 5 are 0.830 for ListQ, 0.790 for IsA and 0.860 for WikiC. Alternatively, in the fourth rows of each vertical portion, the scores at rank 5 are 0.360, 0.350 and 0.370 respectively. The scores indicate that at least one of the top 5 interpretations is correct and specific for about a third of the noun phrases in the gold sets.

Induced Modifiers, Heads and Interpretations: When a candidate interpretation is extracted for a noun phrase, the interpretation effectively induces a particular bracketing over the noun phrase, as it splits it into a modifier and a head. For an ambiguous

Presence of Multiple Bracketings			
Vocabulary	ListQ	IsA	WikiC
Fraction of Noun Phrases	0.110	0.124	0.051

Table 8: Fraction of noun phrases that have some extracted interpretation(s) and contain at least 3 tokens, whose interpretations induce multiple (rather than single) bracketings over interpreted noun phrases. The presence of multiple bracketings for a noun phrase is equivalent to the presence of multiple pairs of a head and a modifier, as induced by the top 10 interpretations extracted for the noun phrase

Noun Phrase → Extracted Interpretations
beatles songs → (songs)$_H$ sung by the (beatles)$_M$, (songs)$_H$ about the (beatles)$_M$
company accounts → (accounts)$_H$ maintained by the (company)$_M$, (accounts)$_H$ owed to a (company)$_M$
florida insurance companies → (insurance companies)$_H$ headquartered in (florida)$_M$, (insurance companies)$_H$ insuring in (florida)$_M$
german food → (food)$_H$ eaten in (germany)$_M$, (food)$_H$ produced in (germany)$_M$, (food)$_H$ that originated in (germany)$_M$
math skills → (skills)$_H$ needed for (math)$_M$, (skills)$_H$ learned in (math)$_M$, (skills)$_H$ gained from studying (math)$_M$
michael jackson song → (song)$_H$ written by (michael jackson)$_M$, (song)$_H$ sung by (michael jackson)$_M$, (song)$_H$ about (michael jackson)$_M$

Table 9: Sample of alternative relevant interpretations extracted among the top 20 interpretations for noun phrases from the ListQ gold set (H=head; M=modifier)

noun phrase, multiple bracketings may be possible, each corresponding to a different interpretation. Interpretations extracted from queries do capture such multiple bracketings, even for phrases from the gold sets, as illustrated in Table 7. Over all noun phrases from the input vocabularies that have some extracted interpretations and contain at least 3 tokens, about 10% (ListQ and IsA) and 5% (WikiC) of the noun phrases have multiple bracketings induced by their top 10 interpretations, as shown in Table 8.

Table 9 shows examples of noun phrases with multiple extracted interpretations that induce identical bracketings, but capture distinct interpretations.

Impact of Variants: Variants of modifiers provide alternatives in extracting candidate interpretations, even when the modifiers from the noun phrases are not present in their original form in the interpretations. For example, the adj-noun variant *ethiopia* of the modifier *ethiopian* leads to the extraction of the interpretation *"runners from ethiopia"* for the noun phrase *"ethiopian runners"*. Similarly, wiki-

Vocab	Variant Type			
	orig-phrase	adj-noun	dist-sim	wiki-templ
Interpretations produced by variant type (not exclusive):				
ListQ	0.453	0.089	0.642	0.017
IsA	0.389	0.121	0.597	0.003
WikiC	0.191	0.097	0.603	0.425
Interpretations produced only by variant type (exclusive):				
ListQ	0.281	0.069	0.450	0.005
IsA	0.299	0.099	0.491	0.001
WikiC	0.086	0.076	0.351	0.225

Table 10: Impact of various types of variants of modifiers, on the coverage of noun phrase interpretations. Computed as the fraction of the top 10 extracted interpretations produced by a particular variant type, and possibly by other variant types (upper portion); or produced only by a particular variant type, and by no other variant types (lower portion) (Vocab=vocabulary of noun phrases)

templ variants *metallica* and *50 cent* of the modifier *art garfunkel*, in the context *"X albums"*, allow for the extraction of the interpretation *"albums sold by art garfunkel"* for the noun phrase *"art garfunkel albums"*, via the interpretations *"albums sold by metallica"* and *"albums sold by 50 cent"*.

Table 10 quantifies the impact of various types of variants, on the coverage of noun phrase interpretations. The scores provided for each variant type correspond to either non-exclusive (upper portion of the table) or exclusive (lower portion) contribution of that variant type towards some extracted interpretations. In other words, in the lower portion, the scores capture the fraction of the top 10 interpretations that are produced only by that particular variant type. Three conclusions can be drawn from the results. First, all variant types contribute to increasing coverage, relative to using only orig-phrase variants. Second, dist-sim variants have a particularly strong impact. Third, wiki-templ variants have a strong impact, but only when the contexts from which they were collected match the context of the noun phrase being interpreted. On the WikiC vocabulary in the lower portion of Table 10, the scores for wiki-templ illustrate the potential that contextual variants have in extracting additional interpretations.

