Boosting Open Information Extraction with Noun-Based Relations

Clarissa Castellã Xavier, Vera Lúcia Strube de Lima
PUCRS, Graduate Program in Computer Science
Porto Alegre, RS, Brazil
E-mail: clarissacastella@gmail.com, vera.strube@pucrs.br

Abstract

Open Information Extraction (Open IE) is a strategy for learning relations from texts, regardless the domain and without predefining these relations. Work in this area has focused mainly on verbal relations. In order to extend Open IE to extract relationships that are not expressed by verbs, we present a novel Open IE approach that extracts relations expressed in noun compounds (NCs), such as (*oil, extracted from, olive*) from "olive oil", or in adjective-noun pairs (ANs), such as (*moon, that is, gorgeous*) from "gorgeous moon". The approach consists of three steps: detection of NCs and ANs, interpretation of these compounds in view of corpus enrichment and extraction of relations from the enriched corpus. To confirm the feasibility of this method we created a prototype and evaluated the impact of the application of our proposal in two state-of-the-art Open IE extractors. Based on these tests we conclude that the proposed approach is an important step to fulfil the gap concerning the extraction of relations within the noun compounds and adjective-noun pairs in Open IE.

Keywords: Information Extraction, Open Information Extraction, Relation Extraction

1. Introduction

According to Nastase et al. (2013) "Every nontrivial text describes interactions and relations". These relations are the connections we perceive among concepts, entities, and events and, also by means of attributes. Those connections can be understood as links between lexemes that, together, reflect a single sense.

Open Information Extraction (Open IE) is an approach to extract relations from texts. It emerged to deal with the heterogeneous nature of the Web and its enormous amount of data, where relations of interest are unpredictable and varied (Banko et al., 2007). Open IE should acquire relations without defining them in advance, regardless the domain.

Wu and Weld (2010) define an Open IE extractor as "a function from a document, d, to a set of triples (arg1, rel, arg2), where arg1 and arg2 are noun phrases and rel is a textual fragment indicating an implicit semantic relation between the two noun phrases".

Open IE state-of-the-art systems (Banko et al., 2007; Wu and Weldt, 2010; Mausam et al., 2012) were designed to extract textual relations that are expressed by verbs. However, as pointed by Etzioni et al. (2011), verbs are not the only way to express relations between nouns in a text. A first attempt to address non-verbal relations gave rise to ClausIE (Corro and Gemulla, 2013), a system that performs non-verbal-mediated extractions for appositions and possessives.

Complementarily, we are interested in alternatives to learn relations as the ones within noun compounds (NCs) and adjective noun pairs (ANs). For example, the AN "raw food" can be interpreted by the relation (food, that is, raw) and the triple (vase, made of, glass) can describe the relation within the NC "glass vase". With this study, we aim to improve Open IE results, providing noun-based extractions.

To accomplish noun-based relation extraction, we use the interpretation approach. As demonstrated by Butnariu et al. (2010), NCs provide a concise means of evoking a relationship between two or more nouns. Transposing this idea to ANs we observe that they also evoke a relation, in this case an attribution relation between the adjective and the qualified noun. Thus, we extend their proposal of interpretation of the relations within NCs, and apply this proposal to ANs as well.

Our method consists of three main steps:

1. Extraction of NCs and ANs:
   identifies noun compounds and adjective-noun pairs present in the input text.
2. NCs and ANs Interpretation:
   interprets the NCs and ANs previously identified. This interpretation consists of associating the compounds to explanatory sentences that are used to enrich the input corpus. For example, in case the NC "neck vein" is detected, the input corpus should be enriched with a corresponding interpretation, such as "vein that comes from the neck".
3. Relation Extraction:
   produces a set of triples describing binary relations extracted from the corpus that was enriched in the previous step.

To confirm the feasibility of this method we created a prototype and evaluated the impact of the application of our proposal in two state-of-the-art Open IE extractors: Ollie (Mausam et al., 2012) and ClausIE (Corro and Gemulla, 2013). From those results we demonstrate that the method increases Ollie's Yield from 1033.2 to 1107.92 and ClausIE Yield from 1685.7 to 1864.98.

