# Attribute Extraction from Conjectural Queries

Marius Paşca
Google Inc.
Mountain View, California
mars@google.com

ABSTRACT
Conjectural search queries (*is python case sensitive, is millennium stadium heated*) embody attempts by Web users to verify whether a particular property (*soluble in water?, case sensitive?, heated?*) does or does not apply to a particular instance (*iodine, python, millennium stadium*). This paper considers such queries to be a data source of attributes of open-domain classes. Conjectural attributes complement attributes encoded in human-compiled knowledge resources or automatically acquired from text by previous methods. They correspond to properties of interest to Web users, which are not necessarily stated in nominal form. Relevant properties of *Chemical elements*, *Programming languages* and *Stadiums* include whether they are *soluble in water*, *flammable* or *ductile*; *case sensitive, platform independent*, or *interpreted*; or *air conditioned, roof retractable* or *heated*, respectively. Experimental results show that relevant, conjectural attributes can be extracted from inherently-noisy queries, for a variety of open-domain classes of interest.

KEYWORDS: Class attributes, open-domain information extraction, Web search queries.

*Proceedings of COLING 2012: Technical Papers*, pages 2177–2190,
COLING 2012, Mumbai, December 2012.

| Examples of Attributes Available in or Extracted from Various Sources, for a Sample of Classes |
|---|
| Food ingredients: |
| W: energy, dietary fiber, solubility in water |
| F: energy per 100g, availability, scientific name, solubility in water |
| $D_1$: species, pounds, cup, kinds, lbs, bowl |
| $D_2$: quality, part, taste, value, portion |
| $Q_1$: nutritional value, health benefits, glycemic index, varieties, calories |
| $Q_C$: gluten free?, safe?, healthy?, vegan?, halal?, fattening?, acidic?, good for skin? |
| Astronomical objects: |
| W: constellation, right ascension, spectral type, rotational velocity, orbital period, mean radius |
| F: category, constellation, age, periapsis, orbital period, mean radius |
| $D_1$: observations, spectrum, planet, spectra, conjunction, transit, temple, surface |
| $D_2$: surface, orbit, bars, history, atmosphere |
| $Q_1$: atmosphere, surface, gravity, diameter, mass, rotation, revolution, moons, radius |
| $Q_C$: bigger than earth?, close to the sun?, circumpolar?, capable of supporting life? |
| Religions: |
| W: fundamentals, texts, deities, sacred sites, schools, people |
| F: founding figures, beliefs, practices, texts, deities, sacred sites |
| $D_1$: teachings, practice, beliefs, religion spread, principles, emergence, doctrines |
| $D_2$: basis, influence, name, truths, symbols, principles, strength, practice, origin, god, defence |
| $Q_1$: basic beliefs, teachings, holy book, practices, rise, branches, spread, sects |
| $Q_C$: monotheistic?, a religion or a way of life?, peaceful?, older than hinduism? |

Table 1: Examples of attributes already explicitly encoded in human-compiled knowledge resources (W=Wikipedia; F=Freebase) or extracted by various methods from text. Some of the entries in the table are also listed in (Van Durme et al., 2008) ($D_1$=from documents (Paşca et al., 2007), $D_2$=from documents (Van Durme et al., 2008), $Q_1$=from queries (Paşca, 2007), $Q_C$=from conjectural queries (this method))

## 1 Introduction

**Motivation**: Current efforts towards injecting structured knowledge into search results place a renewed emphasis on extracting, curating and serving open-domain knowledge. Resources such as Wikipedia (Remy, 2002) and Freebase (Bollacker et al., 2008) contain knowledge about classes (*Chemical elements*, *Programming languages* and *Stadiums*) and their instances (*iodine*, *python*, *millennium stadium*). Knowledge is often represented as properties or attributes of the instances, along with values for those properties. But even the largest human-curated knowledge resources may be missing at least some relevant knowledge, for some or all instances. For example, attributes representing the geographical coordinate or seating capacity are available in both Wikipedia and Freebase for many (e.g., for *millennium stadium*, *bc place*, *georgia dome*), albeit not all (e.g., *jenner park stadium*) instances of *Stadiums*. In contrast, information on whether particular *Stadiums* are *heated* or *roof retractable* is uniformly missing. Search queries such as *"is millennium stadium heated"*, *"is bc place heated"* and *"is the georgia dome heated"* suggest that such information is relevant to Web users. Populating the knowledge resource with the attribute *heated?*, for instances of *Stadiums*, would likely be more useful than, e.g., adding knowledge about whether they are *painted white*. More generally, identifying properties or attributes of interest to Web users but missing from the knowledge resource, helps allocate limited resource-development cycles to areas within the resource where their addition is most beneficial.

**Contributions**: This paper introduces a method for the acquisition of class attributes, from queries in the form *"<be> I A"* (e.g., *"is (georgia dome)$_I$ (heated)$_A$"*) or *"why <be> I A"* (e.g., *"why was the (tiger stadium)$_I$ (demolished)$_A$"*). Through such queries, Web users likely attempt to

verify whether - or why - a particular property (*demolished?*, *heated?*) applies to a particular instance (*tiger stadium*, *georgia dome*). In Table 1, attributes from conjectural queries are different in scope and style from attributes already encoded in human-compiled knowledge resources or attributes automatically acquired by previous methods. Examples include *good for skin?* for *Food ingredients*, *capable of supporting life?* for *Astronomical objects*, or *peaceful?* for Religions. Conjectural attributes cannot be easily captured by, and are therefore complementary to, attributes produced by previous methods, including methods targeting textual fragments in the form *"A of I"* (e.g., *"(seating capacity)$_A$ of (millennium stadium)$_I$"*) occurring in documents (Tokunaga et al., 2005) or queries (Paşca and Van Durme, 2007).

## 2 Extraction of Conjectural Attributes

**Intuitions**: The extraction of attributes from queries starts from the intuition that, if an attribute $A$ is relevant for a class $C$, then users are likely to ask for the value of the attribute $A$, for various instances $I$ of the class $C$ (Paşca, 2007). The submission of fact-seeking queries such as *"what is the (seating capacity)$_A$ of (millennium stadium)$_I$"*, or the more compact *"(seating capacity)$_A$ of (millennium stadium)$_I$"*, is taken as evidence that *seating capacity* is a candidate attribute of the instance *Millennium Stadium*, and transitively a candidate attribute of the class *(Stadiums)$_C$* to which *Millennium Stadium* belongs. Attributes extracted from such queries are usually limited to noun phrases.