Table 11 again quantifies the impact of variant types, but this time on the coverage and, more importantly, accuracy of interpretations extracted for phrases from the gold sets. The scores are computed over the ranked lists of interpretations from the ListQ gold set, as certain types of variants are temporarily disabled in ablation experiments. The upper portion of the table shows results when only

Variant Types				Impact on Precision		
O	A	D	W	Cvg	P@5	C@5
√	-	-	-	74	0.433	0.581
-	√	-	-	16	0.474	0.562
-	-	√	-	66	0.478	0.651
-	-	-	√	2	0.166	0.500
-	√	√	√	73	0.484	0.657
√	-	√	√	97	0.641	0.835
√	√	-	√	83	0.448	0.590
√	√	√	-	99	0.649	0.828
√	-	-	-	74	0.433	0.581
√	√	-	-	81	0.453	0.592
√	-	√	-	96	0.635	0.833
√	-	-	√	76	0.429	0.578
√	√	√	√	100	0.655	0.830

Table 11: Impact of various types of variants of modifiers, on the precision of noun phrase interpretations. Computed over the ListQ gold set, at rank 5 in the ranked lists of extracted interpretations, when various variant types are allowed (√) or temporarily not allowed (-) to produce interpretations (O=orig-phrase variant type; A=adj-noun variant type; D=dist-sim variant type; W=wiki-templ variant type; Cvg=number of noun phrases from the gold set with some interpretation(s) produced by the allowed variant types; P@5=precision at rank 5; C@5=average presence of any interpretations annotated as correct and generic (cg) or correct and specific (cs), among interpretations up to rank 5)

one of the variant types is enabled. It shows that none of the variant types, taken in isolation, can match what they achieve when combined together, in terms of both coverage and accuracy. The middle portion of the table shows results when all but one of the variant types are enabled. Each of the variant types incrementally contributes to higher coverage and accuracy over the combination of the other variant types. The incremental contribution of wiki-templ variants is the smallest. The lower portion of Table 11 gives the incremental contribution of the variant types, relative to using only the orig-phrase variant type. The last row of Table 11 corresponds to all variant types being enabled.

Discussion: Independently of the choice of the textual data source (e.g., documents, queries) from which interpretations are extracted, a noun phrase is intuitively more difficult to interpret if it is relatively more rare or more complex (i.e., longer). Additional experiments quantify the effect, by measuring the correlation between the presence of some extracted interpretations for an input noun phrase, on one hand; and the frequency of the noun phrase as a query (in Table 12), on the other hand. In Table 12,

Vocabulary	Noun Phrases With : Without Interpretation(s)		
	I : ¬I	A_I : $A_{\neg I}$	M_I : $M_{\neg I}$
ListQ	2.14 : 1	2.93 : 1	2.65 : 1
IsA	2.31 : 1	5.76 : 1	3.26 : 1
WikiC	2.60 : 1	3.72 : 1	3.63 : 1

Table 12: Correlation between coverage, measured as the presence of some extracted interpretation(s) for a noun phrase, on one hand; and frequency of the noun phrase as a query, on the other hand (I=number of noun phrases that are queries and have some extracted interpretation(s); ¬I=number of noun phrases that are queries and do not have any extracted interpretation(s); A=average query frequency of noun phrases as queries; M=median query frequency of noun phrases as queries)

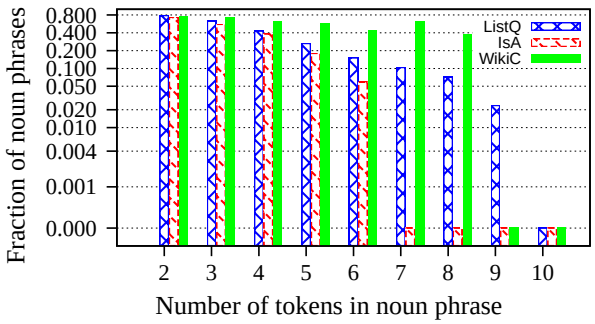

Figure 2: Ability to extract interpretations for noun phrases, as a function of the length of noun phrases. Computed as the fraction of noun phrases from an input vocabulary with a particular number of tokens, for which there are some extracted interpretation(s)

the effect is visible in that query frequency is higher for noun phrases with some extracted interpretations vs. noun phrases with none. For example, the average query frequency is almost three times higher for the former than for the latter, for the ListQ vocabulary. Similarly, in Figure 2, a larger fraction of the input noun phrases with a particular number of tokens have some extracted interpretations, when the number of tokens is lower rather than higher. The effect is somewhat less pronounced for, but still applicable to, the WikiC vocabulary, with some extracted interpretations being present for 75%, 71%, 63%, and 37% of the noun phrases containing 2, 3, 4 and 8 tokens respectively. That a larger fraction of the longer noun phrases can be interpreted in the WikiC vocabulary is attributed to the role of wiki-templ variants in extracting interpretations that would otherwise not be available.