This article is organized in five sections. Section 2 presents Open IE and related work. Section 3 is the core of the paper and details the architecture of the solution proposed, including NCs and ANs extraction and interpretation, and relations extraction. In Section 4 we include information on implementation, experiments and evaluation. Finally, Section 5 brings our conclusions and comments on future work.

2. OpenIE and Related Work

Open IE systems use two main routes to implement relation extraction. The first one is machine learning to automatically learn the patterns from a training corpus. The second one is based on heuristic rules and aims to identify the occurrence of specific patterns in the text.

The TextRunner system, presented in (Banko et al., 2007) introduced this Open IE paradigm. It was followed by WOE systems that are a continuation of TextRunner, including changes on training data. WOEpos and WOEparse systems (Wu and Weld, 2010) are the two versions of WOE, respectively using part-of-speech (POS) tagging and dependency parsing information.

The ReVerb system's design (Etzioni et al., 2011) is based on simple rules that identify verbs expressing relationships. It receives POS tagged and chunked sentences as input. First it identifies the relations and then it extracts the relation arguments. The authors report that ReVerb achieved an AUC (area under precision-recall curve) more than twice as big as those from TextRunner and WOEpos and, 38% higher than the one from WOEparse.

Ollie (Mausam et al., 2012) aims to improve Open IE expanding the syntactic scope to get context information as attribution and clausal modifiers, obtaining a larger number of relations. The system starts with a set of seed tuples from ReVerb to then bootstrap a training set used to learn pattern templates. Next, it applies these patterns into extraction and then, it analyzes the context-window around the tuples in order to add information using a confidence function. The authors report that the system obtains a 1.9 to 2.7 times bigger area under precision-yield curve, if this one is compared to those from ReVerb and WOE systems.

ClausIE (Corro and Gemulla, 2013) detects clauses and clause types. A clause is understood as the part of a sentence that expresses some coherent piece of information, consisting of subject, verb and complement and one or more adverbs. For each input sentence, ClauseIE parses this sentence with a dependency parser; determines the set of clauses; for each clause, it determines the set of coherent derived clauses based on the dependency parsing and lexica; and generates propositions (triples). The results for each of three different datasets are presented as "number of correct extractions/total number of extractions": 500 sentences from ReVerb test dataset - 1706/2975; 200 random sentences from Wikipedia - 598/1001; and 200 random sentences from the New York Times - 696/1303.

Although extending Open IE to learn relations from NCs and ANs is a recent concern, the interest in automatically learning the relations within compounds is not new. James Allen (1995) indicates the importance and the challenge of detecting the semantic relationship between the modifying and the modified noun stating that noun-noun modifiers are notoriously difficult to analyze.

In the next section we propose a simple strategy to improve Open IE extractors by learning these relations.

3. Architecture of the Proposed Solution

Aiming to provide a strategy to boost Open IE systems by extracting relations not expressed by verbs we present a solution graphically represented in Figure 1. It comprises three main components: (1) extraction of NCs and ANs; (2) generation of paraphrases and corpus enrichment; (3) extraction of relations from the enriched corpus (this extraction might be performed by the available Open IE systems). The output is a set of triples describing binary relationships extracted from these compounds.

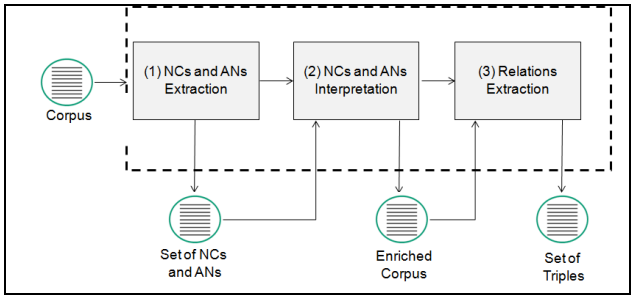

Figure 1: Architecture of the proposed solution

The components in Figure 1 are explained in detail in sections 3.1 to 3.3.

