Dissecting German Grammar and Swiss Passports: Open-Domain Decomposition of Compositional Entries in Large-Scale Knowledge Repositories

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

Hylke Buisman
Google Inc.
Brandschenkestrasse 110
8002 Zurich, Switzerland
hylke@google.com

Abstract

This paper presents a weakly supervised method that decomposes potentially compositional topics (*Swiss passport*) into zero or more constituent topics (*Switzerland*, *Passport*), where all topics are entries in a knowledge repository. The method increases the connectivity of the knowledge repository and, more importantly, identifies the constituent topics whose meaning can be later aggregated into the meaning of the compositional topics. By exploiting evidence within Wikipedia articles, the method acquires constituent topics of Freebase topics at precision and recall above 0.60, over multiple human-annotated evaluation sets.

1 Introduction

Motivation: Human-compiled encyclopedic resources such as Wikipedia [Remy, 2002] organize information into articles on topics pertaining to a seemingly-unbound array of domains of interest. Large-scale knowledge repositories [Bollacker *et al.*, 2008] often initially draw from unstructured and semi-structured text within such articles [Bollacker *et al.*, 2008]. They structure knowledge into graphs containing millions of topics connected by billions of relations or facts.

Challenges associated with the automatic extraction and manual curation of continuously-evolving knowledge cause relevant topics [Lin *et al.*, 2012; Hoffart *et al.*, 2014], and especially relations among topics [Wu and Weld, 2010; West *et al.*, 2014], to be missing from the repositories. Missing relations prevent a topic from being connected to some or all other topics that are essential to understanding its main properties and, ultimately, its meaning. Such "insular" topics are effectively lost, and become "invisible" to any mechanisms that rely on those connections between topics to compute features or, more interestingly, perform inference tasks such as answering queries over structured data [Yao and Van Durme, 2014]. For example, topics such as *German grammar*, *Swiss passport* and *2013-14 Arsenal F.C. season* are effectively disconnected in Freebase [Bollacker *et al.*, 2008] from other topics that are essential to understanding their meaning, such as *German* and *Grammar*, *Switzerland* and *Passport*, or *Arsenal F.C.* and *Season (sports)* etc. More generally, knowledge repositories such as Freebase make no systematic attempt to capture and connect compositional topics, on one hand, with their constituent topics, on the other hand.

Contributions: The weakly supervised method introduced in this paper identifies the constituent topics, if any, whose meaning can be aggregated to define the meaning of a potentially compositional [Downing, 1977] topic. The method decomposes topics into zero (e.g., for *Frank Zappa*), one or more (*Switzerland* and *Passport*, for *Swiss Passport*) constituent topics. The input topic and its extracted constituent topics, if any, are entries in a knowledge repository. Unlike in previous analyses of compositional noun phrases [Nakov and Hearst, 2013; Van de Cruys *et al.*, 2013], the extracted constituents are not ambiguous string entries in the lexical space (e.g., *"German"* for *"German grammar"*). Instead, they are disambiguated entries in the semantic space: *German* is the language rather than the country or nationality for *German grammar*; and *Season* is the season in sports rather than in meteorology or fashion for *2013-14 Arsenal F.C. season*. In contrast to previous work, the constituents are not extracted as a side effect of some other task - namely, the task of extracting lexicalized interpretations (*"issued by"*) of the role that modifiers (*"Swiss"*) play relative to the heads (*"passport"*) of the larger noun phrases [Hendrickx *et al.*, 2013]. To decompose Freebase topics into their constituent topics, if any, the method exploits evidence available within Wikipedia articles aligned with the topics.

2 Decomposition into Constituent Topics

Problem Definition: Let T_F be an entry in a knowledge repository such as Freebase, referring to a topic that may be a relatively generic concept (*Swiss passport*), or a specific instance (*2013-14 Arsenal F.C. season*). The task being addressed is the decomposition of the entry T_F into zero or more other constituent topics C_F (*Swiss*, *Passport*). The constituents C_F compositionally [Downing, 1977] define the meaning of the larger topic T_F [Mitchell and Lapata, 2010].

Intuitions: The open-domain decomposition of topics relies on several intuitions. First, many topics encoded in entries in knowledge repositories are also described in equivalent articles in encyclopedic resources. Indeed, millions of the Freebase topics have equivalent articles in Wikipedia. Second, as more human editors contribute more over time, articles in encyclopedic resources are likely to contain mentions of the constituent topics of the topic being described in the articles.

Indeed, useful encyclopedic articles make at least brief mentions to most, if not all, of the defining properties of topics.

