German Typographers vs. German Grammar: Decomposition of Wikipedia Category Labels into Attribute-Value Pairs

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

ABSTRACT

Given an instance (“*Julieta Pinto*”), most methods for open-domain information extraction focus on acquiring knowledge in the form of either class labels (“*Costa Rican short story writers*”, “*Women novelists*”) referring to concepts to which the instance belongs; or facts (“*nationality: Costa Rica*”) connecting the instance (“*Julieta Pinto*”) to other instances or concepts (“*Costa Rica*”), where the fact and the other instance often take the form of an attribute (“*nationality*”) and a value (“*Costa Rica*”) respectively. From extraction through internal representation and storage, class labels and facts are treated as if they carved out disconnected slices within the larger space of factual knowledge. This paper argues that class labels and facts pertaining to an instance exist in symbiosis rather than as a dichotomy. A constituent (“*Costa Rican*”) within a class label (“*Costa Rican short story writers*”) of an instance may be indicative of a fact (“*nationality: Costa Rica*”) applicable to the instance and vice-versa. As an illustration of the relationship between class labels and facts, the paper introduces an open-domain method for the better understanding of the semantics of class labels in one of the larger and most widely-used repositories of knowledge, namely the categories in the Wikipedia category network. The method exploits the category network to associate constituents (“*Costa Rican*”) within names of Wikipedia categories, with attributes (“*nationality*”) that explain their role.

Keywords

Open-domain information extraction, class labels, facts, attributes

1. INTRODUCTION

Background: The growth of open-domain knowledge in repositories such as Wikipedia has, if not enabled, at least stimulated interest and progress in the area of large-scale knowledge acquisition [3, 13, 7]. In parallel, data within Wikipedia articles, whether taking the form of infoboxes, inter-article links or text itself, has served as a steady source of raw material for training methods for open-domain information extraction [43, 8, 13, 41, 28]. Wikipedia

groups together articles into categories, which are in turn connected to incrementally more general categories, forming what is referred to as the Wikipedia category network. Whether edges in the category network are restricted to hierarchical edges [33, 32, 9] or not, the category network provides fine-grained class labels (i.e., categories) for millions of instances (i.e., articles).

Motivation: As is the case with virtually all class labels automatically extracted for instances from text [5, 44, 9], Wikipedia categories are available for Wikipedia articles only as strings in lexical space. Due to ambiguity, the same lexical constituent (e.g., “*German*”) may play different roles (“*country*” vs. “*nationality*”) vs. “*language*”) in different categories (“*German civil aircraft*”, “*German typographers*”, “*German grammar*”). Even after restricting a lexical constituent (“*chess*”) to a particular disambiguation (*Chess (sport)*, as opposed to *Chess (movie)* or *Chess (magazine)*), the same lexical constituent with the same disambiguation may still play different roles (“*subject matter*” vs. “*sport*”) in different categories (“*Chess books*”, “*Chess coaches*”). Annotating constituents of categories with their roles would be a significant step towards adding a layer of understanding categories beyond mere strings. At a minimum, the constituents and their roles uncover attribute-value properties applying to individual instances of the category, which may not be available in structured form in the articles of the instances. For example, the category “*Costa Rican short story writers*” could be annotated so that the roles of its constituents (“*short story*”) become explicit (“*format*”). The roles point to attribute-value properties (“*format: short story*”) that should apply to instances of the category, where instances are articles linked under the category, such as the article titled (“*Julieta Pinto*”). Yet the meaning of a Wikipedia category, let alone the role of its constituents, is rarely explored [24], nor is it made available to downstream applications.

Contributions: This paper proposes a method to automatically annotate modifier constituents within Wikipedia categories, which are strings in lexical space, with attributes that reveal the constituents’ roles in defining the underlying semantics of the categories. The method is beneficial to Wikipedia and transitively to other repositories derived from Wikipedia, including DBpedia [2], Yago [13] and Wikidata [39]. Experimental results demonstrate that lexical Wikipedia categories, including finer-grained categories containing multiple modifier constituents, can be automatically annotated with attributes at precision and coverage above 0.8 and 0.7 respectively.

2. FROM CLASS LABELS TO FACTS

Artificial Separation of Class Labels vs. Facts: Both human-curated and automatically-extracted knowledge repositories are or-

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

WSDM 2017 February 06-10, 2017, Cambridge, United Kingdom

© 2017 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-4675-7/17/02.

DOI: <http://dx.doi.org/10.1145/3018661.3018662>

ganized around instances (i.e., entities) whose categories, properties, relations and schemas they attempt to represent. Among various types of knowledge that can be compiled about an instance (“*Julieta Pinto*”), research efforts aim at compiling either lexical class labels (“*short story writers*”, “*women writers*”, “*Costa Rican novelists*”) [11, 16, 5, 30, 40], or facts capturing attribute-value properties (“*nationality: Costa Rica*”, “*gender: female*”) [6, 43, 13, 42, 10].

Connecting Class Labels to Facts: The separation of knowledge to be compiled and represented for an instance into either class labels or facts is a practical simplification. It may have encouraged progress in the short term. In the long term, better and more numerous class labels and facts could potentially be extracted for an instance, if the two were treated not separately but in symbiosis. Class labels and facts are often reflections of similar if not identical knowledge, as expressed in the following hypotheses:

Hypothesis 1: If a set of instances shares common properties, then properties are indicative of hypothetical class labels associated with the concept to which the set of instances belongs. Modifier constituents within class labels will compactly refer to the properties.

Hypothesis 2: If a class label is associated with the concept to which a set of instances belongs, then the class label is indicative of hypothetical common properties shared among the set of instances. The properties may be referred to compactly in the form of modifiers within the class labels.

To illustrate, consider a set of instances who are all *Engineers*. In the first hypothesis, if the attribute-value property “*specialty: electrical*” is known to apply to the set of instances, one could infer that a class label such as “*electrical engineers*” should also apply to the set of instances. Conversely, in the second hypothesis, if the class label “*Electrical engineers*” is known to apply to the set of instances, one could infer that the attribute-value property “*specialty: electrical*” should also apply to the set of instances. In the class label, the role played by the modifier constituent (“*electrical*”) relative to the head constituent (“*engineers*”) is precisely to restrict the attribute “*specialty*” to the value “*electrical*”.

Problem Definition: Class labels in general are noun phrases that consist of a head constituent, and zero or more modifier constituents. The task being addressed is the interpretation of the roles played by as many modifier constituents as possible, relative to the head constituent. Interpretations take the form of phrases assigned to modifier constituents, referring to attributes that instances of the class would have, and whose values would be the constituents.