3.1 Extraction of NCs and ANs

The goal here is to extract a list of NCs and ANs from the input textual corpus. We start running a POS tagger and a Noun-Phrase (NP) chunker over the input text. After the identification of the NCs and ANs with the use of a set of patterns, the two lists created are unified to produce a set of ANs and NCs. This set of ANs and NCs is the output of the first extraction step.

Figure 2 illustrates the component architecture.

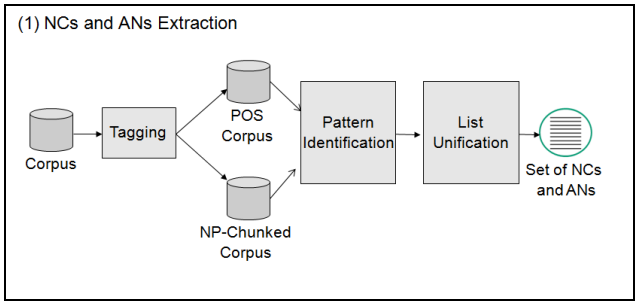

Figure 2: NCs and ANs Extraction Component

```
NNP = ([\\w]*[/ ]*(NNP[/ ]))
NN = ([\\w]*[/ ]*(NN[/ ]|NNS[/ ]))
JJ = ([\\w]*[/ ]*(JJ[/ ]))
JJNN= (JJ NN)
NNNN= (NN NN)
NNPNN= (NNP NN)
Extraction Pattern 1 = NNPNN | NNNN | JJNN
```

Figure 3: POS based patterns to identify NCs and ANs

After POS tagging and NP chunking the input corpus, the sets of patterns shown in Figure 3 and 4 are used to identify NCs and ANs over the labeled text, producing

two lists, one for each set of patterns. Then the two lists are unified producing a set of ANs and NCs, without repetitions, which is the output of the component.

```
sub1 = ((NNS|NN|NNP|JJ) [/ ] [\\w]*)
sub2 = ((NNS|NN) [/ ] [\\w]*)
Extraction Pattern = sub1 [/ ]* sub2
Extraction Pattern 2 = NNPNN | NNNN | JJNN
```

Figure 4: Patterns to identify NCs and ANs in chunked sentences

3.2 Interpretation of NCs and ANs

This component automatically interprets the NCs and ANs identified in the previous stage associating them to explanatory sentences. To accomplish this task we propose a method comprising three steps: (1) automatic interpretation of the compounds; (2) selection of the suitable interpretation in cases where more than one interpretation was generated in the first stage; (3) enrichment of the input corpus with the sentences provided by the second stage.

In the first step NCs and ANs are interpreted in a fully automatic way. In this study we opted for approaches that perform the interpretation of compounds without using pre-defined inventories of relations, since this strategy could limit us to a static list of relationships, not desirable in the Open IE context. A combined strategy, using inventories together with other approaches, might be of interest as well. Next, we describe the interpretation strategies implemented in our method.

3.2.1. Interpretation of ANs

The solution is grounded on the idea that AN pairs are composed by a noun and an adjective describing a certain quality of this noun. From this notion, the first step is designed to identify in the input text the AN pairs. To perform this action we POS tag the text and select the sequences formed by adjective + noun.

Then, for each AN, we produce a sentence in the form "NOUN that is ADJECTIVE". For example, the AN "clear information" is tagged as clear JJ information NN1, JJ being the adjective tag and NN the noun tag. So the sentence information that is clear is associated with the AN "clear information" being available to enrich the original corpus.

3.2.2. Interpretation of NCs

In this work we propose to interpret the NCs with aid of external resources, namely: WordNet2, DBPedia3 and Wiktionary4. Next, we describe the techniques envisaged for each of these resources.

Interpretation with the use of WordNet: if there is a synset in WordNet corresponding to the NC, the definition linked to this synset is then associated with the NC as its interpretation. For instance, the "oil industry" synset definition is "industry that produces and delivers oil". So, this sentence is taken as interpretation of the NC "oil industry".