If an attribute $A$ is relevant for a class $C$, then users are also likely to ask whether the attribute $A$ does or does not apply to various instances $I$ of the class $C$. The method introduced here takes advantage of other kinds of queries, namely conjectural, or truth-verification queries. Conjectural queries ask whether something is true or not. In comparison to fact-seeking queries, conjectural queries provide only weak evidence (conjectures) that the fact being asked about is true or not. Nevertheless, conjectural queries such as *"is (millennium stadium)$_I$ (heated)$_A$"* are taken as weak evidence that *heated?* is a candidate attribute of the instance *Millennium Stadium*, and transitively as stronger evidence a candidate attribute of the class *(Stadiums)$_C$*. Intuitively, the evidence supporting the candidate attribute is stronger for the class than it is for the instance. Users are likely to ask whether an attribute does or does not apply to an instance, based on prior knowledge that the attribute already applies to the class. If users submit the query mentioned earlier in this paragraph, it is likely because they are aware that *Stadiums* may or may not be *heated*, and would like to check whether a particular instance of *Stadiums* is. Collectively, queries asking whether an attribute applies to multiple instances of the same class are indirect evidence that the attribute does in fact apply to the class.

In addition to conjectural queries, the method also takes advantage of explanation-seeking queries, as a less frequent but more reliable source of evidence available in queries. Such queries ask for an explanation of why something is true. Queries such as *"why is (millennium stadium)$_I$ (heated)$_A$"* are intuitively more reliable sources of evidence that *Millennium Stadium*, in particular, and *Stadiums*, in general, are in fact *heated*, than queries like *"is (millennium stadium)$_I$ (heated)$_A$"* are.

**Scope**: Attributes extracted from conjectural queries cover multiple parts of speech, from single-word adjectives (*heated?*) and multiple-word descriptors (*open to the public?*) to noun phrases (*a retractable roof?*, *a 5 star stadium?*). On another dimension, conjectural attributes can capture properties that are objective or subjective. They include the more objective *on netflix?*, *a true story?*, *rated r?* for *Films*; but also the more subjective properties of whether *Films* are *weird?* or *funny?*. Anecdotal evidence of query frequency distribution in query logs suggests that Web users

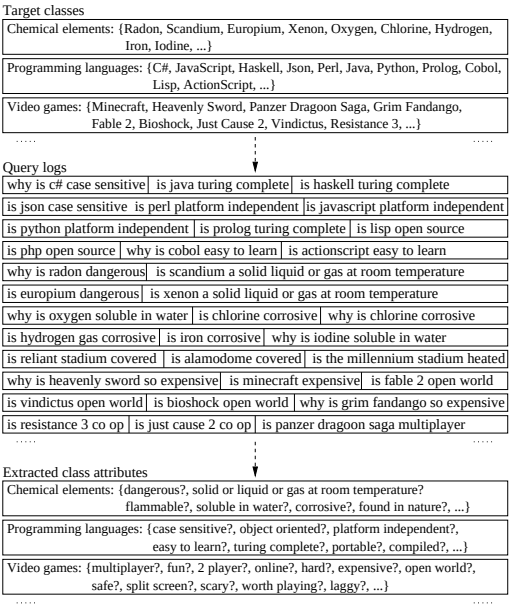


Figure 1: Overview of extraction of conjectural attributes from Web search queries

are sometimes as interested in subjective attributes as they are in objective ones.

Referring back to Table 1, at least some of the conjectural adjectives have a nominal counterpart, which could in principle be extracted by existing extraction methods. The attribute *bigger than earth?* may be thought of as roughly equivalent to *size relative to earth*; *close to the sun?* to *distance to the sun*; *capable of supporting life* to *ability to support life*. However, even large query logs may not contain such noun phrases in the query format expected by previous methods. Moreover, some of the theoretical nominal counterparts of some conjectural attributes would be difficult to even identify manually, let alone extract with previous methods. It is less clear what the nominal counterparts are in the case of *fun?* for *Actors*, *peaceful?* for *Religions*, or *waterproof?* for *Cameras*. To summarize, conjectural attributes are arguably more diverse than attributes from existing resources or previously extracted from text.

**Extraction from Queries**: As illustrated in Figure 1, the proposed method takes as input a set of target classes, each of which is a set of instances; and a set of anonymized queries. Instances may be available as non-disambiguated items, that is, as strings (*python*) whose meaning is otherwise not available; or as disambiguated items, that is, as pointers to knowledge base entries with a particular, disambiguated meaning (*Python (programming language)*).

The method identifies the subset of input queries that are deemed to be conjectural queries. For this purpose, queries are matched against the patterns "*<be> I A*" and "*why <be> I A*", where $I$ and $A$ are a possible instance and an attribute. Other simple patterns were employed in previous

| |
|---|
| Actors, Aircraft, Animated characters, Association football clubs, Astronomical objects, Automobiles, Awards, Battles and operations of World War II, Chemical elements, Cities, Companies, Countries, Currencies by country, Digital cameras, Diseases and disorders, Drugs, Empires, Films, Flowers, Food ingredients, Holidays, Hurricanes in North America, Internet search engines, Mobile phones, Mountains, National Basketball Association teams, National parks, Newspapers, Organizations designated as terrorist, Painters, Programming languages, Religious faiths traditions and movements, Rivers, Skyscrapers, Sports events, Stadiums, Treaties, Universities and colleges, Video games, Wine |

Table 2: Set of 40 Wikipedia categories used as target classes in the evaluation of attributes

work (Paşca and Van Durme, 2007) to extract attributes from text. Queries that match any of the patterns are deemed to be conjectural, but only if the query fragment corresponding to $I$ can be matched to one of the instances of the input target classes. Queries do not need to end in a question mark, in order to be deemed conjectural. Depending on whether instances are non-disambiguated or disambiguated items, the matching from a query fragment to an instance from the input classes consists in either simple string matching; or disambiguation of $I$ in the context of the query. When matching succeeds, the query fragment $A$ is collected as a conjectural attribute for the instance $I$. In turn, attributes of a class are aggregated from attributes of individual instances of the class. The relative ranking among attributes of a class promotes attributes extracted from source queries that a) match many instances of the class, and few instances of any other classes; and b) have higher frequency in query logs.