Given the class label “*Costa Rican short story writers*”, the interpretations of its modifier constituents “*Costa Rican*” and “*short story*” could be attributes such as “*nationality*” and “*format*” respectively. A modifier constituent (“*Costa Rican*”) provides the value of the attribute assigned to it (“*nationality*”). Thus, the task effectively decomposes class labels that apply to instances, into attribute-value properties that should apply to instances. The extracted interpretations act as bridges between class labels and facts, allowing one to potentially infer facts from class labels. Indeed, the interpretations extracted for “*Costa Rican short story writers*” imply that any instance that belongs to the category, such as the article titled “*Julieta Pinto*”, can be associated with facts like “*nationality: Costa Rica*”. Note that the task being addressed is separate from, and cannot be accomplished simply as a by-product of, the lower-level task of disambiguating lexical modifier constituents into semantically grounded constituents. For example, even if the constituent “*chess*” in the class labels “*chess books*” and “*chess coaches*” were correctly disambiguated to the Wikipedia article titled *Chess* (i.e., <http://en.wikipedia.org/wiki/Chess>), that would

still not distinguish the different roles (“*subject matter*”, “*sport*”) that the constituent plays in the respective class labels.

3. FACTS FROM WIKIPEDIA CATEGORIES

Scope: Depending on their source, class labels and their instances may be noun phrases extracted as IsA relations from text [11, 16, 21]; or members of synonym sets in hypernym classes available in WordNet; or categories available in Wikipedia for each article [9]. In the following, the scope of class labels is restricted to Wikipedia categories.

Terminology: Wikipedia organizes each of its articles under one or more Wikipedia categories. Online entries of Wikipedia categories are prefixed by “<http://en.wikipedia.org/wiki/Category:>” and followed by the respective category names. Categories (e.g., “*Costa Rican short story writers*”) consist of constituents, which are contiguous sequences of tokens of the category names. Constituents are nominal or adjectival noun subphrases. Depending on their syntactic role, they may be modifier constituents (“*Costa Rican*”, “*short story*”) or head constituents (“*writers*”). Attributes are noun-phrase annotations (“*nationality*”, “*format*”) extracted and associated with individual constituents (“*Costa Rican*”, “*short story*”) of a category. Attributes explain the role of the constituents (“*nationality*” for “*Costa Rican*”) relative to the category. An attribute, along with a constituent it annotates, together capture an attribute-value property or fact (“*nationality: Costa Rica*”), which is expected to apply to all instances (“*Julieta Pinto*”, “*Carlos Luis Fallas*”) of the category.

Wikipedia category network: Each Wikipedia article (e.g., the article titled “*Julieta Pinto*”) is connected to its ancestor Wikipedia categories, which are parent categories listed at the bottom of the article (“*Costa Rican short story writers*”, “*Costa Rican novelists*”) or, recursively, parent categories of parent categories (“*Costa Rican writers*”, “*Short story writers by nationality*”, “*Costa Rica*”, “*Arts*”, “*Countries*” and many others). Together, the connections among categories and their ancestor or descendant categories form the Wikipedia category network [33], a small portion of which is shown in the upper portion of Figure 1. Since edges in the category network may be but are not necessarily IsA relations, the category network does not organize categories into a clean hierarchy [33, 9].

Acquisition from Wikipedia: The extraction method computes mappings from Wikipedia categories to attributes, in several stages:

- (1) extract attributes for constituents in a Wikipedia category, by aligning the category with its parent categories where possible;
- (2) propagate attributes of constituents from ancestor categories down to descendant categories, if the categories contain the same constituent; and
- (3) propagate attributes of constituents from categories to other categories, if the categories (a) contain the same constituent and (b) are deemed compatible based on pairwise similarity of ngrams containing their head constituents.

A subsequent stage assigns an attribute to a constituent of a category, only if none of the previous stages already assigned an attribute to that constituent in that category. More details on the stages are given below.

Alignment with Parent Categories: Each category to interpret is split into all possible decompositions of three consecutive sequences of tokens $C=[C_1\ C_2\ C_3]$, where the three sequences correspond to a possibly-empty hypothetical prefix C_1 , a hypothetical modifier constituent to be interpreted C_2 , and a hypothetical suffix C_3 . For example, the category “*Costa Rican writers*” is split into $[C_1=\text{“empty”}, C_2=\text{“Costa”}, C_3=\text{“Rican writers”}]$, $[C_1=\text{“empty”}, C_2=\text{“Costa Rican”}, C_3=\text{“writers”}]$, $[C_1=\text{“Costa”}, C_2=\text{“Rican”}, C_3=\text{“writers”}]$. Separately, the parent categories of the category

Figure 1: Overview of extraction of attributes for modifier constituents within Wikipedia categories, by aligning a category with its parent categories that satisfy a particular format (middle portion of the figure) and by propagating attributes of constituents, from ancestor categories down to descendant categories that contain the same constituent (lower portion). Solid lines represent edges in the category network. Dotted lines represent alignments of sequences of tokens and propagation of attributes. Arrows represent assignments of attributes

are similarly split into all possible decompositions of three consecutive sequences of tokens $P=[P_1 P_2 P_3]$, as long as the parent categories are such that P_2 is the string “by”. For example, the parent category “Writers by nationality” is split into [P_1 =“Writers”, P_2 =“by”, P_3 =“nationality”]. If such decompositions of the parent category exist, and C_3 matches P_1 , then C_3 and P_1 are assumed to be head constituents, possibly preceded by other modifiers, of their noun phrases C and P ; C_2 is assumed to be a modifier constituent of the category C ; and P_3 is assumed to be an attribute of the modifier constituent C_2 . The sequences of tokens C_3 and P_1 are considered to be a match, if the latter sequence contains the former sequence after token normalization. For example, the string “Writ-

Figure 2: Overview of extraction of attributes of modifier constituents within Wikipedia categories, by propagating attributes of constituents between categories deemed compatible based on pairwise similarity of ngrams containing their head constituents. Dotted lines and arrows represent pairs of similar phrases and assignments of attributes respectively

ers” from “Writers by nationality” contains the string “writers” from “Costa Rican writers”. The sequence P_3 (“nationality”) is retained as an attribute of the modifier constituent C_2 (“Costa Rican”) of the category “Costa Rican writers”. Additional examples are shown in the middle portion of Figure 1.

Propagation to Descendant Categories: Since a relatively small fraction of categories in Wikipedia have the format “ X by Y ”, not many child categories have attributes assigned to any of their constituents. Taking advantage of longer paths in the Wikipedia category network than only child to parent edges, attributes already assigned to constituents in ancestor categories are propagated down to the same constituents in descendant categories. As illustrated in the lower portion of Figure 1, the attribute “genre” is already assigned in the previous stage to the constituent “animated” in the ancestor category “Animated television series”. The attribute is propagated downward to the same constituent “animated” in the descendant category “1980s American animated television series”.

The same constituent may have different attributes in different ancestor categories. In such cases, multiple candidate attributes may be competing during propagation, for possible assignment to the same constituent in a descendant category. The candidate attributes are ranked in increasing order of the length of the upward path in the category network between the descendant category, on one hand, and the ancestor category to which a candidate attribute is propagated from, on the other hand. The candidate attribute available for the ancestor category with the shortest path is propagated to the descendant category. Because the path from the descendant category “German indie rock groups” is shorter to the ancestor “German musicians” than to the ancestor “German music industry”, it is the candidate attribute from “German musicians”, namely “nationality”, that is propagated to the constituent “German” in the descendant category “German indie rock groups”.

