Demonyms and Compound Relational Nouns in Nominal Open IE

Harinder Pal
Indian Institute of Technology
New Delhi, India
sethi.harinder@gmail.com

Mausam
Indian Institute of Technology
New Delhi, India
mausam@cse.iitd.ac.in

Abstract

Extracting open relational tuples that are mediated by nouns (instead of verbs) is important since titles and entity attributes are often expressed nominally. While appositives and possessives are easy to handle, a difficult and important class of nominal extractions requires interpreting compound noun phrases (e.g., "Google CEO Larry Page"). We substantially improve the quality of Open IE from compound noun phrases by focusing on phenomena like demonyms and compound relational nouns. We release RELNOUN 2.2, which obtains 3.5 times yield with over 15 point improvement in precision compared to RELNOUN 1.1, a publicly available nominal Open IE system.

1 Introduction

Open Information Extraction (Etzioni et al., 2008) systems output relational tuples from text without a pre-specified relational vocabulary by identifying relation phrases present in text. Early work on Open IE (Etzioni et al., 2011) focused on verb-mediated relations that could be expressed using a handful of patterns and still covered substantial information in text. Subsequent research has focused on increasing recall – a noteworthy approach (OLLIE) uses bootstrapping for learning general language patterns (Mausam et al., 2012). Various extensions improve on the amount of linguistic knowledge in the systems – EXEMPLAR (de Sá Mesquita et al., 2013) improves the set of rules on top of dependency parses; Open IE 4.0[1] uses carefully designed rules over semantic role labeling systems (Christensen et al., 2011); several works attempt clause identification or sentence restructuring, thus identifying sentence components and applying extraction rules on top of these components (Schmidek and Barbosa, 2014; Corro and Gemulla, 2013; Bast and Haussmann, 2013). Other approaches include use of lexico-syntactic qualia-based patterns (Xavier et al., 2015), simple sentence-specific inference (Bast and Haussmann, 2014), and a supervised approach using tree kernels (Xu et al., 2013).

While the focus on verbs continues to be common in these Open IE systems, some works have directed attention on noun-mediated relations such as OLLIE (Mausam et al., 2012), RENOUN (Yahya et al., 2014), and RELNOUN.[2] A common observation is that many relations (e.g, *capital of*, *economist at*) are more frequently expressed using nouns, instead of verbs. Common noun-mediated patterns include appositive constructions, possessive constructions, and compound noun phrases (see Table 1 for examples). While most patterns give some syntactic cues for the existence of a relation (such as a comma or a possessive 's), interpreting and extracting tuples from compound NPs is specifically challenging, since they are just a continuous sequence of nouns and adjectives (e.g., "Google CEO Larry Page").

This paper substantially improves the quality of extraction from compound noun phrases. Our work builds on the publicly available RELNOUN system (ver 1.1) and extends it to RELNOUN 2.2, which incorporates three additional sources of recall from compound noun phrases: (1) capitalized relational

[1] *https://github.com/knowitall/openie*

[2] *https://github.com/knowitall/chunkedextractor*

Extractor	Phrase	Extraction
Verb1	Francis Collins is the director of NIH	(Francis Collins; is the director of; NIH)
Verb2	the director of NIH is Francis Collins	(Francis Collins; is the director of; NIH)
Appositive1	Francis Collins, the director of NIH	(Francis Collins; [is] the director of; NIH)
Appositive2	the director of NIH, Francis Collins,	(Francis Collins; [is] the director of; NIH)
Appositive3	Francis Collins, the NIH director	(Francis Collins; [is] the director [of]; NIH)
AppositiveTitle	Francis Collins, the director,	(Francis Collins; [is]; the director)
CompoundNoun	*NIH director Francis Collins*	*(Francis Collins; [is] director [of]; NIH)*
Possessive	NIH's director Francis Collins	(Francis Collins; [is] director [of]; NIH)
PossessiveAppositive	NIH's director, Francis Collins	(Francis Collins; [is] director [of]; NIH)
AppositivePossessive	Francis Collins, NIH's director	(Francis Collins; [is] director [of]; NIH)
PossessiveVerb	NIH's director is Francis Collins	(Francis Collins; is director [of]; NIH)
VerbPossessive	Francis Collins is NIH's director	(Francis Collins; is director [of]; NIH)