Interpretations from Queries vs. Documents: For completeness, additional experiments evaluate the interpretations extracted from queries, relative to a gold standard introduced in (Hendrickx et al., 2013). The gold standard consists of a gold set of 181 compound noun phrases (e.g., *"accounting principle"* and *"application software"*), their manually-assembled gold paraphrases (e.g., *"principle of accounting"*, *"software to make applications"*), and associated scoring metrics referred to as non-isomorphic and isomorphic. Note that, in comparison to the ListQ, IsA and WikiC evaluation sets, the gold standard in (Hendrickx et al., 2013) may contain relatively less popular gold phrases. As many as 45 gold paraphrases are available per gold phrase on average. They illustrate the difficulty of any attempt to manually assemble exhaustive sets of all strings that are valid interpretations of a noun phrase. For example, the gold paraphrases of the gold phrase *blood cell* include *"cell that is found in the blood"*, but not the arguably equally-relevant *"cell found in the blood"*. In addition, more than one human annotators independently provide the same gold paraphrase for only a tenth of all gold paraphrases. See (Hendrickx et al., 2013) for details on the gold standard and scoring metrics. The gold set is added as another input vocabulary to the method proposed here. After inspection of a training set of compound noun phrases also introduced in (Hendrickx et al., 2013), the parameter settings are modified to only retain interpretations whose penalty score is 0.

The isomorphic and non-isomorphic scores reward coverage and accuracy respectively. For the ranked candidate interpretations extracted from queries for the gold set, they are 0.037 and 0.556 respectively. In comparison to previous methods that operate over documents instead of queries, the isomorphic score is much lower for our method (e.g., 0.037 vs. 0.130 (Van de Cruys et al., 2013)). It suggests that queries cannot reliably provide an exhaustive list of all possible strings available in the gold standard for each gold phrase. However, the non-isomorphic score is higher for our method than for the best method operating over documents (i.e., 0.556 vs. 0.548 (Hendrickx et al., 2013)). In fact, the non-isomorphic score using queries would be 0.745 instead of 0.556, if it were computed over only the 135 gold noun phrases with some extracted interpretations. The results suggests that the method proposed here extracts more accurate interpretations from queries, than previous methods extract from

documents. Higher accuracy is preferable in scenarios like Web search, where it is important to accurately trigger structured results.

Error Analysis: The relative looseness of the extraction patterns applied to queries causes interpretations containing undesirable tokens to be extracted. In addition, part-of-speech tagging errors lead to interpretations receiving artificially low penalty scores, and therefore being considered to be of higher quality than they should be. For example, *phd* in the interpretation *"job for phd in chemistry"* is incorrectly tagged as a past participle verb. As a result, the computed penalty score is too low.

Occasionally, the presence of additional tokens within an interpretation is harmless (e.g., *"issues of controversy in society"* for *"controversial issues"*, *"foods allowed on a high protein low carb diet"* for *"high protein low carb foods"*), if not necessary (e.g., *"dances with brazilian origin"* for *"brazilian dances"*, *"artists of the surrealist movement"* for *"surrealist artists"*, *"options with weekly expirations"* for *"weekly options"*). But often it leads to incorrect interpretations (e.g., *"towns of alaska map"* for *"alaska towns"*, *"processes in chemical vision"* for *"chemical processes"*).

Variants of modifiers occasionally lead to incorrect interpretations for a noun phrase, even if the interpretations may be correct for the individual variants. The phenomenon is an instance of semantic drift, wherein variants do share many properties but still diverge in others. Examples are *"words that are bleeped similarly"* extracted for *"bleeped words"* via the variant *bleeped→spelled*. Separately, linguistic constructs that negate or at least alter the desired meaning affect the understanding of text in general and also affect the extraction of interpretations in particular. Examples are *"heaters with no electricity"* for *"electric heaters"*, and *"animal that used to be endangered"* for *"endangered animal"*.

5 Related Work

Relevant interpretations extracted from queries act as a potential bridge between facts, on one hand, and class labels, on the other hand, available for instances. The former might be inferred from the latter and vice versa. There are two previous studies that are relevant to the task of extracting facts from existing noun phrases. First, (Yahya et al., 2014) extract facts for attributes of instances, without requiring the presence of the verbal predicates usually employed (Fader et al., 2011) in open-domain information extraction. Second, in (Nastase and Strube, 2008), relations encoded implicitly within Wikipedia categories are converted into explicit relations. As an example, the relation <*deconstructing harry*, *directed*, *woody allen*> is obtained from the fact that *deconstructing harry* is listed under *"movies directed by woody allen"* in Wikipedia. The method in (Nastase and Strube, 2008) relies on manually-compiled knowledge, and does not attempt to interpret compound noun phrases.