Acquisition from Wikipedia: The open-domain extraction method proposed in this paper exploits features derived from Wikipedia, in order to decompose Freebase topics T_F into their constituent topics C_F. Titles of Wikipedia articles serve as canonical names of the equivalent Freebase topics being decomposed. The decomposition of a Freebase topic operates over the equivalent Wikipedia article, and consists of several **stages**: (1) map the Freebase topic T_F into its equivalent Wikipedia article T_W; (2) extract a noisy set of pairs of a related Wikipedia article R_W and an associated string descriptor D_W, using evidence from the input article T_W; (3) match the string descriptors D_W against input spans (i.e., ngrams) from the article title, to select a subset of candidate constituent articles C_W out of the larger set of related articles R_W; (4) rank candidate constituent articles C_W associated with the same input span, to select the best candidate constituent of each input span, as the computed set of constituent articles of the input article T_W; and (5) map constituent articles C_W from Wikipedia, back into their equivalent constituent topics C_F from Freebase.

Traversal of the Wikipedia Category Network: Each Wikipedia article is connected to its ancestor Wikipedia categories, which are parent categories (e.g., *English football clubs 2013-14 season*) listed at the bottom of the article (*2013-14 Arsenal F.C. season*) or, recursively, parent categories of parent categories (*Seasons*, *Structure* and many others). Together, the connections among categories form the Wikipedia category network [Ponzetto and Strube, 2007].

For each Wikipedia category in the category network, a set of representative articles is computed. A representative article of a category has the category as one of its parent categories; and its title, after normalization, is identical to the name of the category. Normalization may include operations such as morphological normalization, stemming and conversion to lowercase. The set of representative articles of a category may be empty. For example, the representative articles computed for the categories *Seasons* and *Passports* are {*Season*, *Season (society)*, *Season (sports)*} and {*Passport*} respectively.

Extraction of Related Articles and Their Descriptors: From the content of the article T_W, a noisy set of pairs of a potentially related article R_W and a descriptor for R_W is collected from two types of features. First, each outgoing internal link in T_W, that is, each hyperlink from T_W to another Wikipedia article R_W, is taken as evidence that R_W may be related to T_W. The anchor text of the hyperlink, and the title itself of R_W, are collected as descriptors of the related article R_W. For example, outgoing internal links in the article *2013-14 Arsenal F.C. season* produce candidate related articles that include *Arsenal F.C.* (with the descriptor *"Arsenal"*), *2014 FA Cup Final* (descriptor *"2014 final"*) and *Nike, Inc.* (descriptor *"Nike"*). Second, the parent Wikipedia categories of the article T_W recursively lead to ancestor categories, which in turn lead to other related articles. Specifically, each representative article R_W of each ancestor category of T_W is collected as a related article of T_W. The title of the article R_W is also its descriptor. Candidate related articles collected via parent categories may include *Arsenal F.C.* (descriptor *"Arsenal F.C."*), *Season (society)*, *Season (sports)* and *Structure*.

Extraction of Candidate Constituents: After the canonical name of topic T_F, i.e., the title of the article T_W, is split into all possible input spans (i.e., ngrams), each input span is compared to each descriptor D_W of a related constituent article R_W. If an input span and a descriptor are identical after normalization of the strings, then the related article R_W is selected as a candidate constituent C_W of T_W. For example, the descriptors *"Season (society)"*, *"Season (sports)"* and *"Arsenal F.C."* match the input spans *"season"*, *"season"* and *"Arsenal F.C."* respectively. Therefore, related articles such as *Season (society)*, *Season (sports)* and *Arsenal F.C.* are among those selected as candidate constituents of *2013-14 Arsenal F.C. season*.

Ranking of Candidate Constituents: When multiple constituent articles C_W are associated with the same input span from the title of the article T_W, the constituents are ranked relative to another, based on their individual semantic similarity to T_W. The semantic similarity between two Wikipedia articles can be computed using the Wikipedia category network [Strube and Ponzetto, 2006], or comparing the vectors of tokens and other features from the two articles. The most semantically similar constituent article, if any, is retained for each input span of the article T_W being decomposed. For example, among the candidate constituents *Season (society)* and *Season (sports)* associated with the input span *"season"*, the latter is more similar to *2013-14 Arsenal F.C. season*. After mapping the Wikipedia article T_W and its Wikipedia constituent articles C_W back into their equivalent Freebase topics T_F and C_F, the decomposition of the topic T_F is the set of constituent topics C_F, if any.

3 Experimental Setting

Raw Data Sources: The experiments rely on English entries in snapshots as of January 2015 for three raw data sources, namely Freebase, Wikipedia and WordNet [Fellbaum, 1998].

Derived Data Sources: The experiments use only a fraction of the data available for each entry in the raw data sources.

The entire Freebase snapshot is reduced to a set of 4,301,415 mappings from a Freebase topic id, also referred to as a Freebase mid [Bollacker *et al.*, 2008] (e.g., */m/02wvpn5*), to the title (e.g., *Swiss passport*) of the Wikipedia article (e.g., http://en.wikipedia.org/wiki/Swiss_passport) from which the Freebase topic was initially created. Freebase topics not mapped to any Wikipedia articles are discarded.