Table 1: RELNOUN 1.1 extractors along with the example phrases and corresponding extractions

nouns, (2) demonyms, the adjectives used to identify residents of a location (e.g., 'Japanese' for 'Japan'), and (3) compound relational nouns (see Table 2 for examples). Compared to its predecessor, RELNOUN 2.2 triples the yield with over 15 point improvement in precision. Our code is freely downloadable.[2]

2 Background on Nominal Open IE

Probably the earliest work on Nominal Open IE is OLLIE, which is a pattern learning approach based on a bootstrapped training data using high precision verb-based extractions (Mausam et al., 2012). It identified that nominal IE can't be completely syntactic, and, at the least, a list of relational nouns (e.g, *mother, director, CEO, capital*) is needed for high precision extraction. OLLIE is superseded by RELNOUN, which is a rule-based extractor incorporating most of the high precision learnings of OLLIE.

A third work on nominal Open IE is RENOUN (Yahya et al., 2014). RENOUN builds a comprehensive list of relational nouns using bootstrapping over query logs and text. It then uses seed patterns to extract data and then uses these as source of distant supervision for additional pattern learning. Unfortunately, neither their list of relational nouns, nor their final extractor are available. Moreover, it is hard to reproduce their list of nouns since most researchers don't have access to query logs. Hence, we build upon the publicly available RELNOUN system.

RELNOUN 1.x series is a set of POS and NP-Chunk patterns defined to extract a high precision subset of noun-mediated extractions (see Table 1). The input to RELNOUN is a set of relational nouns, which are extracted using bootstrapping – these include words which are common headnouns for X in "is a X of" patterns, as well as words which are within the 'Person' subclass in WordNet hierarchy (Miller, 1995).[3] We added a couple of missing patterns and made small modifications to the previous RELNOUN release (version 1.0.9) to increase its coverage and precision. The resulting system RELNOUN 1.1, acts as the baseline for our work.

Our analysis of RELNOUN 1.1 revealed significant missed recall when extracting from compound noun phrases such as "*Mechelen Mayor Bart Somers*", "*Chinese president Hu Jintao*", or "*United States health minister Levitt*". The desired extractions are (Bart Somers, [is] Mayor [of], Mechelen), (Hu Jintao, [is] president [of], China), and (Levitt, [is] health minister [of], United States). We attribute this to three important missing phenomena in RELNOUN 1.1 when extracting from compound noun phrases – capitalized relational nouns ('Mayor'), demonyms ('Chinese'), and compound relational nouns ('health minister'). Note that here a compound *relational noun* occurs within a larger compound noun phrase. RELNOUN 2.2 improves the analysis for all these three categories.

3 RELNOUN 2.2

RELNOUN 1.1 does not extract from phrases containing capitalized (NNP) relational nouns (e.g., "Mechelen Mayor Bart Somers") even though, at times, that is grammatically correct whereas uncapitalized nouns are not.[4] The main reason for this

[3] RELNOUN was not published as a research paper. Some of the system details are based upon personal communication with the main engineer, Michael Schmitz.

[4] http://blog.esllibrary.com/2012/11/01/when-do-we-

Phrase	RELNOUN 1.1	RELNOUN 2.2
"*United States President Obama*"	(Feliks, [is] historian [of], Seattle)	(Obama, [is] President [of], United States)
"*Seattle historian Feliks*"		(Feliks, [is] historian [from], Seattle)
"*Japanese foreign minister Kishida*"		(Kishida, [is] foreign minister [of], Japan)
"*GM Deputy Chairman Lutz*"		(Lutz, [is] Deputy Chairman [of], GM)

Table 2: Comparison of RELNOUN 1.1 and RELNOUN 2.2 on some phrases

choice is that allowing extractions from compound noun phrases with capitalized relational nouns can lead to a large number of false positives. On further analysis we observe three major categories of errors:

1. Organization names: Erroneous tuples are extracted when the compound NP is the name of an organization. For example, it extracts (Association, [is] Banker [of], New York) from the phrase *"New York Banker Association"*.
2. Demonyms: A common error is when the title is preceded by a demonym. For e.g., it extracts (Angela Merkel; [is] Chancellor [of]; German) from *"German Chancellor Angela Merkel"*.
3. Compound Relational Nouns: The relational noun with a pre-modifier often confuses the extractor. For example, *"Prime Minister Modi"* yields (Modi, [is] Minister [of], Prime).