Interpretation with the use of Wiktionary: if there is an entry corresponding to a certain NC in the dictionary, in that case its definition is taken as the NC interpretation. For instance, Wiktionary defines the NC "blood pressure" as "the pressure exerted by the blood against the walls of the artery|arteries and veins". So, that definition is assumed as a NC interpretation.

Interpretation with the use of DBPedia: if there is a resource in DBPedia corresponding to the NC, its definition is assumed as the NC interpretation. For instance, the resource corresponding to the NC blood pressure in DBPedia is "blood pressure (bp) is the pressure exerted by circulating blood upon the walls of blood vessels, and is one of the principal vital signs. when used without further specification, blood pressure usually refers to the arterial pressure of the systemic circulation". So, that definition is associated to the NC as interpretation.

3.3 Extraction of Relations

Given a corpus enriched with NCs and ANs interpretations, the last component performs the extraction of relations from this corpus using an Open IE extractor that ought be able to learn verbal relations from the corpus. This component should produce a set of triples representing binary relationships extracted from the corpus. Any available Open IE extractor can be used at this stage, provided that it runs over a textual corpus enriched with the NCs/ANs interpretations.

4. Concerns on Implementation, Experiments and Evaluation

A prototype was developed to assess the proposed approach. We used as input the same test corpus used by ReVerb 5 (Etzione et al., 2011), composed of 500 sentences from the web. The OpenNLP6 tagger was used to label the corpus. The prototype was implemented in Java and the executable file and source code are available at https://sites.google.com/site/clarissacastella/nlp-tools. Two linguists acted as judges in order to assess the results. We were interested in evaluating the impact of our solution when it is applied over two Open IE extractors: Ollie and ClauseIE. Examining the relations extracted from the original (not enriched) corpus, ClausIE and Ollie learned 2298 and 1260 triples each, respectively. Applying our approach and running those systems over the enriched corpus, ClausIE extracted 618 extra relations and Ollie found 202 extra relations.

Mausam et al. (2012) propose the use of Yield as a metric to evaluate Open IE systems. Yield is calculated multiplying the total number of extractions by Precision. Considering the original corpus, Yield for each system is:

[1] Penn Treebank II Tags

[2] http://wordnet.princeton.edu/

[3] http://dbpedia.org/

[4] http://en.wiktionary.org/

[5] Downloaded from http://reverb.cs.washington.edu

[6] http://opennlp.apache.org/

1033.2 for Ollie and 1685.7 for ClausIE. When using the corpus enriched according to the present proposal, Yield becomes 1107.92 for Ollie and 1864.98 for ClausIE (see Figure 5).

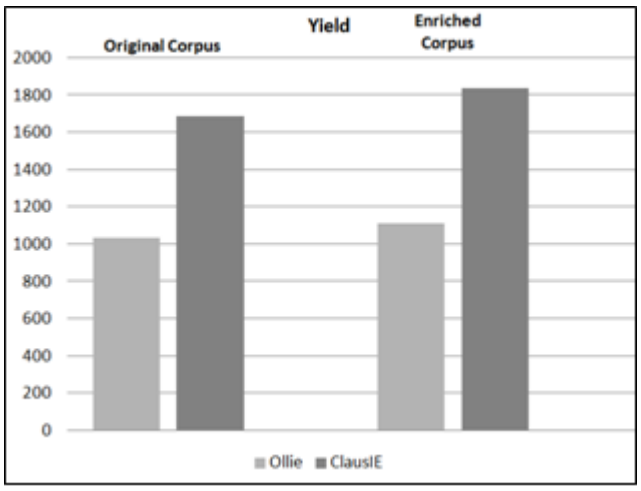

Figure 5: Increase of extractions by system

Regarding relations extracted from ANs, Ollie was not able to get any triple from those. This fact can be due to the lack of a specific pattern to capture the relation from the textual expression adopted in our proposal. ClausIE learned 53 relations from ANs, such as the triple (destinations, are, Canadian) from the sentence "destinations that are Canadian". Unfortunately ClausIE missed several extractions of triples from the ANs. For instance, for the AN "comic book" interpreted as *"*book that is comic*"* ClausIE generated the triple (that , is , comic) using the determiner "that" in place of the noun *"comic*" in the first attribute. Certainly, textual adaptations over the generated sentences that enrich the corpus, may lead to better results in this case. A further study may conduct to a clearer expression of the attribution relation as well.