The raw content of Wikipedia articles is distilled into mappings from an article title (e.g., *2013-14 Arsenal F.C. season*) to: a) 21 outgoing internal links on average, collected as pairs of the anchor text (*Arsenal*) in the article and the title of the target Wikipedia article (*Arsenal F.C.*); and b) 8 parent categories on average, collected as Wikipedia categories listed at the bottom of the article (e.g., *Arsenal F.C. seasons*). Consistent with treatment in previous work [Ponzetto and Strube, 2007], Wikipedia categories containing any of the subphrases *article(s)*, *category(ies)*, *infobox(es)*, *pages*, *redirects*, *stubs*, *templates*, *wikiproject* and *use mdy dates* are deemed to have internal bookkeeping as sole purpose, and therefore are discarded. The other categories are retained as

Evaluation Target Topic		Golden Constituent Topics	
Sample from WikiA:			
/m/0m_j	Amino acid	*/m/0pq0*	Amine
		/m/0hqs	Acid
/m/01lyc2	DNA microarray	*/m/011b83nf*	Microarray
		/m/026w5	DNA
/m/01j2fy	German grammar	*/m/039dj*	Grammar
		/m/04306rv	German language
/m/0fjms	Jabba the Hutt	*/m/067m2n*	Hutt (Star Wars)
/m/05rfpk	Nursing theory	*/m/05fh2*	Nursing
		/m/07kk5	Theory
/m/02pvbny	Sneezing powder	*/m/01hsr_*	Sneeze
		/m/0gzznm	Powder (substance)
/m/0102lx22	Trevor Moran	*(None)*	(None)
Sample from SemE:			
/m/0793nt	Rubber glove	*/m/09kmv*	Natural rubber
		/m/0174n1	Glove
/m/0hpg7	Blood cell	*/m/019jw*	Blood
		/m/01cbd	Cell (biology)
Sample from ComP:			
/m/011bm38q	Bank card	*/m/017ql*	Bank
		/m/09vh0m	Payment card
/m/01xzx	Computer vision	*/m/01m3v*	Computer
		/m/01k1vd	Visual perception

Table 1: Sample of entries from the human-annotated evaluation sets. An entry consists of a target topic and zero or more golden constituent topics that compositionally define the meaning of the topic. The target topics and golden constituent topics are uniquely identified via their Freebase mids. For clarity, the table shows the title of the equivalent Wikipedia article next to each *mid*.

part of the Wikipedia category network. Traversal of the category network gives mappings from an article title to 1,450 ancestor categories on average, that is, parent categories or, recursively, parent categories of parent categories. Separately, the text within each Wikipedia article is converted into a word embedding vector [Weston *et al.*, 2013]. The semantic similarity for a pair of Wikipedia articles is approximated as the cosine similarity between their word embedding vectors. A higher score, e.g., higher for (*2013-14 Arsenal F.C. season*, *Season (sports)*) vs. for (*2013-14 Arsenal F.C. season*, *Season (society)*), is indicative of higher similarity between the articles in a pair.

The relations labeled Derivationally-Related, Value-Of and Related-Noun in WordNet are the source of mappings from 6,139 adjectives to one or more noun phrases (e.g., (*"Swiss"*→*"Switzerland"*), (*"olfactory"*→*"olfaction"*)).

Parameter Settings: During the collection of outgoing internal links of a Wikipedia article, only the first 100 links, in the order in which they occur in the article, are considered. String comparison is performed after the strings are normalized through removal of portions within parentheses (e.g., *"Season (society)"* to *"Season"*), stemmed [Porter, 1980] and converted to lowercase.

Evaluation Sets: Several human-annotated evaluation sets serve the purpose of computing precision and recall of the extracted constituent topics. Entries in the evaluation sets uniformly consist of a target topic, i.e., a Freebase topic whose constituents, if any, must be extracted; and a set of zero or more manually-selected golden constituent topics, i.e, Freebase topics that compositionally define the meaning of the target topic. As shown in Table 1, a truly compositional target topic is associated with the golden constituents from which its meaning can be compositionally derived. In contrast, a target topic deemed non-compositional has no golden constituents in the evaluation set. All topics in the evaluation sets are represented as Freebase mids. Over a sample of 100 target topics, whose golden constituent components were manually annotated by two annotators, the inter-annotator agreement was 85.2%.

Evaluation Set	Count			
	Target Topics			Constituents
	All	Empty	Non Empty	All
WikiA	250	133	117	371
SemE	32	0	32	53
ComP	14	0	14	26

Table 2: Size of evaluation sets, computed along multiple dimensions. The number of golden constituent topics is computed across all target topics of an evaluation set (Empty=number of target topics with no golden constituent topics; Non Empty=number of target topics with one or more golden constituent topics).