## 3 Experimental Setting

**Textual Data Sources**: The experiments rely on a random sample of around 500 million fully-anonymized Web search queries in English. Each query is available independently from other queries, and is accompanied by its frequency of occurrence in the query logs.

**Target Classes**: The attributes extracted from queries are evaluated over a set of 40 target classes included in Table 2. In an effort to reuse experimental setup proposed in previous work, each of the 40 manually-compiled classes introduced in (Paşca, 2007) is mapped into the Wikipedia category that best matches it. For example, the evaluation classes *Actor*, *Mountain*, *Movie*, *Religion* and *TerroristGroup* from (Paşca, 2007) are mapped into the Wikipedia categories *Actors*, *Mountains*, *Films*, *Religious faiths traditions and movements* and *Organizations designated as terrorist* respectively. Note that the name of the Wikipedia category only serves as a convenience label for its target class, and is not otherwise exploited in any way during the evaluation. Instead, a target class consists in a sample set of Wikipedia articles selected from all articles listed under the respective category in Wikipedia, or listed under sub-categories of the respective category. For example, the target class *Automobiles* includes the Wikipedia articles titled *Chevrolet Tahoe*, *Jaguar XJ220* etc., as illustrated in Table 3. Due to noise within Wikipedia categories, spurious instances such as *Vibrating alert* (for *Mobile phones*) or *Veljko Rus* (for *Universities and colleges*) are occasionally present in the target classes, which makes the evaluation set more realistic than artificially clean. The resulting set of 40 target classes contains an average of 14,974 instances per class.

**Extraction Parameters**: A tagger links query fragments to their disambiguated, corresponding Wikipedia instances (i.e., to Wikipedia articles). The tagger is simplified to select the longest instance mentions in case of multiple, overlapping possible mentions. Depending on the sources of textual data available for training, any taggers (Cucerzan, 2007; Ratinov et al., 2011; Pantel et al., 2012) that disambiguate text fragments relative to Wikipedia entries can be employed.

The application of extraction patterns to disambiguated queries identifies matching source queries, which contain a Wikipedia instance and a candidate attribute. There are almost 1 million

| Class: Examples of Instances |
|---|
| Actors: Bobby Todd, Chunky Pandey, Cody Estes, Emma Cleasby, Koen Crucke, Lauro Delgado, Leandra Leal, Leo Hallerstam, Lindsay Sloane, Margit Saad, Renu Saikia |
| Animated characters: Boris Badenov, Flying Dutchman (SpongeBob SquarePants), Kang and Kodos, Little Red-Haired Girl, Road Runner (video game), Strafe (Transformers), Subaru Sumeragi, Teletraan I, Zor (Robotech) |
| Astronomical objects: 65 Andromedae, Beta Lacertae, Dione (moon), Eta Corvi, Jewel Box (star cluster), Lacaille 8760, Messier 90, Mu Pegasi, Radio relics, S/2011 J 2 |
| Automobiles: Chevrolet Tahoe, Ford CD2 platform, Jaguar XJ220, Lancia Dikappa, Mercedes-Benz Vario, Mitsubishi Mizushima, Rolls-Royce Phantom VI, Simca Esplanada, Tata Pixel |
| Awards: Academy Award, Dan Immerfall, Kalyna Roberge, Pulitzer Prizes, Justin Winsor Prize (library), Palme d'Or, Phajol Moolsan, Wallace Roney |
| Chemical elements: Aluminium, Bromine, Californium, Group 12 element, Group 7 element, Ilmenium, Lanthanum, Nebulium, Period 3 element, Polonium, Unbiunium, Yttrium, Zirconium |
| Companies: Consilient, Dorfan, Enercell, Fixafone, Gigaset Communications, Ingman, N3V Games, Quakers Yard and Merthyr Railway, Southeastern Airlines, TV Senado |
| Digital cameras: Fujifilm FinePix T-series, Fujifilm FinePix X100, General Imaging, Kodak DCS 300 series, Nikon Coolpix series, Nikon D60, Olympus E-400, Pentax K10D, Polaroid Z340 |
| Diseases and disorders: Anthropophilia in animals, Bladder spasm, Hyperlysinemia, Inappropriate sinus tachycardia, Lymphoid Leucosis, Pantothenate kinase-associated neurodegeneration, Repetitive strain injury, Xwrits |
| Films: I Am Trying to Break Your Heart: A Film About Wilco, It Takes Two (1995 film), Mockingbird Don't Sing, Moster fra Mols, Prayers for Bobby, Roped, Strings (2004 film), Ten Thousand Years Older |
| Flowers: Acanthephippium mantinianum, Alcea, Dahlia 'Moonfire', Evergreen rose, Flower frog, Geranium caespitosum, Gilliflower, Hyacinth (plant), Iris douglasiana, Lavandula stoechas |
| Hurricanes in North America: 1856 Last Island hurricane, 1948 Miami hurricane, Effects of Hurricane Isabel in New York and New England, Hurricane Bertha (2008), Hurricane Charley, Hurricane Jova (2011) |
| Mobile phones: HTC Desire, LG Rumor, LG VX9400, Motorola A910, Nokia 1100, Nokia 7230, Nokia N71, Samsung S5560, Samsung SGH-T809, Samsung SPH-A460, Serene (phone), Vibrating alert |
| Mountains: Boulder Hills, Cerro San Luis Obispo, Lumiere Peak, Mount Erymanthos, Mount Shindainichi, Mynydd Mawr, Peters Mountain, Poroto Mountains, Remsspitze, Steel Peak, Stob Coire an Laoigh, Stoodley Pike |
| National Basketball Association teams: Chicago Bulls, Cleveland Cavaliers, Dallas Mavericks, Golden State Warriors, Houston Rockets, Indiana Pacers, Miami Heat, Philadelphia 76ers |
| National parks: Bu Gia Map National Park, Defileul Jiului National Park, Dooragan National Park, Great Basalt Wall National Park, Orang National Park, Tortuguero National Park |
| Newspapers: 7days, Chicago Sun-Times, Chojoongdong, Church Times, Corriere del Trentino, Government Gazette (Greece), L'Ordine Nuovo, Le Propagateur Catholique, Novato Advance, Seattle Medium, Venad Pathrika |
| Painters: Chester Harding (painter), Domingo Antonio Velasco, Gifford Beal, Giovanni Ambrogio de Predis, Lech Rzewuski, Ronnie Landfield, Tarcisio Merati, Tyler Vlahovich, Wolfgang Bauer (artist) |
| Programming languages: ALGOL 68, Brutus2D, C11 (C standard revision), CORC, Ease (programming language), GiNaC, IBM Basic assembly language, KANT (software), NESL, Obliq, RoboLogix, Rubinius |
| Religious faiths traditions and movements: Astral body, Chen Tao ("True Way"), Fudoki, Gnostic Mass, Ife, Lutheran Church in Singapore, Omnism, Pneumatic (Gnosticism) |
| Rivers: Arctic Red River, Deadwood River, Kuzumaru Dam, Ounce River, Pingo River, Presumpscot River, Reno (river), Sand Creek (Denver, Colorado), Viehmoorgraben, Zebracu River |
| Skyscrapers: 100 Montgomery Street, 15 Penn Plaza, 19 South LaSalle Street, 595 Market Street, EQT Plaza, Edelweiss (skyscraper), Transamerica Tower (Baltimore) |
| Wine: Acqui Terme, Cinque Terre, Costers del Segre, Elqui River, Los Palacios (Vino de la Tierra), Norte de Granada, Santorini, Sovana DOC, Vinho Verde, Wachau wine |