Propagation to Compatible Categories: Propagation to descendant categories requires the ancestor and the descendant to be connected by a path in the category network. The alternative described in Figure 2 does not. Instead, it propagates attributes between pairs

Phrase: Ranked List of Most Similar Phrases
engineers: [technicians, design engineers, software engineers, engineers and scientists, technologists, systems engineers, project managers, chemists, geologists, scientists and engineers, .., engineering, ..]
engineering: [systems engineering, engineering technology, science and engineering, engg, science, electrical engineering, system engineering, science engineering, civil engineering, chemistry, ..., engineers, ..]
music: [classical music, dance, jazz, musicmusic, classic music, songs, christian music, dance music, acoustic music, electronic music, ..., soundtracks, ..]
musicians: [artists, performers, singers, instrumentalists, songwriters, dancers, jazz musicians, musical groups, drummers, writers, ..]
musical groups: [music groups, musicians, sportspeople, dance groups, orchestras, vocalists, choirs, songwriters, jazz groups, ..]
soundtracks: [movie soundtracks, film scores, movie scores, film soundtracks, theme tunes, theme songs, original soundtracks, music soundtracks, movie themes, albums, ..]
traditional music: [folk music, traditional folk music, traditional irish music, irish music, celtic music, traditional dances, irish traditional music, traditional songs, indigenous music, ethnic music, ..]

Table 1: Sample of ranked lists of similar phrases available for various phrases

of categories regardless of the position of the categories in the category network, as long as the categories are deemed compatible.

As a prerequisite to assessing compatibility among categories, distributionally similar phrases [18, 19, 29] and their scores are collected in advance. Distributional similarities serve as a proxy for estimating category compatibility. Two categories are compatible, if (a) they share a common modifier constituent, and (b) any pair of ngrams that occur in the respective categories after the modifier constituent, and contain the head tokens, are distributionally similar. In Figure 2, “*Traditional pop soundtracks*” and “*African traditional music*” are deemed compatible because they contain the same modifier constituent “*traditional*”, and the ngram *soundtracks* from the first category is distributionally similar to the ngram *music* from the second category. If two categories are compatible, an attribute (“*genre*”) already assigned to a constituent (“*traditional*”) in the first category, but not to the same constituent in the second category, is a candidate attribute for propagation from the first to the second category.

If a category is compatible with multiple categories, multiple candidate attributes may be propagated for assignment to the same constituent of the category. The candidate attributes are ranked in decreasing order of the maximum similarity score between an ngram in the category, on one hand, and an ngram in a compatible category, on the other hand. The candidate attribute with the highest similarity score is propagated to the category. Between the attributes “*type*” available for the category “*Electrical standards*” and “*specialty*” available for the category “*Electrical engineers*”, the latter is selected for the constituent “*Electrical*” in “*Electrical engineering*”, because the ngram “*engineering*” is more similar to “*engineers*” than to “*standards*”.

4. EXPERIMENTAL SETTING

Data Sources: The experiments rely on English entries in a snapshot of Wikipedia articles as of September 2015. In the snapshot, a Wikipedia article (e.g., titled “*Julieta Pinto*”) is mapped to 8 parent categories on average, available as Wikipedia categories listed at the bottom of the article (e.g., “*Costa Rican short story writers*”). In accordance to pre-processing steps suggested in previous work [33], Wikipedia categories containing any of the subphrases *article(s)*, *category(ies)*, *infobox(es)*, *pages*, *redirects*, *stubs*, *tem-*

<i>Target Category: Constituent→Golden Attribute</i>
<i>Sample from CatN evaluation set:</i>
African traditional music: African→continent
African traditional music: traditional→genre
Blues Hall of Fame inductees: Blues Hall of Fame→OutOfWiki
Brazilian footballers: Brazilian→nationality
Cash Money Records albums: Cash Money Records→record label
Casio products: Casio→company
Driving licences: Driving→OutOfWiki
Foodborne illnesses: Foodborne→medium
Komnenos dynasty: Komnenos→OutOfWiki
Mtume albums: Mtume→artist
Roman consuls: Roman→period
<i>Sample from CatL evaluation set:</i>
1980s American animated television series: 1980s→decade
1980s American animated television series: American→country
1980s American animated television series: animated→genre
1980s American animated television series: television→medium
Central Oklahoma Bronchos football coaches: football→sport
Central Oklahoma Bronchos football coaches: Central Oklahoma Bronchos→team
Columbia Aircraft Manufacturing Corporation aircraft: Columbia Aircraft Manufacturing corporation→manufacturer
Costa Rican short story writers: Costa Rican→nationality
Costa Rican short story writers: short story→format
IUCN Red List critically endangered species: critically endangered→conservation status
IUCN Red List critically endangered species: IUCN Red list→OutOfWiki
Proposed liquefied natural gas terminals: Proposed→condition
Proposed liquefied natural gas terminals: liquefied natural gas→commodity
Walt Disney Records video albums: Walt Disney Records→record label
Walt Disney Records video albums: video→medium

Table 2: Sample of entries from the human-annotated evaluation sets. An entry consists of a target category (i.e., Wikipedia category), a constituent of the category, and a golden attribute for the constituent (OutOfWiki=golden attributes that are out-of-Wikipedia, i.e., that are not part of the set of possible attributes Y from categories in the format “ X by Y ” available anywhere in Wikipedia)

plates, *wikiproject*, *use mdy dates*, *lists* and *stubs* are deemed to have internal bookkeeping as sole purpose, and therefore are discarded. The other categories are retained as part of the Wikipedia category network. Traversal of the category network gives mappings from a category to 397 descendant categories on average, that is, child categories or, recursively, child categories of child categories. Despite the relatively high average number of descendants, the median is zero, indicating that most categories have no further child or descendant categories.

A random sample of around 200 million documents in English from a Web snapshot is used to construct a phrase similarity repository following [19, 29]. As shown in Table 1, for various phrases (e.g., “*music*”), the repository contains ranked lists of their most distributionally similar phrases, e.g., [classical music, dance, jazz, classic music, songs, ..., soundtracks, ..].

Evaluation Sets: The precision and recall of the attributes extracted for categories is measured over multiple human-annotated

Evaluation Set	Count		
	Target Categories	Golden Attributes	
	All	All	OutOfWiki
CatN	100	190	21
CatL	50	130	12
(All)	150	320	33

Table 3: Size of evaluation sets. The number of golden attributes is computed across all target categories of an evaluation set (OutOfWiki=golden attributes that are out-of-Wikipedia, i.e., that are not part of the set of possible attributes Y from categories in the format “ X by Y ” available anywhere in Wikipedia)

evaluation sets. Table 2 gives examples of entries in the evaluation sets. Entries consist of a target category, i.e., a Wikipedia category for which attributes must be extracted for its constituents; a manually-selected constituent, which is a textual span over the target category for which an attribute must be extracted; and a manually-assigned golden attribute for the constituent. For example, for the target category “*Costa Rican short story writers*”, the evaluation sets contain as golden attributes “*nationality*” for “*Costa Rican*”, and “*format*” for “*short story*”.