The evaluation sets differ from one another with respect to the origin of their target topics. In the first evaluation set, **WikiA**, target topics are a random sample of Freebase topics for which some Wikipedia article is available. The selection of target topics in the second and third evaluation sets, **SemE** and **ComP**, starts from 181 and 35 compound noun phrases introduced in [Hendrickx *et al.*, 2013] and [Davis, 1997] respectively to evaluate the alternative task of extracting paraphrases (e.g., *"machine used for washing clothes"*) of compound noun phrases (*"washing machine"*). For both SemE and ComP, the Freebase topic, if any, corresponding to the desired sense of each compound noun phrase is identified manually. Only when a matching Freebase topic is identified manually for a compound noun phrase, the topic is added as a target topic to the SemE or ComP evaluation sets. For example, */m/01xzx* (*Computer vision*) is added as a target topic for the compound noun phrase *"computer vision"*; */m/01_m9b* (*Rock candy*) but not */m/098l9z* (*Rock Candy* the song) is added for *"rock candy"*; and no target topics are added for the phrases *"liberation struggle"* or *"amateur record"*. As shown in Table 2, the SemE and ComP evaluation sets are comparatively smaller, whereas WikiA is larger. In addition, while all target topics in SemE and ComP have one or more golden constituent topics, some of the target topics in WikiA are non-compositional, and therefore have no golden constituents in the evaluation set. Overall, WikiA contains more entries than gold standards previously introduced for other tasks related to compositionality [Hendrickx *et al.*, 2013].

4 Evaluation Results

Evaluation Procedure: Let S be a subset of entries from an evaluation set, relative to which the performance of extracted

Score	Evaluation Set		
	WikiA	SemE	ComP
As an average over set of golden constituents from evaluation set:			
Average per-constituent precision	0.757	0.777	0.823
Average per-constituent recall	0.765	0.622	0.583
As an average over set of entries from evaluation set:			
Average per-entry precision	0.861	0.671	0.821
Fraction of target entries with max precision	0.819	0.625	0.785
Average per-entry recall	0.852	0.625	0.642
Fraction of target entries with max recall	0.796	0.531	0.428

Table 3: Precision and recall of constituent topics extracted for each target topic (i.e., each entry) from the evaluation sets. Computed as average per-constituent and per-entry scores.

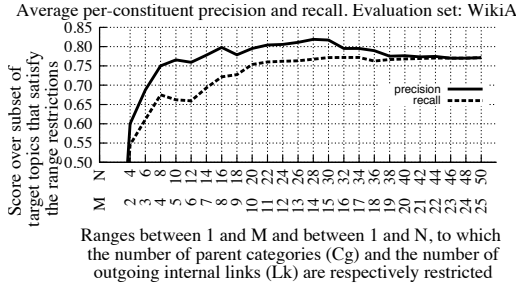

Figure 1: Impact of amount of evidence available through various features from Wikipedia, on per-constituent precision and recall of constituent topics extracted for target topics from the WikiA evaluation set. Computed over variable subsets of target topics. Each subset is obtained from the set of target topics from WikiA, by restricting to those target topics whose Wikipedia articles have a number of parent categories (Cg) and a number of outgoing internal links (Lk) that fall within certain ranges.

constituent topics must be evaluated. As a reminder, an entry consists of a target topic and its golden constituent topics, if any. A pre-requisite to estimating accuracy and coverage is to compute intermediate precision and recall metrics, by comparing the combined set of constituent topics extracted for all target topics from S, with the combined set of golden constituent topics available for all target topics from S:

- $P_S = |E \wedge G|_S \,/\, (|E \wedge G|_S + |E \wedge \neg G|_S)$
- $R_S = |E \wedge G|_S \,/\, (|E \wedge G|_S + |\neg E \wedge G|_S)$

where E, ¬E are constituent topics that are or are not extracted for a particular target topic; G, ¬G are constituent topics that are or are not among the golden constituent topics of the particular target topic; and $|C|_S$ is the count of constituent topics over S that satisfy a particular constraint C, such as the count of constituent topics that are extracted but are not golden constituent topics in the case of $|E \wedge \neg G|_S$. In other words, intermediate precision is the ratio of extracted constituent topics that are golden; and intermediate recall is the ratio of golden constituent topics that are extracted. If no constituent topics are extracted and no golden constituent topics are present over all entries from S, the target topics in all entries from S are necessarily non-compositional and (correctly) no constituent topics are extracted for any entry. In this case, the intermediate metrics P_S and R_S are both set to the maximum value, i.e., 1.