Table 3: Examples of instances for a sample of target classes, where an instance is represented by the title of its Wikipedia article

such matching queries. In turn, Wikipedia instances are separately associated with the classes (Wikipedia categories) to which they belong, including the target classes described earlier. Transitively, attributes extracted from queries are thus associated to target classes. If an attribute co-occurs in the source queries with fewer than five, distinct instances of a class, then the attribute is deemed unreliable and therefore removed from the set of candidate attributes of the class.

| Label | Value | Examples of Attributes |
|---|---|---|
| vital | 1.0 | Painters: famous when he was alive? |
| | | Programming languages: compiled? |
| | | Stadiums: covered? |
| okay | 0.5 | Painters: born in italy? |
| | | Programming languages: useful? |
| | | Stadiums: safe? |
| wrong | 0.0 | Painters: stolen? |
| | | Programming languages: affectionate pets? |
| | | Stadiums: good college? |

Table 4: Correctness labels manually assigned to attributes extracted for various classes

| Class | Precision of Extracted Attributes | | | |
|---|---|---|---|---|
| | %vital | %okay | %wrong | Score |
| Animated characters | 64 | 32 | 4 | 0.80 |
| Companies | 84 | 0 | 16 | 0.84 |
| Empires | 68 | 20 | 12 | 0.78 |
| Hurricanes in North America | 52 | 44 | 4 | 0.74 |
| Organizations designated as terrorist | 56 | 8 | 36 | 0.60 |
| Programming languages | 96 | 4 | 0 | 0.98 |
| Wine | 44 | 4 | 52 | 0.46 |
| ... | | | | |
| Avg-All-Classes | 78 | 12 | 10 | 0.84 |

Table 5: Accuracy of attributes extracted from conjectural queries, over the set of 40 target classes

## 4 Evaluation Results

**Attribute Accuracy**: A sample of 25 attributes from each target class is manually assigned correctness labels. Following previously introduced methodology, an attribute is marked as *vital*, if it must be present amongrepresentative attributes of the class; *okay*, if it provides useful but non-essential information; and *wrong*, if it is incorrect (Paşca, 2007). When attributes are not *vital*, the choice between the labels *okay* vs. *wrong* often depends on whether the instances mentioned in source queries have been disambiguated to the correct entries in Wikipedia or not. To compute the precision score over a set of attributes, the correctness labels are converted to numeric values as shown in Table 4. Precision is the sum of the correctness values of the attributes, divided by the number of attributes.

Table 5 summarizes the resulting precision scores over the evaluation set of target classes. The scores vary from one class to another, for example 0.46 for *Wine* but 0.98 for *Programming languages*. The average score is 0.84, indicating that attributes extracted from conjectural queries have encouraging levels of accuracy. Table 6 shows examples of attributes extracted for some of the target classes, whereas Table 7 contains examples of source queries from which various attributes are extracted. Attributes extracted from *why*-prefixed source queries tend to be individually more reliable that attributes from *is*-prefixed queries. As noted earlier, this is because *why*-prefixed queries tend to request an explanation for something already known to be true to the questioner, instead of merely asking whether something is true or not.

**Error Analysis**: Erroneous attributes are extracted mainly due to two reasons. First, the presence of noisy instances in an input target class, illustrated earlier in Table 3, causes the extraction of attributes that may be relevant to individual instances but not to the class. For example, a significant number of instances in the target classes *Awards* and *Wine* are incorrect, because Wikipedia lists the categories *Award winners* and *Wine regions* as sub-categories of *Awards* and *Wine* respectively.