The first set of errors is easy to fix. We create a list of 160 organization words and filter out any extractions where arg1 has an organization word. The list is created by extracting the most frequent last words from the list of organizations on Wikipedia. They include words like 'Committee', 'Limited', 'Group', and 'Association'. This ORG filtering improves the precision slightly, with almost no impact to recall.

The next two subsections detail our approaches for incorporating knowledge about demonyms and compound relational nouns in RELNOUN.

3.1 Demonyms

Demonyms are words derived from the name of a location and are used to identify residents or natives of that location. Typical examples include 'Israeli', 'Japanese', and 'South African' for residents from Israel, Japan, and South Africa, respectively. We first parse a list of demonyms from Wikipedia to populate a table of (Location, Demonym) pairs.[5] We expand this table with information from additional geographical websites[6] leading to a total of 2,143 base entries. The demonyms (and locations) can frequently take region-related pre-modifiers. When checking for demonyms, we allow them to be preceded by 'North', 'South', 'East', 'West', 'Northern', 'Southern', 'Eastern', 'Western', and 'Central'.

To extract relationships expressed via demonyms appropriately (e.g. *"German Chancellor Angela Merkel"*), we simply check whether arg2's headnoun is in our demonym table, and if it is then we replace it with its corresponding location from the table. Demonym replacements are also needed for compound NPs without capitalization, for example, *"German chancellor Angela Merkel"*. This requires another small extension to RELNOUN – allowing arg2 to be a JJ when it is in the demonym list (typically arg2s can only be an NNP).

In addition to compound NPs, the demonym replacement can be useful for other patterns from Table 1 also. For example, AppositivePossessive and PossessiveAppositive both benefit from this (for example, *"Angela Merkel, German Chancellor"* and *"German Chancellor, Angela Merkel*).

Domicile vs. Title Classification: Demonyms are used to denote two common relationships. First, arg1 may be directly related to the location or the government of the location through the relational noun. For example, *"United States President Obama"* – Obama is the President of the country of (or govt. of) United States. A second usage simply suggests that the location is arg1's *domicile*, i.e., arg1 is a native of, lives in, or has another substantial connection with the location. For example, *"Canadian pitcher Andrew Albers"* only denotes a domicile – Albers is not a player *of* Canada! Ideally, we would like to extract (Andrew Albers, [is] player [*from*], Canada), instead of [of].

We manually create a small list of ten relational nouns, which represent heads and other high-posts

capitalize-president/ says *"Use a capital when the title directly precedes the name"*

[5]https://en.wikipedia.org/wiki/Demonym

[6]http://www.geography-site.co.uk, http://everything2.com, http://geography.about.com

System	Precision	Yield
OLLIE-NOUN	0.29	136
RELNOUN 1.1	0.53	60
+ NNP relational nouns	0.37	100
+ ORG filtering	0.39	100
+ demonyms	0.52	158
+ compound relational nouns	0.69	209

Table 3: Precision and Yield (#correct extractions) for each system on a dataset of 2000 random Newswire sentences.

of the govt. of a city, state and country (includes 'king', 'president', 'mayor', 'governor') and use only those for titles. All other relational nouns are assumed to be in the domicile sense. While not perfect, the resulting extractions are often accurate, for example, (Sepoy Kanshi, [is] soldier [from], India) is accurate even if he is also a soldier of India.

3.2 Compound Relational Nouns

We now extend RELNOUN to handle *compound relational nouns* such as 'health minister', 'foreign secretary', and 'vice president'. We first observe that for all extractors in Table 1, except CompoundNoun there are lexical indicators ('of', ',', or possessive marker) to segment a relational noun. However, because CompoundNoun pattern is simply a sequence of nouns, segmenting relational nouns is harder and it is a common source of errors.

Segmenting compound relational nouns is relatively easy if it is followed by a demonym, since the demonym can help segment the left boundary (e.g., "*Indian Prime Minister Modi*"). However, the real challenge is in non-demonym cases – disambiguating whether "*GM Vice Chairman Bob Lutz*" should result in (Bob Lutz, [is] Vice Chairman [of], GM) or (Bob Lutz, [is] Chairman [of], GM Vice).