From the intermediate metrics from above, two types of scores are computed. In the case of **per-constituent** scores, intermediate precision and recall are used directly:

- Per-constituent-Precision$_S$ = P_S
- Per-constituent-Recall$_S$ = R_S

where S could be one of the evaluation sets WikiA, SemE or ComP. In contrast, **per-entry** scores apply intermediate metrics individually for subsets S_i that each contain a different entry from S; then average the values over the number of subsets, which is identical to the number of entries in S:

- Per-entry-Precision$_S$ = $(\sum_{i=1..|S|} P_{Si}) \,/\, |S_i|$
- Per-entry-Recall$_S$ = $(\sum_{i=1..|S|} R_{Si}) \,/\, |S_i|$

In particular, per-constituent and per-entry scores over entire evaluation sets can be computed by substituting one of the WikiA, SemE or ComP sets for S.

Precision and Recall: Table 3 summarizes the accuracy and coverage of the constituent topics extracted for the evaluation sets. Five conclusions can be drawn from the results. First, as shown in the upper part of the table, average per-constituent precision is above 0.75 across all evaluation sets. Second, WikiA contains some non-compositional target topics whereas ComP and SemE do not. Nevertheless, the presence of non-compositional target entries in WikiA does not cause significant reductions in per-constituent precision for WikiA, relative to precision for SemE or ComP. Thus, the method is able to relatively accurately extract constituent topics when it should, as it is able to not extract any constituent topics when it should not. Third, the average per-constituent recall is lowest (i.e., 0.583) for ComP and highest (0.765) for WikiA. The scores confirm that it is more difficult to extract all relevant constituents for truly compositional topics than for non-compositional topics. Fourth, in the lower part of the table, per-entry scores vary more significantly from one evaluation set to another. Fifth, at least two thirds (i.e., 0.671 for SemE) of the target entries have extracted constituents that are all correct; and all golden constituents are extracted for at least two fifths (0.428, for ComP) of the target entries.

Strength of Evidence from Wikipedia: The ability of the proposed method to decompose topics into their constituents depends on whether the method does or does not have access to enough of the evidence it requires. The evidence takes the form of the two features from Wikipedia, namely parent categories and outgoing internal links. Figure 1 quantifies the impact on per-constituent precision and recall, when restricting the set of target entries from WikiA to only the target topics whose number of parent categories and outgoing links fall under certain ranges. For example, in the leftmost point on the horizontal axis in the figure, the scores are computed over the subset of target topics from WikiA whose equivalent Wikipedia articles contain between 1 and 2 parent categories, and between 1 and 4 outgoing internal links. At the rightmost point, the scores are computed over target topics from WikiA for which the respective ranges are between 1 and 25, and between 1 and 50 respectively. Access to only a limited amount of evidence, at the leftmost point in the graph, gives lower precision and recall. Moving from the leftmost point in the graph towards the right, as the amount of evidence increases,

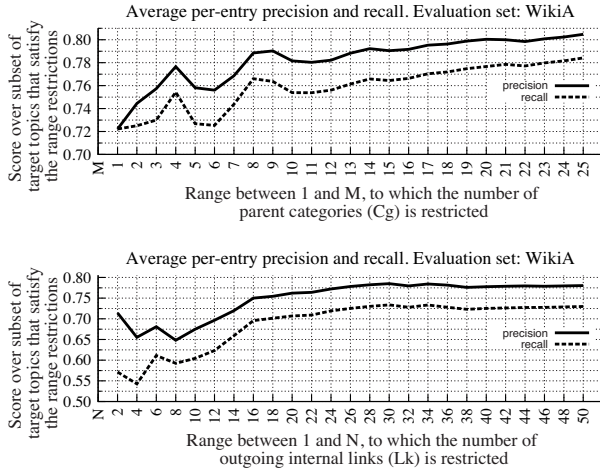

Figure 2: Impact of amount of evidence available through various features from Wikipedia, when the other features are temporarily not allowed to contribute, on per-entry precision and recall of constituent topics extracted for the WikiA evaluation set. Computed over variable subsets of target topics. Each subset is obtained from the set of target topics from WikiA, by restricting to those target topics whose Wikipedia articles have a number of parent categories (Cg) (upper graph) or a number of outgoing internal links (Lk) (lower graph) that fall within certain ranges.

precision and recall increase. The increase is initially rapid, then gradual, until a plateau is reached in the right half of the graph. The plateau indicates that more parent categories and outgoing internal links, beyond a sufficiently high number, do not translate into better accuracy or coverage.

Isolating Evidence from Wikipedia: The graphs in Figure 2 isolate the impact on per-entry precision and recall scores, when one of the two features from Wikipedia is temporarily disabled. In the upper graph, parent categories are enabled and outgoing internal links are disabled (i.e., not collected), whereas the opposite is true in the lower graph. From left to right, the scores are computed over subsets of target topics from WikiA whose equivalent Wikipedia articles contain between 1 and iteratively more parent categories (upper graph) or outgoing internal links (lower graph). Both precision and recall tend to increase towards the right, as more target topics, whose Wikipedia articles have more parent categories or outgoing internal links, are included in the evaluation. The absolute values of per-entry precision and recall are different between the two graphs, demonstrating that the two features do not contribute equally to extraction.