| Class: Extracted Attributes |
|---|
| Association football clubs: playing today?, in debt?, a big club?, in the champions league?, in europe?, in fifa 12?, wearing black armbands?, for sale?, going to sign anyone?,.. |
| Automobiles: a good car?, reliable?, front wheel drive?, a good first car?, rear wheel drive?, all wheel drive?, safe?, a girl car?, fast?, 4 wheel drive?, awd?, a sports car?, discontinued?, worth it?, a chick car?, good in snow?, good on gas?, parts expensive?, expensive to maintain?, discontinued?, being discontinued?,.. |
| Awards: married?, famous?, a vegetarian?, overrated?, broke?, cancelled?, left handed?, based on a true story?, playing tonight?, having a baby?, appropriate for kids?, a good show?,.. |
| Battles and operations of World War II: a success?, fought?, necessary?, a turning point?, an important battle?, important to the allies?, a failure?, in world war 2?, successful?, inevitable?, called that?, important to the us?, planned?,.. |
| Chemical elements: dangerous?, flammable?, a metal nonmetal or metalloid?, a solid?, magnetic?, reactive?, toxic?, malleable?, a metalloid?, poisonous?, man made?, rare?, harmful?, a transition metal?, explosive?, an element?, found in nature?, conductive?, ductile?, paramagnetic?, a gas?, safe?, a solid liquid or gas at room temperature?, expensive?, hazardous?, soluble?, stable?, a conductor?, a mineral?, a cation or anion?, a compound?, combustible?, soluble in water?, brittle?, diamagnetic?, corrosive?, diatomic?,.. |
| Companies: a good company to work for?, a good place to work?, free?, worth it?, a scam?, a good brand?, a public company?, legit?, expensive?, reliable?, publicly traded?, open on thanksgiving?, hiring?, down?, expensive?, a franchise?, a fortune 500 company?, going out of business?, open on labor day?, open on easter?, open today?, legitimate?, a good investment?, a word?, dead?, real?, open on sunday?, for sale?, slow?, an american company?, in trouble?,.. |
| Currencies by country: strong?, pegged?, pegged to us dollar?, backed by gold?, overvalued?, going down?, rising?, strong?, strengthening?, a good investment?, appreciating?, depreciating?, a convertible currency?, a proper noun?, capitalized?, convertible?,.. |
| Digital cameras: a good camera?, full frame?, discontinued?, worth it?, worth it?, a professional camera?, waterproof?, a professional camera?, fx or dx?, sdhc compatible?, worth the money?, out of stock?, full frame sensor?, made in japan?, weather sealed?, worth buying?, a good camera for beginners?, being replaced?, easy to use?, weather sealed?, good for video?, a full frame dslr?,.. |
| Diseases and disorders: hereditary?, contagious?, curable?, genetic?, dangerous?, fatal?, serious?, painful?, common?, deadly?, a disability?, a disease?, treatable?, life threatening?, permanent?, inherited?, infectious?, reversible?, rare?, preventable?, a sign of pregnancy?, an autoimmune disease?, dominant or recessive?, communicable?, chronic?, an std?, a mental illness?,.. |
| Films: on netflix?, a true story?, rated r?, a good movie?, scary?, on dvd?, appropriate for kids?, rated pg-13?, funny?, on demand?, for kids?, a remake?, still in theaters?, a disney movie?, in english?, out?, sad?, worth watching?, accurate?, in theaters?, gory?,.. |
| Food ingredients: gluten free?, safe?, bad for you?, good for you?, healthy?, vegan?, safe during pregnancy?, dangerous?, harmful?, toxic?, halal?, fattening?, bad for dogs?, poisonous?, natural?, vegetarian?, acidic?, edible?, good for health?, kosher?, kosher for passover?, unhealthy?, soluble in water?, paleo?, organic?, polar?, flammable?, alkaline?, safe to eat?, good for skin?, a carbohydrate?, safe for pregnant women?, addictive?, good for diabetics?, good for hair?, safe for babies?, a compound?,.. |
| Internet search engines: safe?, down?, free?, legit?, important?, legal?, better than google?, not working?, a meta search engine?, effective?, reliable?, accurate?, case sensitive?, profitable?, account free?, blocked?, a buy?,.. |
| Mobile phones: a good phone?, 3g?, a smartphone?, triband?, available in india?, android?, worth it?, worth buying?, gsm?, better than iphone?, symbian?, touch screen?, compatible with mac?, a smart phone?, 4g?, worth buying?, wifi?, 4g?, discontinued?, quad band?, coming to verizon?, a world phone?, better than iphone 4?, free?, dual sim?, unlocked?,.. |
| National parks: established?, safe?, a world heritage site?, famous?, unique?, expensive?, busy in may?, good for nightlife?, hot in september?, in danger?, in europe?, nice?, open in august?,.. |
| Newspapers: conservative?, liberal or conservative?, reliable?, biased?, right wing?, a reliable source?, a tabloid?, left wing?, a broadsheet?, credible?, a daily newspaper?, a magazine?, a national newspaper?, a scholarly source?, free?, real?, important?, app free?, fake?, reputable?, a peer reviewed journal?, a reputable source?, a scholarly journal?,.. |
| Programming languages: case sensitive?, object oriented?, easy to learn?, open source?, still used?, a scripting language?, compiled?, easy?, compiled or interpreted?, installed?, fast?, object oriented language?, oop?, platform independent?, turing complete?, compatible with windows 7?, safe?, popular?, slow?, developed?, a good language?, obsolete?, portable?, secure?, thread safe?,.. |
| Rivers: polluted?, safe?, clean?, safe to swim in?, famous?, brown?, flooding?, tidal?, man made?, navigable?, sustainable?, a good place to live?, worth visiting?, drying up?,.. |
| Stadiums: covered?, a dome?, air conditioned?, indoors?, heated?, roof open?, a retractable roof?, safe?, grass?, haunted?, demolished?, complete?, fixed?, open to the public?, open today?, still standing?, turf?, worth it?,.. |
| Treaties: signed?, legally binding?, fair?, controversial?, a failure?, successful?, necessary?, needed?, written?, significant?, a law?, significant?, effective?, introduced?, written?, working?,.. |
| Universities and colleges: a party school?, a public or private university?, aacsb accredited?, a good school for nursing?, fully accredited?, a good university?, jesuit?, a good school for business?, test optional?, a four year college?, a graduate school?, a reputable school?, a safe campus?, a state college?, mba accredited?, a top school?, considered ivy league?, in a bad neighborhood?, accredited in canada?, going to a bowl game 2011?, quarter or semester?, single choice early action?,.. |
| Video games: multiplayer?, free?, worth buying?, fun?, 2 player?, online?, hard?, out?, for mac?, rated m?, on ps3?, on pc?, safe?, split screen?, compatible with windows 7?, on xbox 360?, scary?, free roam?, worth playing?, region free?, co op?, on wii?, compatible with vista?, expensive?, open world?, not working?, real?, on psp?, laggy?,.. |
| Wine: a grape?, expensive?, safe?, worth visiting?, famous?, important?, an island?, hot in october?, sweet?, a dry wine?, a region?, in the eu?, in tuscany?, nice?, red or white wine?, italian?, a champagne?, a city or country?, a dessert wine?,.. |