We bootstrap a list of common relational noun prefixes over a large text corpus. We collect all words (from 1.1 million sentences of ClueWeb12 corpus) that precede relational nouns and are in the same NP chunk, are of appropriate POS (JJ, NN, NNP, etc), are not a known demonym and don't end in a possessive. We keep all such prefixes of frequency greater than 20, resulting in a list of 5,606 prefixes. We use these prefixes to segment the final relational nouns in compound NPs. Even though noisy, this list serves our purpose since common prefixes are already present, and it is used only for non-demonym CompoundNoun extractor.

4 Experiments

Our goal is to compare RELNOUN 1.1 with RELNOUN 2.2. We randomly sample 2,000 sentences from Newswire and run both (and other intermediate systems) on them. We ask two annotators (students) to tag if the sentence asserted or implied an extraction. Our inter-annotator agreement is 0.97 and we retain the subset of extractions on which the annotators agree for further analysis. Note that while precision and yield (number of correct extractions) can be naturally computed by tagging extractions, estimating recall is challenging, as it requires annotators to tag all possible extractions from these sentences. Following previous work (Mausam et al., 2012), we report yield, since recall is proportional to yield and suffices for system comparisons.

Table 3 reports the results. We find that OLLIE's noun patterns have a good yield but poor precision, whereas RELNOUN 1.1 has a decent precision of 0.53, but the yield is much lower. Allowing NNP relational nouns in RELNOUN 1.1 has a 66% increase in yield, but precision takes a severe hit. ORG filtering only helps a little bit in improving precision. Handling of demonyms has a huge impact since not only does yield increase by another 58%, precision goes up 13 points too. Finally, incorporating compound relational nouns adds another 51 extractions with a significant (17 point) improvement in precision. Overall, RELNOUN 2.2 has a 3.5x increase in yield with a 16 point increase in precision, making it a substantial improvement on the existing RELNOUN 1.1 system.

5 Conclusions

An important subtask of Open IE is Nominal Open IE, within which dealing with compound NPs is particularly challenging owing to complications arising due to organization names, demonyms and compound relational nouns. We release RELNOUN 2.2, a significant improvement over the publicly available RELNOUN 1.1 system, which deals with each of these challenges. The main approach for improvement uses a combination of rule-based systems and semantic lists, bootstrapped automatically from text corpus and also compiled manually over the Web. RELNOUN 2.2 has 3.5x more yield with a substantial increase of 16 precision points.

Acknowledgments

The work was supported by Google language understanding and knowledge discovery focused research grants and a Bloomberg grant. We thank Michael Schmitz, the author of the original RELNOUN system, for helping with details of the algorithm.

References

Hannah Bast and Elmar Haussmann. 2013. Open information extraction via contextual sentence decomposition. In *2013 IEEE Seventh International Conference on Semantic Computing, Irvine, CA, USA, September 16-18, 2013*, pages 154–159.

Hannah Bast and Elmar Haussmann. 2014. More informative open information extraction via simple inference. In *Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, pages 585–590.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2011. An analysis of open information extraction based on semantic role labeling. In *Proceedings of the 6th International Conference on Knowledge Capture (K-CAP 2011), June 26-29, 2011, Banff, Alberta, Canada*, pages 113–120.

Luciano Del Corro and Rainer Gemulla. 2013. Clausie: clause-based open information extraction. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 355–366.

Filipe de Sá Mesquita, Jordan Schmidek, and Denilson Barbosa. 2013. Effectiveness and efficiency of open relation extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 447–457.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Commun. ACM*, 51(12):68–74.

Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam. 2011. Open information extraction: The second generation. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 3–10.

Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 523–534.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Jordan Schmidek and Denilson Barbosa. 2014. Improving open relation extraction via sentence restructuring. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 3720–3723.

Clarissa Castellã Xavier, Vera Lúcia Strube de Lima, and Marlo Souza. 2015. Open information extraction based on lexical semantics. *J. Braz. Comp. Soc.*, 21(1):4:1–4:14.

Ying Xu, Mi-Young Kim, Kevin Quinn, Randy Goebel, and Denilson Barbosa. 2013. Open information extraction with tree kernels. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 868–877.

Mohamed Yahya, Steven Whang, Rahul Gupta, and Alon Y. Halevy. 2014. Renoun: Fact extraction for nominal attributes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 325–335.