The human-annotation task consists in first identifying the constituents of a target category, then selecting the most relevant golden attribute corresponding to each constituent. The golden attribute is selected by freely navigating the Wikipedia category network. The attribute must be one of the ngrams Y that appear as a suffix in any Wikipedia category in the format “ X by Y ”, anywhere in the category network. There are 1,181 such unique ngrams, which are accessible as a set during human annotation. Occasionally, none of the available ngrams is deemed appropriate as a golden attribute for a particular constituent. For example, no ngram is found to properly convey the desired attribute for the constituents “*State*” or “*3D printed*” within the target categories “*State schools*” and “*3D printed firearms*”. For such constituents, an attribute denoted as out-of-Wikipedia is selected as golden attribute for the constituent topic.

To select the target categories of an evaluation set, the set of Wikipedia categories is filtered to discard categories that contain any of a set of 65 most common prepositions (e.g., *about*, *for*), or a token introducing a relative clause (*that*, *who*). Such prepositions or tokens could otherwise serve as syntactic clues towards delimiting individual constituents from one another, and possibly hint at the roles played by the constituents as well, which would make the overall task easier. Categories containing fewer than two tokens (e.g., *Montreal*) are also discarded. Out of the resulting set of 338,487 categories, an initial random sample of Wikipedia categories is drawn for manual inspection. A Wikipedia category from the selected sample is either retained as a target category, or discarded if it corresponds in fact to an instance (“*PSV Eindhoven*”, “*American Airlines*”). Ideas from [38, 46] might help automate the otherwise manual inspection. Out of the categories drawn in the sample and manually inspected, 14% are discarded. The manual inspection ends once a sample containing the desired number of target categories has been retained from the initial sample. The procedure yields two evaluation sets, which differ from each other with respect to how their respective initial random samples are drawn. In the case of the first evaluation set, **CatN**, the initial random sample is drawn from all available Wikipedia categories. For the second evaluation set, **CatL**, the sample is drawn from Wikipedia categories that contain at least 5 tokens. As summarized in Table 3,

the target categories in the evaluation sets are 100 Wikipedia categories (CatN) and 50 relatively more specific Wikipedia categories (CatL). Overall, the evaluation sets contain a comparable number of entries to gold standards previously introduced for the evaluation of tasks related to compositionality analysis [28].

Parameter Settings: In the first extraction stage, during the comparison of sequences of tokens in a child category vs. parent category, all tokens are converted to lowercase and stemmed using the Porter stemmer [34]. In all stages, pairs of sequences of tokens or ngrams or, more generally, strings are compared after the strings are converted to lowercase. In the third stage, ngrams from a category and a potentially compatible category are considered to be similar to each other, if their distributional similarity score, in the phrase similarity repository collected in advance, is above 0.01. The ngrams are collected around the head tokens of the categories, which are identified with a syntactic parser [31] but could have otherwise been approximated to be the last tokens in the categories.

Baseline Methods: Besides the method proposed here, which is denoted R_{ie} for convenience, several alternative methods are considered.

An alternative method corresponds to deriving attributes as introduced in [24]. Edges in the Wikipedia category network, such as the edge between the category “*Miles Davis albums*” and its parent “*Albums by artist*”, are used in [24] to derive attributes such as “*artist*” for “*Miles Davis*”. The resulting method, which acts as a baseline and is denoted R_{dc} , roughly corresponds to taking the method R_{ie} being proposed here and temporarily modifying it to enable only its first stage, which aligns categories with parent categories; while disabling its remaining stages, which propagate to descendant and to compatible categories.

For completeness, another alternative method takes advantage of relations between Wikipedia articles, as extracted from Wikipedia and already available in WikiNet [25]. The relations in WikiNet (e.g., “*Julieta Pinto: nationality: Costa Rica*”) connect pairs of Wikipedia articles, through predicates (e.g., “*nationality*”) that explicitly state how the first and second article in a pair (the articles titled “*Julieta Pinto*” and “*Costa Rica*”) are related to each other. The alternative method, which acts as a baseline and is denoted R_{nt} , takes WikiNet data as it was graciously made publicly available by its authors, as input; and produces attribute annotations over Wikipedia categories, as output. For this purpose, it propagates WikiNet relations from Wikipedia articles to their parent Wikipedia categories. Concretely, for each relation, it replaces the first article (“*Julieta Pinto*”) in the relation, with each of its parent Wikipedia categories (“*Costa Rican short story writers*”). Simultaneously, all possible consecutive sequences of tokens (“*Costa*”, “*Costa Rican*” etc.) of each parent category, on one hand, are compared against the string descriptions (“*Costa Rica*”) available in WikiNet for the second article (“*Costa Rica*”) in the relation, on the other hand. During comparison, strings are normalized, and adjectives converted into their nominal counterparts (“*Costa Rican*” into “*Costa Rica*”). If the strings match, the sequence of tokens (“*Costa Rican*”) is deemed to be a constituent of the Wikipedia category (“*Costa Rican short story writers*”); whereas the predicate of the relation (e.g., “*nationality*”) is deemed to be a candidate attribute of the constituent. Different child articles (e.g., “*Julieta Pinto*”, “*Carlos Luis Fallas*”) of a Wikipedia category (“*Costa Rican short story writers*”) are involved in different relations in WikiNet (“*Julieta Pinto: nationality: Costa Rica*”; “*Julieta Pinto: organization: Costa Rica*”; “*Carlos Luis Fallas: nationality: Costa Rica*”). Therefore, the same constituent “*Costa Rican*” of a given Wikipedia category may become associated with multiple candidate attributes (“*nationality*”, “*organization*”). For

<i>Target Category: Constituent→Extracted Attribute (Correct?)</i>
Argentine short story writers: Argentine→country (C)
British soul singers: British→nationality (C), soul→genre (C)
Dell laptops: Dell→location (W)
French jazz composers: French→organization (W), jazz→genre (C)
Oklahoma elections: Oklahoma→location (C)
Sri Lankan Buddhist monks: Sri Lankan→nationality (C), Buddhist→religion (C)
Winx Club films: Winx Club→location (W)

Table 4: Attribute annotations extracted by the baseline method R_{nt} for a sample of categories from the evaluation sets (C=correct; W=wrong)

a given category and a given constituent of the category, the candidate attribute with the largest number of contributing child articles (8, in case of the candidate “nationality”) is retained and extracted as an attribute of the constituent (“Costa Rican”) of the category (“Costa Rican short story writers”).

5. EVALUATION RESULTS

Evaluation Procedure: Given an evaluation set, the tuples of a target category from the set, a constituent and an extracted attribute, on one hand; are compared with the tuples of a target category, a golden constituent and a golden attribute, on the other hand. Micro-averaged precision is the ratio of extracted tuples that are golden; and micro-averaged recall is the ratio of golden tuples that are extracted. In contrast, macro-averaged precision and macro-averaged recall are first computed separately for each target category, then averaged over all target categories.