A very important issue that must be addressed concerns the meaning of the compounds. For example, the NC "orthopedic surgery" was interpreted as "orthopedic surgery or orthopedics (also spelled orthopaedic surgery and orthopaedics in british english) is the branch of surgery concerned with conditions involving the musculoskeletal system. orthopedic surgeons use both surgical and nonsurgical means to treat musculoskeletal trauma, sports injuries, degenerative diseases, infections, tumors, and congenital disorders". This kind of problem that comes with the textual explanation and its length, was mainly observed within interpretations taken from DBPedia. This suggests us reconsider the way we use DBPedia or other similar resources in future implementations of the method.

It is also important to have in mind that ANs are much more frequent than NCs, so that their presence contributes with the largest amount of the extra triples. If the Open IE system is tuned to capture these relations, they make good difference regarding the results. The relations extracted from NCs show to bring a not so numerous but quite significant contribution to the quality of the results.

5. Conclusion and Future Work

This paper introduces a solution to boost the performance of Open IE systems with the extraction of non-verbal relations contained in NCs and ANs. The approach consists of three steps: detection of NC and ANs, interpretation of these compounds in view of corpus enrichment and extraction of relations from the enriched corpus.

We built a prototype and evaluated the impact of the application of our solution into different Ollie and ClausIE Open IE extractors. For that we used the enriched corpus as input for the extractors. We conclude that the proposed approach is an important step to fulfill the gap concerning the extraction of relations within the noun compounds and adjective-noun pairs in Open IE. However, this solution is strongly dependent on the available resources as well as the techniques used in the interpretation of compounds, which still have a large room for improvement.

As future work, we plan to enhance the interpretation component with other approaches. We believe that the introduction of a co-reference resolution task is mandatory to improve the quality of the results of any Open IE method. Pronominal arguments present in the triples that represent relations claim for a broader analysis that would allow a more significant extraction. Efforts toward cleaner results (with fewer cases of duplication) are also important. Without these additional steps, the relations extracted tend to be less informative or even useless for certain applications.

6. Acknowledgement

Clarissa Castellã Xavier thanks CAPES for the PhD scholarship.

7. References

Allen J. (1995). Natural Language Understanding (2nd Ed.). Benjamin-Cummings Publ. Co., Inc., Redwood City, CA, USA.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead M., Etzioni O. (2007). Open information extraction from the web. In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI '07), Rajeev Sangal, Harish Mehta, and R. K. Bagga (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2670-2676.

Butnariu, C., Kim, S. N., Nakov, P., Séaghdha, D. O., Szpakowicz, S., Veale, T. (2010). SemEval-2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 39-44.

Corro L. D., Gemulla R. (2013). ClausIE: clause-based open information extraction. In Proceedings of the 22nd International Conference on World Wide Web (WWW

'13). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 355-366.

Etzioni O., Fader A., Christensen, J., Soderland, S., Mausam. (2011). Open information extraction: the second generation. In Proceedings of the 22nd International Joint Conference on Artificial Intelligence - Volume One (IJCAI '11). AAAI Press 3-10.

Mausam, Schmitz M., Bart, R., Soderland, S., Etzioni, O. (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 '12). Association for Computational Linguistics, Stroudsburg, PA, USA, 523-534.

Nastase, V., Nakov, P., Séaghdha, D. Ó., Szpakowicz, S. (2013). "Semantic Relations between Nominals". Series: Synthesis Lectures on Human Language Technologies. Morgan&Claypool Publishers.

Wu F., Weld D. S. (2010). Open information extraction using Wikipedia. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 118-127.