Table 6: Examples of attributes extracted from queries

Consequently, *Awards* and *Wine* respectively contain many people names and geographical regions as their instances. Noisy instances in target classes could be removed or reduced, by filtering the

| Source Query Containing a Mention of an Instance and an *Attribute* | Instance | Class |
|---|---|---|
| are e36 m3 *reliable* | BMW M3 | Automobiles |
| is the chevy avalanche *reliable* | Chevrolet Avalanche | Automobiles |
| why is the honda civic *reliable* | Honda Civic | Automobiles |
| is chrysler neon *a good car* | Chrysler Neon | Automobiles |
| is honda element *a good car* | Honda Element | Automobiles |
| are honda accords *rear wheel drive* | Honda Accord | Automobiles |
| is a mini cooper s *rear wheel drive* | Mini | Automobiles |
| is dodge charger *rear wheel drive* | Dodge Charger (LX) | Automobiles |
| is the jaguar xf *rear wheel drive* | Jaguar XF | Automobiles |
| are ford mustangs *expensive to maintain* | Ford Mustang | Automobiles |
| are bmw 3 series *expensive to maintain* | BMW 3 Series | Automobiles |
| is range rover *expensive to maintain* | Range Rover | Automobiles |
| are boxsters *expensive to maintain* | Porsche Boxster | Automobiles |
| is a range rover *expensive to maintain* | Range Rover | Automobiles |
| are 350z *expensive to maintain* | Nissan 350Z | Automobiles |
| why was the battle of el alamein *important to the allies* | Second Battle of El Alamein | Battles and operations of World War II |
| why was the north africa campaign *important to the allies* | North African Campaign | Battles and operations of World War II |
| is europium *dangerous* | Europium | Chemical elements |
| why is radon *dangerous* | Radon | Chemical elements |
| is scandium *a solid liquid or gas at room temperature* | Scandium | Chemical elements |
| is xenon *a solid liquid or gas at room temperature* | Xenon | Chemical elements |
| is iron *corrosive* | Iron | Chemical elements |
| is hydrogen gas *corrosive* | Hydrogen | Chemical elements |
| is chlorine *corrosive* | Chlorine | Chemical elements |
| why is chlorine *corrosive* | Chlorine | Chemical elements |
| why is oxygen *soluble in water* | Oxygen | Chemical elements |
| why is iodine *soluble in water* | Iodine | Chemical elements |
| why are the green bay packers *publicly traded* | Green Bay Packers | Companies |
| is quicken loans *publicly traded* | Quicken Loans | Companies |
| why is starbucks *a good company to work for* | Starbucks | Companies |
| why is southwest airlines *a good company to work for* | Southwest Airlines | Companies |
| is henkel *a good company to work for* | Henkel | Companies |
| is sodexo *a good company to work for* | Sodexo | Companies |
| is porter cable *a good brand* | Porter-Cable | Companies |
| is chevrolet *a good brand* | Chevrolet | Companies |
| why is coca cola *a good brand* | Coca-Cola | Companies |
| why is nike *a good brand* | Nike, Inc. | Companies |
| is singapore dollar *backed by gold* | Singapore dollar | Currencies by country |
| is the philippine peso *backed by gold* | Philippine peso | Currencies by country |
| is iraq dinar *a convertible currency* | Iraqi dinar | Currencies by country |
| why is the malaysian ringgit *pegged to us dollars* | Malaysian ringgit | Currencies by country |
| why is hong kong dollar *pegged to us dollar* | Hong Kong dollar | Currencies by country |
| is singapore dollar *pegged to us dollar* | Singapore dollar | Currencies by country |
| is indian rupee *pegged to us dollar* | Indian rupee | Currencies by country |
| is the d90 *weather sealed* | Nikon D90 | Digital cameras |
| is d700 *weather sealed* | Nikon D700 | Digital cameras |
| is nikon d90 *weather sealed* | Nikon D90 | Digital cameras |
| is nikon d300 *full frame sensor* | Nikon D300 | Digital cameras |
| why is gothika *rated r* | Gothika | Films |
| why was the book of eli *rated r* | The Book of Eli | Films |
| why was the matrix *rated r* | The Matrix | Films |

Table 7: Examples of source queries from which candidate attributes are extracted for various target classes

category-to-category edges from Wikipedia into a category taxonomy (Ponzetto and Strube, 2007; Ponzetto and Navigli, 2009) prior to assembling the target classes from categories. Second, the incorrect disambiguation of instances in queries causes attributes of a different instance with a similar name to be associated with the wrong class.

| Conjectural Attribute → Possible Mappings to Equivalent Nominal Attributes | |
|---|---|
| Astronomical objects: bigger than earth? → size relative to earth$^{A}$, size$^{P}$, volume$^{W,F,P}$ | Astronomical objects: capable of supporting life? → ability to support life$^{A}$ |
| Astronomical objects: circumpolar? → ∅ | Food ingredients: fattening? → energy$^{W,F}$, calories$^{W,F}$ |
| Chemical elements: man made? → manufacturing process$^{P}$ | Chemical elements: toxic? → toxicity$^{P}$ |
| Chemical elements: soluble in water? → solubility in water$^{W,F,P}$ | Chemical elements: a solid liquid or gas at room temperature? → phase$^{W,F,P}$ |
| Chemical elements: conductive? → conductivity$^{P}$, electrical conductivity$^{A}$, electrical resistivity$^{W,F}$ | Automobiles: reliable? → reliability$^{P}$ |
| Automobiles: front wheel drive? → driveline$^{F}$ | Automobiles: safe? → safety rating$^{P}$ |
| Mobile phones: smartphone? → function as smartphone$^{A}$ | Mobile phones: compatible with mac? → compatibility with mac$^{P}$ |
| Cities: dog friendly? → dog friendliness$^{A}$ | Cities: nice place to live? → attractiveness as a place to live$^{A}$ |
| Cities: cheap? → cost of living$^{P}$ | Companies: publicly traded? → stock symbol$^{P}$ |
| Companies: fortune 500 company? → ∅ | Companies: for sale? → availability for sale$^{A}$ |
| Companies: a good place to work? → ranking$^{P}$, employee benefits$^{P}$ | Films: on netflix? → availability on netflix$^{A}$, netflix title$^{F}$, providers$^{P}$ |
| Films: sad? → genre$^{F,P}$ | National parks: busy in may? → number of visitors in may$^{A}$, may visitors statistics$^{A}$, visitor statistics$^{P}$ |
| National Basketball Association teams: losing money? → cash flow$^{A}$, financial statement$^{P}$, debt, | Newspapers: conservative? → preference for conservative views$^{A}$, political alignment$^{W}$, political bias$^{P}$ |
| Newspapers: reputable? ? → reputation$^{A}$ | Programming languages: case sensitive? → case sensitivity$^{A}$ |
| Programming languages: obsolete? → ∅ | Treaties: binding? → ∅ |