The computed scores are lower or higher, depending on whether the golden attributes marked as out-of-Wikipedia are considered or ignored during the computation. In a somewhat conservative approach, out-of-Wikipedia attributes are considered during the computation, although the proposed method cannot extract them. Therefore, the scores take into account, and are already lowered by the inability of the method to select a relevant attribute for a particular constituent, when in fact there is no such attribute in the set of possible attributes Y from categories in the format “ X by Y ” available anywhere in the Wikipedia category network.

Determining if extracted attributes are golden is automatic, and consists in simple, full string comparison of extracted attributes versus golden attributes. The exception is the baseline method R_{nt} . Its set of extracted attributes is drawn from a set different the gold attributes - namely, from the set of predicates of WikiNet relations rather than the set of possible attributes Y from categories in the format “ X by Y ”. Therefore, it would be unfair to automatically mark an attribute extracted by R_{nt} as wrong during evaluation, only because it is different than its corresponding golden attribute. Instead, only for the baseline method R_{nt} and only for the purpose of evaluation, each attribute extracted by R_{nt} for a constituent of a target category is manually compared to the corresponding golden attribute. If the extracted attribute (“location”) is identical to, or manually found to plausibly approximate the golden attribute (“state”), then the extracted attribute is manually marked as correct for the target category (“Oklahoma elections”), and therefore receives full credit during evaluation. In case of disagreement during the manual comparison, on whether an extracted attribute is or is not a plausible approximation, the more lenient comparison (i.e., it is in fact a plausible approximation) prevails, and the extracted attribute receives full credit. The manual comparison of extracted versus golden attributes is incorporated into the evaluation of R_{nt}

Eval Set	Precision and Recall Scores			
	Score			
	Micro Averaged		Macro Averaged	
	P	R	P	R
CatN	0.886	0.695	0.917	0.691
CatL	0.808	0.746	0.841	0.725
(All)	0.851	0.715	0.898	0.699

Table 5: Precision and recall of attributes extracted for constituents of each target category (i.e., each entry) from the evaluation sets. Computed as micro and macro-averaged scores (P=precision; R=recall)

Sequence of Stages	Precision and Recall Scores					
	Eval Set					
	CatN			CatL		
	P	R	F	P	R	F
Micro-averaged scores:						
S_1	0.924	0.384	0.543	1.000	0.146	0.255
S_1+S_2	0.898	0.605	0.723	0.833	0.500	0.625
S_1+S_3	0.804	0.605	0.690	0.792	0.585	0.673
$S_1+S_2+S_3$	0.886	0.695	0.779	0.808	0.746	0.776
Macro-averaged scores:						
S_1	0.947	0.443	0.604	1.000	0.178	0.302
S_1+S_2	0.928	0.614	0.739	0.895	0.496	0.638
S_1+S_3	0.836	0.611	0.706	0.822	0.591	0.688
$S_1+S_2+S_3$	0.917	0.691	0.788	0.841	0.725	0.779

Table 6: Impact of various extraction stages on the precision and recall of extracted attributes. Computed over each evaluation set, when the three extraction stages, denoted S_1 , S_2 and S_3 , are enabled or temporarily disabled (P=precision; R=recall; F=balanced F-score)

in two steps: (1) manually inspect each extracted attribute of R_{nt} , and potentially replace it with the corresponding golden attribute if the former is identical to or is a plausible approximation of the latter; and (2) automatically compare possibly-replaced extracted attributes and golden attributes, as full strings. Table 4 shows comparison outcomes, for attributes extracted by R_{nt} for a sample of target categories.

The baseline method R_{nt} takes as input static WikiNet data, as derived and made available in [25] from an earlier snapshot of Wikipedia. Some of the target categories from the evaluation sets may be missing from the earlier Wikipedia snapshot, and therefore from WikiNet data. In practice, 15 of 150 (CatN) and 18 of 50 (CatL) of the target categories are missing from WikiNet data. For fairness to R_{nt} , evaluation metrics are computed after excluding the missing categories from the evaluation sets.

Annotation of Attributes over Categories: Table 5 summarizes the accuracy and coverage of the attributes extracted for Wikipedia categories by the proposed method R_{ie} . Micro and macro averaged precision and recall vary from one evaluation set to another. Micro-averaged scores are consistently above 0.800 for precision, and above 0.690 for recall, across evaluation sets. Macro-averaged scores vary between 0.841 (for the CatL evaluation set) and 0.917 (for CatN) in the case of precision, and between 0.691 (for CatN) and 0.725 (for CatL) in the case of recall. Judging from recall scores, the relatively longer categories from the CatL evaluation set are slightly easier to annotate than the relatively shorter categories from CatN, but with lower accuracy as indicated by precision scores.

Isolating the Effect of Attribute Propagation: To quantify the effect of the various extraction stages on the extracted attributes, Table 6 gives detailed micro-averaged and macro-averaged precision

<i>Target Category: Constituent→Extracted Attribute (Stage)</i>
<i>Attributes extracted for a sample of categories from CatN:</i>
African traditional music: African→continent (S ₂), traditional→genre (S ₃)
Arizona Wildcats men’s basketball coaches: Arizona Wildcats→team (S ₂), basketball→sport (S ₂), men’s→type (S ₃)
Blues Hall of Fame inductees: (none extracted)
Brazilian footballers: Brazilian→nationality (S ₂)
Cash Money Records albums: Cash Money Records→record label (S ₁)
Casio products: Casio→company (S ₁)
Komnenos dynasty: (none extracted)
Mtume albums: Mtume→artist (S ₁)
Roman consuls: Roman→culture (S ₂)
<i>Attributes extracted for a sample of categories from CatL:</i>
Central Oklahoma Bronchos football coaches: football→sport (S ₂), Central→country (S ₃), Central Oklahoma Bronchos→team (S ₃), Oklahoma→state (S ₃)
Columbia Aircraft Manufacturing Corporation aircraft: Columbia Aircraft Manufacturing Corporation→manufacturer (S ₁), manufacturing→industry (S ₂)
Costa Rican short story writers: Costa Rican→nationality (S ₁), short story→format (S ₂)
IUCN Red List critically endangered species: critically endangered→conservation status (S ₃), endangered→conservation status (S ₃)
Proposed liquefied natural gas terminals: Proposed→condition (S ₂)
Walt Disney Records video albums: Walt Disney Records→record label (S ₁), Disney→owner (S ₂), video→medium (S ₃)

Table 7: Attribute annotations extracted for a sample of categories from the evaluation sets (S₁, S₂ and S₃ are the first, second and third stages of the extraction method)

and recall scores as well as balanced F-scores over the evaluation sets, when individual stages are enabled or temporarily disabled in ablation experiments. The individual stages, described earlier in Section 3, are denoted in the table as S₁, for the first stage that aligns categories with their parent categories; S₂, for the second stage that propagates from ancestor to descendant categories; and S₃, for the third stage that propagates from categories to compatible categories based on ngram similarity.