Table 4 gives per-constituent and per-entry precision and recall scores over the entire evaluation sets, when either both parent categories (Cg) and outgoing internal links (Lk) are enabled, or only one of them (Cg or Lk) is temporarily enabled. The results support multiple conclusions. First, with one exception (namely, per-entry recall over ComP), recall is higher when using Cg and Lk together than when using either one of them separately. Second, evidence from outgoing internal links, as opposed to evidence from parent categories,

Score	Feature		Evaluation Set		
	Cg	Lk	WikiA	SemE	ComP
As an average over set of golden constituents from evaluation set:					
Average per-constituent precision	-	√	0.810	0.920	1.000
	√	-	0.742	0.760	0.642
	√	√	0.757	0.777	0.823
Average per-constituent recall	-	√	0.413	0.511	0.500
	√	-	0.602	0.422	0.409
	√	√	0.765	0.622	0.583
As an average over set of target topics from evaluation set:					
Average per-entry precision	-	√	0.777	0.562	0.714
	√	-	0.806	0.546	0.535
	√	√	0.861	0.671	0.821
Fraction of target entries with max precision	-	√	0.770	0.562	0.714
	√	-	0.770	0.531	0.500
	√	√	0.819	0.625	0.785
Average per-entry recall	-	√	0.728	0.484	0.607
	√	-	0.786	0.468	0.428
	√	√	0.852	0.625	0.642
Fraction of target entries with max recall	-	√	0.680	0.406	0.500
	√	-	0.725	0.375	0.285
	√	√	0.796	0.531	0.428

Table 4: Impact of various types of evidence (features) from Wikipedia, on the precision and recall of extracted constituent topics. Computed over each evaluation set, when various types of evidence are allowed (√) or temporarily not allowed (-) to contribute towards the extraction of constituents (Cg=evidence from parent categories in the Wikipedia category network; Lk=evidence from Wikipedia outgoing internal links).

produces more constituent topics that are also more accurate. Per-constituent precision is uniformly higher for Lk than for Cg, over all evaluation sets. In addition, per-entry precision is also higher for Lk than for Cg, over SemE and ComP.

That Lk gives lower per-entry precision than Cg over WikiA in Table 4 may seem a contradiction, but in fact is consistent with results on the other evaluation sets. Given that Lk tends to produce more constituent topics than Cg, Lk is likely to give lower precision than Cg on non-compositional target topics. As shown earlier in Table 2, this should affect some entries from WikiA, but none from SemE or ComP.

Results from Table 4 suggest that, if the target topics being decomposed are an unknown mix of non-compositional and compositional topics, then enabling both outgoing internal links and parent categories is likely to give the best results. If the input topics are known to be compositional, and if accuracy is more important than coverage, then enabling only outgoing internal links is a possible strategy.

Discussion: Several types of errors, causing precision or recall losses, are relatively more frequently encountered. Noisy edges in the Wikipedia category network cause Wikipedia articles to be transitively connected upwards to many spurious ancestor categories. In turn, the connections may cause incorrect constituent topics to be extracted. For example, because the category *Civics* can be reached in upward paths starting from many Wikipedia articles, the constituent *Civics* (as in society and politics) is incorrectly extracted for the topic

Honda Civic Hybrid. Even if a relevant category is reached, it is still not usable unless a representative article for that category has also been identified. For example, the category *Football teams* is not listed as one of the parent categories of its arguably representative article *Football team*. While the category *Football teams* can be reached upwards from the article of the topic *Switzerland national football team*, the article *Football team* is not connected to it. Therefore, the method fails to extract the relevant constituent *Football team* for the topic *Switzerland national football team*.

Besides the category network, outgoing internal links in Wikipedia articles may also cause incorrect constituents to be extracted. The naming of the *Chevrolet Impala* car model was inspired by the *Impala* species of antelopes, and the Wikipedia article for the former points this out, including a link to the article for the latter. But this does not make *Impala* (the antelope) a relevant constituent topic for the car model, or at least the evaluation takes the pessimistic view that it does not, thus penalizing the extraction of the constituent *Impala* as a precision loss. When the extracted constituent is quite semantically similar to the golden constituent, yet not identical, the extracted constituent receives no credit. Examples of such near misses are extracting *Receptor (biochemistry)* instead of the golden *Sensory receptor* as a constituent of the topic *Olfactory receptor*; or extracting *Presidency* instead of *President* as a constituent of the topic *President of the United States*. Some of the near misses are caused by sub-optimal ranking of candidate constituents by the similarity model using word embeddings, where *Presidency* may be assigned a higher similarity score than *President* is, relative to the topic *President of the United States*.