Since relevant interpretations paraphrase the noun phrases which they interpret, a related area of research is paraphrase acquisition (Madnani and Dorr, 2010; Ganitkevitch et al., 2013). Previous methods for the acquisition of paraphrases of compound noun phrases (Kim and Nakov, 2011; Van de Cruys et al., 2013) operate over documents, and may rely on text analysis tools including syntactic parsing (Nakov and Hearst, 2013). In contrast, the method proposed here extracts interpretations from queries, and applies part of speech tagging. Queries were used as a textual data source in other tasks in open-domain information extraction (Jain and Pennacchiotti, 2010; Pantel et al., 2012).

6 Conclusion

Interpretations extracted from queries explain the roles that modifiers play within longer noun phrases. Current work explores the interpretation of noun phrases containing multiple modifiers (e.g., *"(french)$_{M1}$ (healthcare)$_{M2}$ (companies)$_{H}$"* by separately interpreting *"(french)$_{M1}$ (companies)$_{H}$"* and *"(healthcare)$_{M2}$ (companies)$_{H}$"*); the grouping of lexically different but semantically equivalent interpretations (e.g., *"dances of brazilian origin"*, *"dances from brazil"*); the collection of more variants from Wikipedia and other resources; the incorporation of variants of heads (*physicists→scientists* for interpreting the phrase *"belgian physicists"*), which likely need to be more conservatively applied than for modifiers; and the use of query sessions, as an alternative to sets of disjoint queries.

Acknowledgments

The paper benefits from comments from Jutta Degener, Mihai Surdeanu and Susanne Riehemann. Data extracted by Haixun Wang and Jian Li is the source of the IsA vocabulary of noun phrases used in the evaluation.

References

M. Banko, Michael J Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. 2007. Open information extraction from the Web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, pages 2670–2676, Hyderabad, India.

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.

B. Dalvi, W. Cohen, and J. Callan. 2012. Websets: Extracting sets of entities from the Web using unsupervised information extraction. In *Proceedings of the 5th ACM Conference on Web Search and Data Mining (WSDM-12)*, pages 243–252, Seattle, Washington.

P. Downing. 1977. On the creation and use of English compound nouns. *Language*, 53:810–842.

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)*, pages 1535–1545, Edinburgh, Scotland.

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

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)*, pages 945–955, Baltimore, Maryland.

J. Ganitkevitch, B. Van Durme, and C. Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Association for Computational Linguistics (NAACL-HLT-13)*, pages 758–764, Atlanta, Georgia.

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.

I. Hendrickx, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Szpakowicz, and T. Veale. 2013. SemEval-2013 task 4: Free paraphrases of noun compounds. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval-14)*, pages 138–143, Atlanta, Georgia.

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*, 194:28–61.

A. Jain and M. Pennacchiotti. 2010. Open entity extraction from Web search query logs. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING-10)*, pages 510–518, Beijing, China.

N. Kim and P. Nakov. 2011. Large-scale noun compound interpretation using bootstrapping and the Web as a corpus. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP-11)*, pages 648–658, Edinburgh, Scotland.

Z. Kozareva and E. Hovy. 2010. A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP-10)*, pages 1110–1118, Cambridge, Massachusetts.

D. Lin and X. Wu. 2009. Phrase clustering for discriminative learning. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP-09)*, pages 1030–1038, Singapore.

N. Madnani and B. Dorr. 2010. Generating phrasal and sentential paraphrases: a survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

P. Nakov and M. Hearst. 2013. Semantic interpretation of noun compounds using verbal and other paraphrases. *ACM Transactions on Speech and Language Processing*, 10(3):1–51.

V. Nastase and M. Strube. 2008. Decoding Wikipedia categories for knowledge acquisition. In *Proceedings of the 23rd National Conference on Artificial Intelligence (AAAI-08)*, pages 1219–1224, Chicago, Illinois.

P. Pantel, T. Lin, and M. Gamon. 2012. Mining entity types from query logs via user intent modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL-12)*, pages 563–571, Jeju Island, Korea.

T. Van de Cruys, S. Afantenos, and P. Muller. 2013. MELODI: A supervised distributional approach for free paraphrasing of noun compounds. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval-14)*, pages 144–147, Atlanta, Georgia.

R. Wang and W. Cohen. 2009. Automatic set instance extraction using the Web. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP-09)*, pages 441–449, Singapore.

M. Yahya, S. Whang, R. Gupta, and A. Halevy. 2014. ReNoun: Fact extraction for nominal attributes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP-14)*, pages 325–335, Doha, Qatar.

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)*, pages 956–966, Baltimore, Maryland.