Table 8: Examples of mappings from conjectural attributes to their equivalent nominal attributes, if any. Nominal attributes present in Wikipedia, Freebase or among non-conjectural attributes extracted with previous methods (Paşca et al., 2007), are marked as W, F and/or P respectively; otherwise, they are marked A. A lack of equivalent nominal attributes is marked as ∅

**Relation to Nominal Attributes**: Table 8 reviews a sample of conjectural attributes, from the point of view of the availability of equivalent noun-phrase attributes into which conjectural attributes can be mapped. The nominal attributes would correspond to attributes available in human-compiled resources or that may be extracted by previous methods. Out of a sample of 200 conjectural attributes, 72% are manually found to have possible mappings into equivalent nominal attributes, of which 28% fully preserve the meaning (e.g., *safe?→safety rating*) but 44% only partially approximate the meaning (e.g., *bigger than earth?→volume*). When possible mappings to nominal attributes exist, whether fully or partially meaning-preserving, 18% (Wikipedia), 22% (Freebase) and 33% are present among the attributes available for instances of the respective classes in Wikipedia, Freebase or among the top 500 attributes returned by the method from (Paşca, 2007). About 31% of the conjectural attributes in the sample are found to not have a nominal equivalent, and an additional 28% have a possible nominal equivalent that is not present in any of the respective resources. Overall, the analysis suggests that conjectural attributes are not already captured by, and therefore complement, existing or previously extracted attributes.

**Relation to Unary Attributes**: To our knowledge, (Van Durme et al., 2008) is the only previous attribute extraction method allowing for the extraction of non-nominal attributes. It parses document sentences, leading to so-called unary attributes of classes, such as *trapped*, *dangerous*, and *unfortunate* for *Animals*. The attributes are collected from document sentences, roughly by identifying adjectival modifiers of mentions of the class (“*..the animals were trapped..*”, “*..the trapped*

| Class | Extracted Attributes |
|---|---|
| Painter | $D_U$: famous, romantic, distinguished, celebrated, well-known, pre-raphaelite, flemish, dutch, abstract |
| | $Q_C$: famous when he was alive?, an important artist?, dead?, born in italy?, rich or poor?, left handed? |
| Animal | $D_U$: dead, trapped, dangerous, unfortunate, intact, hungry, wounded, tropical, sick, favourite |
| | $Q_C$: good with children?, aggressive?, smart?, dangerous?, hard to train?, endangered?, nocturnal? |
| Drug | $D_U$: dangerous, powerful, addictive, safe, illegal, experimental, effective, prescribed, harmful, hallucinatory |
| | $Q_C$: safe during pregnancy?, addictive?, dangerous?, over the counter?, a controlled substance?, legal?, effective? |
| Apple | $D_U$: red, juicy, fresh, bad, substantive, stuffed, shiny, ripe, green, baked |
| | $Q_C$: good for baking?, tart?, cooking apple?, sweet?, healthy?, good for canning?, biennial?, gmo? |
| Earthquake | $D_U$: disastrous, violent, underwater, prolonged, powerful, popular, monstrous, fatal, famous, epic |
| | $Q_C$: predicted?, destructive?, man made?, deadly?, devastating?, common?, normal? |

Table 9: Comparison between unary attributes ($D_U$) extracted from documents as described in (Van Durme et al., 2008), and conjectural attributes ($Q_C$) extracted from queries as described in the current paper. To extract comparable attributes, the classes from (Van Durme et al., 2008), shown in the first column, are manually mapped into their equivalent Wikipedia categories *Painters*, *Animals*, *Drugs*, *Apple cultivars* and *Earthquakes* respectively

*animals*..”). In comparison to (Van Durme et al., 2008), our method takes noisy queries as input, rather than clean document sentences from reliable documents such as news articles. The comparison in Table 9, between attributes extracted for a sample of classes, highlights a few additional differences between the two methods. First, the attributes extracted in (Van Durme et al., 2008) seem to be limited to single-word adjectives, whereas conjectural attributes accommodate other phrases too. Second, the different intuitions behind the two methods determine what kind of attributes one would expect to see extracted. Unary attributes from (Van Durme et al., 2008) often capture transient states or events in which the respective classes are involved, like *Animals* being *trapped*, or *Apples* being *ripe* or *baked*. Such states or events may be relevant in the particular context of the document containing the source sentences. But they are not necessarily generally-relevant, context-independent properties that Web users would likely inquire about. Indeed, adding the property *trapped* for the class *Animals* to Wikipedia, and filling in its corresponding values for all instances (kinds) of *Animals*, would likely have little value. In contrast, conjectural attributes like *aggressive?* or *nocturnal?* are information-seeking, and often refer to permanent, context-independent rather than transient, context-dependent properties of the instances of the class *Animals*.

## 5 Related Work

A variety of methods address the more general task of acquisition of open-domain relations from text, e.g., (Zhu et al., 2009; Carlson et al., 2010; Fader et al., 2011; Lao et al., 2011). In order to acquire class attributes in particular, a common strategy is to first acquire attributes of instances, then aggregate or propagate (Talukdar and Pereira, 2010) attributes, from instances to the classes to which the instances belong. The identification of relevant instances within queries is related to the task of word sense disambiguation (Ponzetto and Navigli, 2010).