The scores of the regular, fully-fledged run, in which all extraction stages S₁+S₂+S₃ are enabled, are shown in the last rows of the upper and lower portions of Table 6. Comparatively, enabling only the first stage S₁ and disabling any propagation in stages two and three, in the first rows of the upper and lower portions of Table 6, gives higher precision, but at a significant loss in coverage. Also enabling either the second S₁+S₂ or third stage S₁+S₃ on top of the first stage, in the second and third rows of the table, increases coverage at the expense of lower precision. When comparing the scores for S₁+S₂ and S₁+S₃ on the CatN evaluation set, both micro-averaged and macro-averaged scores suggest S₁+S₃ does not give as good results as S₁+S₂, when enabled on top of the first stage. Indeed, enabling the second stage increases recall as much as when enabling the third stage, but results in a smaller reduction in precision. Scores over the CatL evaluation appear mixed. The combination S₁+S₃ leads to a higher loss of precision than S₁+S₂, but also higher gain in coverage. But if the first and second stages are enabled S₁+S₂ in the second rows, the option exists to also enable the third stage S₁+S₂+S₃ in the last rows. Doing so achieves not just higher recall than when enabling the first and third stages S₁+S₃ in the third rows, as expected; but also higher precision. The results suggest that enabling all stages gives the best overall results; and that the second stage, between the first and the third stage, is important in achieving the results.

Table 7 shows the attributes extracted for a few categories from the evaluation sets, as well as the stages that extract the attributes. **Comparison to Baseline Methods:** Table 8 compares the proposed method R_{ie} to the alternative, baseline methods. Across all evaluation sets, and for both micro-averaged and macro-averaged scores, the balanced F-scores are uniformly lower for the baseline methods than for the proposed method. For the baseline method R_{dc}, its scores are copied from the scores reported in Table 6, when

<i>Precision and Recall Scores</i>									
<i>Eval Set</i>	<i>Score</i>								
	<i>P</i>			<i>R</i>			<i>F</i>		
	R _{dc}	R _{nt}	R _{ie}	R _{dc}	R _{nt}	R _{ie}	R _{dc}	R _{nt}	R _{ie}
Micro-averaged scores:									
CatN	0.924	0.697	0.886	0.384	0.360	0.695	0.543	0.475	0.779
CatL	1.000	0.615	0.808	0.146	0.293	0.746	0.255	0.397	0.776
(All)	0.962	0.677	0.851	0.265	0.343	0.715	0.399	0.455	0.777
Macro-averaged scores:									
CatN	0.947	0.786	0.917	0.443	0.377	0.691	0.604	0.510	0.788
CatL	1.000	0.642	0.841	0.178	0.277	0.725	0.302	0.387	0.779
(All)	0.960	0.750	0.898	0.377	0.352	0.699	0.541	0.479	0.786

Table 8: Comparative precision and recall of attributes extracted for constituents of each target category (i.e., each entry) from the evaluation sets. Computed for the proposed method (R_{ie}) and for baseline methods R_{dc} and R_{nt} (P=precision; R=recall; F=balanced F-score)

only the first stage S₁ is enabled. The balanced F-scores in Table 8 are lower for the baseline methods R_{dc} and the WikiNet [25]-based R_{nt}, than for the proposed method R_{ie}.

Extraction of Attribute-Value Properties: Attributes computed and assigned by the proposed method to Wikipedia categories, along with the constituents to which they are assigned, are indicative of attribute-value properties that should apply to all instances of the category. To estimate the usefulness of the annotations over categories as attribute-value properties, a random Wikipedia article listed under each target category from the evaluation sets is selected. Separately, out of the attributes extracted for each target category, a random attribute is selected along with the constituent to which it is assigned. Using a set of 6,139 mappings from adjectives to one or more noun phrases collected from data in WordNet, adjectival constituents are converted into their nominal counterparts (“German” into “Germany”). After conversion, the pair of an extracted attribute and its constituent become a candidate attribute-value property that should apply to the instances of the category, including the selected Wikipedia article. The content of the article selected for a category is manually inspected, to ascertain whether a) the article contains evidence that the attribute-value property is

Category	Attribute-Value Property	Article	C?	N?
Costa Rican short story writers	nationality:Costa Rica	Alfonso Chase	Yes	Yes
Dean bass guitars	manufacturer:Dean	Dean Metalman ML	Yes	No
Ice Prince songs	artist:Ice Prince	Aboki (song)	Yes	No
Roman consuls	culture:Roman	Marcus Fabius Dorsuo	No	N/A
South Korean carom billiards players	nationality:South Korea	Sang Lee	Yes	Yes
Swedish football clubs 2000 season	sport:football	2000 Helsingborgs IF season	Yes	Yes

Table 9: Examples of judgments of correctness and novelty for attribute-value properties derived from the attribute annotations of categories, as extracted by the proposed method. The article relative to which the candidate property is judged is shown through its title. A candidate attribute-value property is a pair that consists of the extracted attribute, and the constituent for which the attribute is extracted. Adjectival constituents are converted to nominal form. (C?=attribute-value property is correct?; N?=attribute-value property is not already available in structured form in the article?)

correct; and b) the evidence is available in structured form, as part of the infobox of the article.

Table 9 gives examples of candidate properties derived from attribute annotations over categories, and their judgments relative to articles that belong to the category. Over the larger sample of tuples of a category, an article and a candidate attribute-value property, 84% of the candidate attribute-value properties are judged correct. Out of the properties deemed correct, 42% are judged to not already exist in structured form (including infoboxes) in the respective articles. The results suggest that, in addition to adding a layer of understanding on top of the string surface forms of the categories, attribute annotations are also useful in inferring attribute-value properties that apply to instances of the categories but are not already available in Wikipedia.

Discussion: Several types of errors cause losses in precision or recall. The propagation of attributes in the second and third stage, in this order, significantly increase coverage over the first stage. As expected, the increase in coverage comes at the expense of lower accuracy. For example, the annotation “*Disney*” → “*owner*” is incorrectly propagated in the second stage, from the ancestor category “*Disney franchises*” to the descendant category “*Walt Disney Records video albums*”. In part, such errors are caused by the category network containing edges that are not *IsA* relations [33, 9].

In the first stage, the alignment of the three consecutive sequences of tokens, into which the categories and their parent categories are split, produces attributes of relatively high precision. However, parent categories in the desired format “*X by Y*” are occasionally not aligned to any of their child categories. For example, the parent category “*Mozambican women by occupation*” is not aligned with any of its child categories “*Mozambican women artists*”, “*Mozambican women in politics*”, “*Mozambican sportswomen*” and “*Mozambican women writers*”. Similarly, no alignment is found for the parent category “*Serbian jazz musicians by instrument*” relative to any of its child categories “*Serbian jazz clarinetists*”, “*Serbian jazz*

guitarists”, “*Serbian jazz keyboardists*” or “*Serbian jazz trumpeters*”. While alignment in such cases would be more difficult, the cases suggest that the recall of the first stage could be increased incrementally, by considering the alignment of sequences of tokens other than the second in a child category and the first in a parent category.