5 Related Work

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 constituents words [Mitchell and Lapata, 2010; Socher *et al.*, 2013]. 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, methods such as [Lin, 1999] collect vocabularies of non-compositional phrases, to rule out any subsequent attempts to possibly translate those phrases based on their words. Our method can also be used to indirectly produce a similar vocabulary of noun phrases deemed non-compositional, for example by collecting titles of Wikipedia articles equivalent to topics for which no constituent topics are extracted. For ranking candidate constituent topics, alternatives to word embeddings include [Joshi *et al.*, 2014].

If a noun phrase is known to be compositional, other methods use text in document collections to extract lexicalized interpretations (*"from"*, *"issued by"*) of the role that a modifier (*"Swiss"*) plays relative to the head (*"passport"*) of the larger noun phrase (*"Swiss passport"*) [Hendrickx *et al.*, 2013; Nakov and Hearst, 2013]. Noun phrases are often assumed to contain one modifier and one head, for a total of only two constituent phrases, or sometimes even only two words [Kim and Nakov, 2011]. In comparison, our method is more general, as it accommodates longer topics such as *2013-14 Arsenal F.C. season*, which may contain more than two constituents. It is also complementary, in that it extracts constituents, rather than interpretations of the role of a constituent relative to the main (i.e., head) constituent. Crucially, the topics and their constituents extracted by our method are disambiguated entries (i.e., topics) in the semantic space, rather than ambiguous string entries in the lexical space (e.g., *"passport"* for *"Swiss passport"*). It is trivial to convert semantic topics into lexical entries, for example by selecting the names of the topics (*"Swiss passport"* and *"passport"*, from *Swiss passport* and *Passport*). Conversely, automatically linking lexical entries to the semantic topics corresponding to their meaning is difficult and prone to errors [Pantel and Fuxman, 2011], especially for inherently ambiguous strings occurring within Web documents of arbitrary quality.

Within the larger area of open-domain information extraction [Etzioni *et al.*, 2011; Mausam *et al.*, 2012; Hoffart *et al.*, 2013; Yao and Van Durme, 2014], methods that extract pairs of phrases in IsA relations from text are potentially useful to the task addressed by our method. Intuitively, the presence of IsA pairs like (*"Swiss passport"*, *"passport"*), whose phrases share the same head, suggests that the head, and possibly the remainder (*"Swiss"*) of the larger phrase, are lexical constituents of the larger phrase *"Swiss passport"*. However, such an approach would again be limited, at least initially, to noun phrases containing only two constituents; would produce lexical, rather than semantic constituents; and would perform poorly on noisy IsA relations extracted from text. In a sense, part of our method is a reflection of this idea, applied to pairs of categories from the Wikipedia category network, rather than to IsA pairs of phrases extracted from text. In fact, it turns an often-cited weakness of the category network, namely the presence of non-IsA relations among its edges [Ponzetto and Strube, 2007; Flati *et al.*, 2014], into a strength. Such edges, along upward paths such as *Swiss passport* $\rightarrow$ *Politics of Switzerland* $\rightarrow$ *Switzerland*, enable the extraction of constituents that would otherwise be unreachable in a "pure" IsA hierarchy.

Previous studies illustrate Wikipedia's role in knowledge acquisition [Nastase and Strube, 2008; Wu and Weld, 2010; Hoffart *et al.*, 2013] and information retrieval [Hu *et al.*, 2009; Scaiella *et al.*, 2012]. In particular, [Nastase and Strube, 2008] is related to topic decomposition. In [Nastase and Strube, 2008], the Wikipedia category *Albums* may be automatically assigned to the Wikipedia article titled *Kind of Blue*, based on the availability of the parent category *Miles Davis albums* for the article, and transitively the availability of its own parent category *Albums by artist*. Although this is not explored in [Nastase and Strube, 2008], the assignment implies that the topic, if any, to which the category *Miles Davis albums* corresponds can be partially decomposed into the topic to which the category *Albums* corresponds.

Our focus on finding constituent topics distinguishes our method from other work on expanding knowledge resources, where the expansion consists in filling in new relations for existing properties of existing topics [Wu and Weld, 2010; Dong *et al.*, 2014; West *et al.*, 2014].

6 Conclusion

Evidence in encyclopedic resources such as Wikipedia enables the open-domain decomposition of entries in knowledge repositories such as Freebase into constituents. Current work explores the extraction of interpretations of the role that various constituent topics play relative to the larger topic; additional signals for better distinguishing between fully compositional and non-compositional topics; and sources of additional vocabularies of candidate constituents, when some of the desirable constituents (e.g., *Warning signals*, *Filter*) that would best decompose a topic (*Hong Kong rainstorm warning signals*, *Air filter*) are absent even from large-scale resources such as Wikipedia or Freebase.