Data available within Web documents, from which attributes are extracted in previous work, includes unstructured (Tokunaga et al., 2005; Paşca et al., 2007), structured (Raju et al., 2008) and semi-structured text (Yoshinaga and Torisawa, 2007), layout formatting tags (Wong et al., 2008), itemized lists or tables (Cafarella et al., 2008). Another source of attributes is data in human-compiled encyclopedia (Wu et al., 2008; Cui et al., 2009), including infoboxes and category labels (Suchanek et al., 2007; Nastase and Strube, 2008; Wu and Weld, 2008) associated with

Wikipedia articles (Remy, 2002). The role of Web search queries, as an alternative textual data source to Web documents in open-domain information extraction, has been investigated in the tasks of attribute extraction (Paşca, 2007), as well as in collecting labeled (Sekine and Suzuki, 2007; Pennacchiotti and Pantel, 2009) or unlabeled sets of related instances (Jain and Pennacchiotti, 2010) and ranking of class labels already extracted from text (Billerbeck et al., 2010).

## 6 Conclusion

Collectively, queries that inquire whether an attribute applies to individual instances or not are evidence that the attribute in question does apply to classes to which the instances belong. The resulting conjectural attributes have encouraging accuracy, and complement attributes available in manually-compiled resources or automatically extracted with previous methods. Current work investigates the role of repositories of class labels (*heated stadiums*) extracted from text for various instances (*Millennium Stadium*), as an additional source of evidence towards extracting class attributes (*heated?*); and the acquisition of (mostly binary) values of conjectural attributes for individual instances.

## References


Billerbeck, B., Demartini, G., Firan, C., Iofciu, T., and Krestel, R. (2010). Ranking entities using Web search query logs. In *Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries (ECDL-10)*, pages 273–281, Glasgow, Scotland.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 International Conference on Management of Data (SIGMOD-08)*, pages 1247–1250, Vancouver, Canada.

Cafarella, M., Halevy, A., Wang, D., Wu, E., and Zhang, Y. (2008). WebTables: Exploring the power of tables on the Web. In *Proceedings of the 34th Conference on Very Large Data Bases (VLDB-08)*, pages 538–549, Auckland, New Zealand.

Carlson, A., Betteridge, J., Wang, R., Hruschka, E., and Mitchell, T. (2010). Coupled semi-supervised learning for information extraction. In *Proceedings of the 3rd ACM Conference on Web Search and Data Mining (WSDM-10)*, pages 101–110, New York.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing (EMNLP-07)*, pages 708–716, Prague, Czech Republic.

Cui, G., Lu, Q., Li, W., and Chen, Y. (2009). Automatic acquisition of attributes for ontology construction. In *Proceedings of the 22nd International Conference on Computer Processing of Oriental Languages*, pages 248–259, Hong Kong.

Fader, A., Soderland, S., and Etzioni, O. (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.

Jain, A. and Pennacchiotti, M. (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.

Lao, N., Mitchell, T., and Cohen, W. (2011). Random walk inference and learning in a large scale knowledge base. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP-11)*, pages 529–539, Edinburgh, Scotland.

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

Paşca, M. (2007). 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.

Paşca, M. and Van Durme, B. (2007). What you seek is what you get: Extraction of class attributes from query logs. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, pages 2832–2837, Hyderabad, India.

Paşca, M., Van Durme, B., and Garera, N. (2007). The role of documents vs. queries in extracting class attributes from text. In *Proceedings of the 16th International Conference on Information and Knowledge Management (CIKM-07)*, pages 485–494, Lisbon, Portugal.

Pantel, P., Lin, T., and Gamon, M. (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.

Pennacchiotti, M. and Pantel, P. (2009). Entity extraction via ensemble semantics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP-09)*, pages 238–247, Singapore.

Ponzetto, S. and Navigli, R. (2009). Large-scale taxonomy mapping for restructuring and integrating Wikipedia. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-09)*, pages 2083–2088, Pasadena, California.

Ponzetto, S. and Navigli, R. (2010). Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, pages 1522–1531, Uppsala, Sweden.

Ponzetto, S. and Strube, M. (2007). Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI-07)*, pages 1440–1447, Vancouver, British Columbia.

Raju, S., Pingali, P., and Varma, V. (2008). An unsupervised approach to product attribute extraction. In *Proceedings of the 31st International Conference on Research and Development in Information Retrieval (SIGIR-08)*, pages 35–42, Singapore.

Ratinov, L., Roth, D., Downey, D., and Anderson, M. (2011). Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-11)*, pages 1375–1384, Portland, Oregon.

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

Sekine, S. and Suzuki, H. (2007). Acquiring ontological knowledge from query logs. In *Proceedings of the 16th World Wide Web Conference (WWW-07), Posters*, pages 1223–1224, Banff, Canada.

Suchanek, F., Kasneci, G., and Weikum, G. (2007). 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.

Talukdar, P. and Pereira, F. (2010). Experiments in graph-based semi-supervised learning methods for class-instance acquisition. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, pages 1473–1481, Uppsala, Sweden.

Tokunaga, K., Kazama, J., and Torisawa, K. (2005). 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.

Van Durme, B., Qian, T., and Schubert, L. (2008). Class-driven attribute extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*, pages 921–928, Manchester, United Kingdom.

Wong, T., Lam, W., and Wong, T. (2008). An unsupervised framework for extracting and normalizing product attributes from multiple Web sites. In *Proceedings of the 31st International Conference on Research and Development in Information Retrieval (SIGIR-08)*, pages 35–42, Singapore.

Wu, F., Hoffmann, R., and Weld, D. (2008). Information extraction from Wikipedia: Moving down the long tail. In *Proceedings of the 14th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-08)*, pages 731–739.

Wu, F. and Weld, D. (2008). Automatically refining the Wikipedia infobox ontology. In *Proceedings of the 17th World Wide Web Conference (WWW-08)*, pages 635–644, Beijing, China.

Yoshinaga, N. and Torisawa, K. (2007). Open-domain attribute-value acquisition from semi-structured texts. In *Proceedings of the 6th International Semantic Web Conference (ISWC-07), Workshop on Text to Knowledge: The Lexicon/Ontology Interface (OntoLex-2007)*, pages 55–66, Busan, South Korea.

Zhu, J., Nie, Z., Liu, X., Zhang, B., and Wen, J. (2009). StatSnowball: a statistical approach to extracting entity relationships. In *Proceedings of the 18th World Wide Web Conference (WWW-09)*, pages 101–110, Madrid, Spain.