Affecting recall more significantly is the absence from Wikipedia of some of the desirable attributes from any parent categories in the format “*X by Y*”, let alone from alignments of such parent categories with their child categories. If an attribute desirable for a constituent in a category is not available for the same constituent or even for any other constituents in other categories, then the attribute cannot be potentially propagated among categories. For example, the second and third stages cannot potentially propagate and assign an attribute to the constituent “*vacuum tube*” in the category “*IBM vacuum tube computers*” from the same constituent in other categories, because the first stage does not assign any attribute to “*vacuum tube*” in any other category. Similarly, the hypothetical attribute “*website*” cannot be potentially propagated to the constituent “*YouTube*” in the category “*YouTube videos*”, if no “*X by Y*” category exists in Wikipedia that allows “*website*” to be collected from the placeholder *Y*.

The absence of some desirable attributes, from the set of possible attributes that can be extracted, places an undesirable upper limit on extraction recall. But the phenomenon is just one of the factors contributing towards overall extraction recall. Unlike in the method R_{ie} proposed here, attributes extracted by the baseline method R_{nt} , from among predicates of WikiNet relations, are not restricted to possible attributes *Y* from categories in the format “*X by Y*”. Nevertheless, R_{ie} has not just higher recall than the baseline R_{nt} , but also higher precision.

6. RELATED WORK

Analysis of Compositional Noun Phrases: The extraction of attributes for constituents of Wikipedia categories implies that the categories are compositional noun phrases. Studies on the role of compositionality in understanding noun phrases and other phrases share the view that the semantics of a concept denoted by a compositional phrase is effectively defined by, and can be computed from, the semantics of the concepts denoted by its constituent words [22, 37]. Conversely, since the meaning of non-compositional noun phrases such as “*hot dogs*” and “*red tape*” has little to do with the meaning of their constituent words, it is useful to identify non-compositional phrases [17], to rule out any subsequent attempts to possibly translate those phrases based on their words.

If a noun phrase is known to be compositional, lexical interpretations (“*from*”, “*born in*”) can be extracted from text to explain the role that a modifier (“*Costa Rican*”) plays in the larger noun phrase (“*Costa Rican writers*”) [23]. Noun phrases are often assumed to contain one modifier and one head, for a total of only two constituent phrases [27], or sometimes even only two words [15]. In comparison, the extraction method proposed here accommodates longer phrases with more than two modifier constituents, such as “*Costa Rican short story writers*”. It is also complementary, in that it extracts attributes (“*nationality*”) rather than interpretations (“*born in*”) for modifier constituents.

Switching focus from ambiguous to already-disambiguated compositional noun phrases, the method introduced in [28] decomposes compositional Freebase topics (“*Swiss passport*”) into constituent Freebase topics (*Switzerland*, *Passport*). Because it requires the input items to be Freebase topics, it is not applicable to class labels or Wikipedia categories, most of which have no equivalent topics in Freebase or articles in Wikipedia. Even if the method were appli-

cable to Wikipedia categories, the disambiguation of constituents (“chess”) into Freebase topics (*Chess (sport)*) alone would still not necessarily distinguish between the different roles (“topic”, “sport”) of the same constituent with the same disambiguation in different Wikipedia categories (“*Chess books*”, “*Chess coaches*”). Methods for semantic parsing of noun phrases [4] do apply to compositional noun phrases including class labels and Wikipedia categories. They produce logical forms that connect related constituents to one another, but do not annotate constituents with attributes corresponding to their roles.

Open-Domain Information Extraction: Within the larger area of open-domain information extraction [8, 21, 13], a significant body of work [1, 16, 5, 30, 36] acquires large repositories of pairs of an instance and a lexical class label from arbitrary text or, less frequently, from Wikipedia [9]. The lexical class labels could benefit from an extension of our method to operate beyond Wikipedia categories. The annotation of class labels with attributes would be a significant step away from flat string representations of class labels, towards more structured, semantic representations.

Extraction from Wikipedia: Previous studies already illustrate Wikipedia’s role in knowledge acquisition [24, 45, 43, 13] and information retrieval [14, 35]. In [25], the analysis of Wikipedia categories relies on data from infoboxes rather than the category network, and does not attempt to systematically identify the role of constituents within categories. Alternatively, [24] proposes the use of edges in the Wikipedia category network, such as the edge between the category “*Miles Davis albums*” and its parent “*Albums by artist*”, to derive attributes such as “*artist*” for “*Miles Davis*”. Our method builds upon the same intuition, scaling it up to assign more attributes to more constituents of more Wikipedia categories. Associating Wikipedia categories with attributes is potentially useful to knowledge repositories built upon Wikipedia [2, 13, 39]; and to extraction methods operating over Wikipedia data [43, 8, 13, 41, 28].

Expansion of Knowledge Repositories: Attributes extracted for Wikipedia constituents correspond to attribute-value properties (e.g., “*nationality: Costa Rica*”) that may be missing for at least some of the instances (“*Julietta Pinto*”) of Wikipedia categories (“*Costa Rican short story writers*”). Other work aims specifically at expanding knowledge repositories, where the expansion fills in new relations for existing properties of existing topics [43, 7, 42], fills in new types for existing topics [26], or adds new topics [20, 12].

7. CONCLUSION

Open-domain information extraction tends to treat class labels vs. attribute-value properties as knowledge to be separately extracted for various instances. As a counterpart to that strategy, this paper proposes a connected view of class labels and attributes. Class labels and attributes of an instance are not just inter-connected, but constituents in relevant class labels are often indicative of relevant attributes and vice-versa. In an application of the connected view to Wikipedia, attribute annotations over Wikipedia categories provide a layer of understanding beyond the string surface of the categories. It is a step towards decoding the meaning of tens or hundreds of thousands of lexical class labels within human curated repositories like Wikipedia in particular, or automatically extracted repositories in general. Possible extensions of the work explore:

(a) the role of infoboxes of Wikipedia articles listed under a category, as an additional source of candidate attributes for annotating the category;

(b) the potential for using Wikipedia ancestor categories as an alternative source of candidate attributes, for example using the ancestors of “*American record labels*” or “*Record labels*” of the cat-

egory “*Walt Disney Records*”, as candidate attributes for the constituent “*Walt Disney Records*” within the category “*Walt Disney Records video albums*”;

(c) more generally, the potential for using hypernyms from separate repositories extracted from text, as an alternative source of candidate attributes, especially for constituents with no candidate attributes currently extracted from any category in the Wikipedia category network;

(d) learning an annotation model for Wikipedia categories that can be transferred to annotate lexical class labels from repositories other than Wikipedia, regardless of whether such repositories contain any class labels in the format “*X by Y*”;

(e) using decompositions of categories into disambiguated constituents, to further improve the annotations; and

(f) besides the annotation of existing class labels with new attributes, the complementary direction of deriving new class labels from existing attributes.

8. ACKNOWLEDGMENTS

The author would like to thank Susanne Riehemann for comments on an earlier version of the paper.