Acknowledgments

The authors thank Norman Casagrande, for access to similarity data for Wikipedia articles; Daniel Furrer, for assistance in assembling the WikiA evaluation set; and Janara Christensen and Jutta Degener, for comments on an earlier draft.

References

[Bollacker *et al.*, 2008] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *SIGMOD-08*, 2008.

[Davis, 1997] E. Davis. The meaning of noun phrases, 1997. Retrieved from http://commonsensereasoning.org/problem_page.html#nounphrases.

[Dong *et al.*, 2014] 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*, 2014.

[Downing, 1977] P. Downing. On the creation and use of English compound nouns. *Language*, 53, 1977.

[Etzioni *et al.*, 2011] O. Etzioni, A. Fader, J. Christensen, S. Soderland, and Mausam. Open information extraction: The second generation. In *IJCAI-11*, 2011.

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

[Flati *et al.*, 2014] T. Flati, D. Vannella, T. Pasini, and R. Navigli. Two is bigger (and better) than one: the Wikipedia Bitaxonomy project. In *ACL-14*, 2014.

[Hendrickx *et al.*, 2013] I. Hendrickx, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Szpakowicz, and T. Veale. SemEval-2013 task 4: Free paraphrases of noun compounds. In *SemEval-13*, 2013.

[Hoffart *et al.*, 2013] 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, 2013.

[Hoffart *et al.*, 2014] J. Hoffart, Y. Altun, and G. Weikum. Discovering emerging entities with ambiguous names. In *WWW-14*, 2014.

[Hu *et al.*, 2009] J. Hu, G. Wang, F. Lochovsky, J. Sun, and Z. Chen. Understanding user's query intent with Wikipedia. In *WWW-09*, 2009.

[Joshi *et al.*, 2014] M. Joshi, U. Sawant, and S. Chakrabarti. Knowledge graph and corpus driven segmentation and answer inference for telegraphic entity-seeking queries. In *EMNLP-14*, 2014.

[Kim and Nakov, 2011] N. Kim and P. Nakov. Large-scale noun compound interpretation using bootstrapping and the Web as a corpus. In *EMNLP-11*, 2011.

[Lin *et al.*, 2012] T. Lin, Mausam, and O. Etzioni. No noun phrase left behind: Detecting and typing unlinkable entities. In *EMNLP-CoNLL-12*, 2012.

[Lin, 1999] D. Lin. Automatic identification of non-compositional phrases. In *ACL-99*, 1999.

[Mausam *et al.*, 2012] Mausam, M. Schmitz, S. Soderland, R. Bart, and O. Etzioni. Open language learning for information extraction. In *EMNLP-CoNLL-12*, 2012.

[Mitchell and Lapata, 2010] J. Mitchell and M. Lapata. Composition in distributional models of semantics. *Cognitive Science*, 34(8), 2010.

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

[Nastase and Strube, 2008] V. Nastase and M. Strube. Decoding Wikipedia categories for knowledge acquisition. In *AAAI-08*, 2008.

[Pantel and Fuxman, 2011] P. Pantel and A. Fuxman. Jigs and lures: Associating web queries with structured entities. In *ACL-11*, 2011.

[Ponzetto and Strube, 2007] S. Ponzetto and M. Strube. Deriving a large scale taxonomy from Wikipedia. In *AAAI-07*, 2007.

[Porter, 1980] M. Porter. An algorithm for suffix stripping. *Program*, 14(3), 1980.

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

[Scaiella *et al.*, 2012] U. Scaiella, P. Ferragina, A. Marino, and M. Ciaramita. Topical clustering of search results. In *WSDM-12*, 2012.

[Socher *et al.*, 2013] R. Socher, J. Bauer, C. Manning, and A. Ng. Parsing with compositional vector grammars. In *ACL-13*, 2013.

[Strube and Ponzetto, 2006] M. Strube and S. Ponzetto. WikiRelate! computing semantic relatedness using Wikipedia. In *AAAI-06*, 2006.

[Van de Cruys *et al.*, 2013] T. Van de Cruys, S. Afantenos, and P. Muller. MELODI: A supervised distributional approach for free paraphrasing of noun compounds. In *SemEval-13*, 2013.

[West *et al.*, 2014] R. West, E. Gabrilovich, K. Murphy, S. Sun, R. Gupta, and D. Lin. Knowledge base completion via search-based question answering. In *WWW-14*, 2014.

[Weston *et al.*, 2013] J. Weston, A. Bordes, O. Yakhnenko, and N. Usunier. Connecting language and knowledge bases with embedding models for relation extraction. In *EMNLP-13*, 2013.

[Wu and Weld, 2010] F. Wu and D. Weld. Open information extraction using Wikipedia. In *ACL-10*, 2010.

[Yao and Van Durme, 2014] X. Yao and B. Van Durme. Information extraction over structured data: Question Answering with Freebase. In *ACL-14*, 2014.