9. REFERENCES

- [1] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the Web. In *IJCAI-07*, pages 2670–2676, 2007.
- [2] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - a crystallization point for the Web of data. *Journal of Web Semantics*, 7(3):154–165, 2009.
- [3] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. Hruschka, and T. Mitchell. Toward an architecture for never-ending language learning. In *AAAI-10*, pages 1306–1313, 2010.
- [4] E. Choi, T. Kwiatkowski, and L. Zettlemoyer. Scalable semantic parsing with partial ontologies. In *ACL-15*, pages 1311–1320, 2015.
- [5] B. Dalvi, W. Cohen, and J. Callan. Websets: Extracting sets of entities from the Web using unsupervised information extraction. In *WSDM-12*, pages 243–252, 2012.
- [6] D. Davidov and A. Rappoport. Extraction and approximation of numerical attributes from the Web. In *ACL-10*, pages 1308–1317, 2010.
- [7] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, and K. Murphy. Knowledge Vault: A Web-scale approach to probabilistic knowledge fusion. In *KDD-14*, pages 601–610, 2014.
- [8] O. Etzioni, A. Fader, J. Christensen, S. Soderland, and Mausam. Open information extraction: The second generation. In *IJCAI-11*, pages 3–10, 2011.
- [9] T. Flati, D. Vannella, T. Pasini, and R. Navigli. Two is bigger (and better) than one: the Wikipedia Bitaxonomy project. In *ACL-14*, pages 945–955, 2014.
- [10] M. J. P. G. Angeli and C. Manning. Leveraging linguistic structure for open domain information extraction. In *ACL-15*, pages 344–354, 2015.
- [11] M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING-92*, pages 539–545, 1992.
- [12] J. Hoffart, Y. Altun, and G. Weikum. Discovering emerging entities with ambiguous names. In *WWW-14*, pages 385–396, 2014.
- [13] J. Hoffart, F. Suchanek, K. Berberich, and G. Weikum. 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, 2013.
- [14] J. Hu, G. Wang, F. Lochovsky, J. Sun, and Z. Chen. Understanding user’s query intent with Wikipedia. In *WWW-09*, pages 471–480, 2009.
 - [15] N. Kim and P. Nakov. Large-scale noun compound interpretation using bootstrapping and the Web as a corpus. In *EMNLP-11*, pages 648–658, 2011.
 - [16] Z. Kozareva, K. Voevodski, and S. Teng. Class label enhancement via related instances. In *EMNLP-11*, pages 118–128, 2011.
 - [17] D. Lin. Automatic identification of non-compositional phrases. In *ACL-99*, pages 317–324, 1999.
 - [18] D. Lin and P. Pantel. Concept discovery from text. In *COLING-02*, pages 1–7, 2002.
 - [19] D. Lin and X. Wu. Phrase clustering for discriminative learning. In *ACL-IJCNLP-09*, pages 1030–1038, 2009.
 - [20] T. Lin, Mausam, and O. Etzioni. No noun phrase left behind: Detecting and typing unlinkable entities. In *EMNLP-CoNLL-12*, pages 893–903, 2012.
 - [21] Mausam, M. Schmitz, S. Soderland, R. Bart, and O. Etzioni. Open language learning for information extraction. In *EMNLP-CoNLL-12*, pages 523–534, 2012.
 - [22] J. Mitchell and M. Lapata. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429, 2010.
 - [23] P. Nakov and M. Hearst. Semantic interpretation of noun compounds using verbal and other paraphrases. *ACM Transactions on Speech and Language Processing*, 10(3):1–51, 2013.
 - [24] V. Nastase and M. Strube. Decoding Wikipedia categories for knowledge acquisition. In *AAAI-08*, pages 1219–1224, 2008.
 - [25] V. Nastase and M. Strube. Transforming Wikipedia into a large scale multilingual concept network. *Artificial Intelligence*, 194:62–85, 2013.
 - [26] A. Neelakantan and M. Chang. Inferring missing entity type instances for knowledge base completion: New dataset and methods. In *NAACL-HLT-15*, 2015.
 - [27] M. Paşca. Interpreting compound noun phrases using Web search queries. In *NAACL-HLT-15*, pages 335–344, 2015.
 - [28] M. Paşca and H. Buisman. Dissecting German grammar and Swiss passports: Open-domain decomposition of compositional entries in large-scale knowledge repositories. In *IJCAI-15*, pages 896–902, 2015.
 - [29] P. Pantel, E. Crestan, A. Borkovsky, A. Popescu, and V. Vyas. Web-scale distributional similarity and entity set expansion. In *EMNLP-09*, pages 938–947, 2009.
 - [30] P. Pasupat and P. Liang. Zero-shot entity extraction from Web pages. In *ACL-14*, pages 391–401, 2014.
 - [31] S. Petrov, P. Chang, M. Ringgaard, and H. Alshawi. Uptraining for accurate deterministic question parsing. In *EMNLP-10*, pages 705–713, 2010.
 - [32] S. Ponzetto and R. Navigli. Large-scale taxonomy mapping for restructuring and integrating Wikipedia. In *IJCAI-09*, pages 2083–2088, 2009.
 - [33] S. Ponzetto and M. Strube. Deriving a large scale taxonomy from Wikipedia. In *AAAI-07*, pages 1440–1447, 2007.
 - [34] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
 - [35] U. Scaiella, P. Ferragina, A. Marino, and M. Ciarmita. Topical clustering of search results. In *WSDM-12*, pages 223–232, 2012.
 - [36] V. Shwartz, Y. Goldberg, and I. Dagan. Improving hypernymy detection with an integrated path-based and distributional method. In *ACL-16*, pages 2389–2398, 2016.
 - [37] R. Socher, J. Bauer, C. Manning, and A. Ng. Parsing with compositional vector grammars. In *ACL-13*, pages 455–465, 2013.
 - [38] F. Suchanek, G. Kasneci, and G. Weikum. YAGO: a core of semantic knowledge unifying WordNet and Wikipedia. In *WWW-07*, pages 697–706, 2007.
 - [39] D. Vrandečić and M. Krötzsch. Wikidata: A free collaborative knowledge base. *Communications of the ACM*, 57:78–85, 2014.
 - [40] C. Wang, K. Chakrabarti, Y. He, K. Ganjam, Z. Chen, and P. Bernstein. Concept expansion using Web tables. In *WWW-15*, pages 1198–1208, 2015.
 - [41] Z. Wang, Z. Li, J. Li, J. Tang, and J. Pan. Transfer learning based cross-lingual knowledge extraction for Wikipedia. In *ACL-13*, pages 641–650, 2013.
 - [42] R. West, E. Gabrilovich, K. Murphy, S. Sun, R. Gupta, and D. Lin. Knowledge base completion via search-based question answering. In *WWW-14*, pages 515–526, 2014.
 - [43] F. Wu and D. Weld. Open information extraction using Wikipedia. In *ACL-10*, pages 118–127, 2010.
 - [44] W. Wu, H. Li, H. Wang, and K. Zhu. Probase: a probabilistic taxonomy for text understanding. In *SIGMOD-12*, pages 481–492, 2012.
 - [45] Y. Yan, N. Okazaki, Y. Matsuo, Z. Yang, and M. Ishizuka. Unsupervised relation extraction by mining Wikipedia texts using information from the Web. In *ACL-IJCNLP-09*, pages 1021–1029, 2009.
 - [46] C. Zirn, V. Nastase, and M. Strube. Distinguishing between instances and classes in the Wikipedia taxonomy. In *ESWC-08*, pages 376–387, 2008.