Improving Open Information Extraction using Domain Knowledge

Cheikh Kacfah Emani, Catarina Ferreira da Silva, Bruno Fiès, Parisa Ghodous

SDSW, Workshop on surfacing the deep and the social web, co-located with ISWC 2014, 13th International semantic web conference, October 19-23, 2014, Trentino, Italy

Improving Open Information Extraction using Domain Knowledge

Cheikh Kacfah Emani[1,2], Catarina Ferreira Da Silva[2], Bruno Fiès[1], and Parisa Ghodous[2]

[1] CSTB, 290 route des Lucioles, BP 209, 06904 Sophia Antipolis, France
[2] Université Lyon 1, LIRIS, CNRS, UMR5205, F-69622, France

Abstract. Open Information Extraction (OIE) aims to identify all the possible assertions within a sentence. Recent and thus the most efficient OIE-tools use the grammatical dependencies or the syntactic tree of the sentence to perform extraction. When they provide a wrong extraction it is mainly due to parsing errors. In this paper, we propose to handle these parsing errors before doing OIE itself. To achieve our goal we focus on multi-word expressions (MWE). They represent more than 45% of wrong extractions. We show how the MWE-problem can be handle in a given domain and how MWE-unbreakable property is a good filter for OIE.

1 Introduction

In recent years, researchers have tackled the problem of Open Information Extraction in different manner: from machine learning [8] to the exploitation of sentence structure [7], [2]. This last type of approaches obtains the best results. Unfortunately, their OIE-tools (exploiting grammatical dependencies [7] and syntactic tree [2]) sometimes output incorrect tuples. These wrong extractions are mainly due to parsing errors. Indeed these approaches take advantage of the syntactic tree or grammatical dependencies provided by a parser. Consequently, a good way to improve Open Information Extraction is to handle parsing errors before the extraction stage itself. To achieve this goal, we have decided to *handle multi-word expressions* (MWE). A MWE is a phrase, made up of a set of words, which has a precise meaning and is unbreakable. "MWE-errors" represent more than 45% of parsing errors. We propose an algorithm to shorten multi-word expressions. We have evaluated our proposals in a given domain, which is law texts in building engineering construction, and show how we outperform existing tools. Indeed, in a given domain, multi-word expressions are easy to handle: domain terms, recurrent domain-independent terms, named entities, etc. Our goal is discussed through the following agenda. Initially, a brief state of the art on OIE is presented (Sect. 2). Next, we detail our contribution (Sect. 3). Finally, our algorithms are tested on a set of sentences issue from law in the field of engineering construction (Sect. 4.1) and we discuss our early results (Sect. 4.2).

2 Related work

As mentioned in the introduction, we want to perform "Open" Information Extraction, but from sentences which describe a *precise field*. This constraint makes us have pieces of information about *terminology* as described in next sections. Nevertheless, our ideas can be used in an "open" manner (MWE will be mainly named entities, formulae, etc.) and thus be compared to "traditional" OIE-systems.
During the recent years, many systems were developed to perform OIE. It is the case of `ReVerb` [8], `OLLIE` [9], `ClausIE` [7] and CSD-IE [2,3]. `ReVerb` by means of efficient heuristics, focused on *incoherent* and *uninformative* triples. Unfortunately, relations extracted by `ReVerb` were necessarily verb-based. This is the main reason why `OLLIE`, developed by the same group of researchers, was provided. In addition to be able to identify non verb-driven facts, `OLLIE` aims to provide the *context/condition*, if existing, in which the extracted fact can *be considered true*. These two previous tools are machine learning-based. The most recent approach does not need any additional resource. They only exploit result of a standard parser. `ClausIE` uses grammatical typed dependencies and `CSD-IE` the syntactic tree of the input sentence. These two tools dissect each piece of the result they get from the parsing tool. Consequently, if a dependency is wrong or a sub-tree is incorrectly labelled in the syntactic tree these OIE-tools may provide inaccurate extraction. This is why we propose to make some preprocessing operations before OIE itself. The details of these tasks are given in the next section.

3 Handle Multi-Word Expressions

Researchers commonly agree that parsing errors lead to major incorrect extractions in Open Information Extraction. In a sample set of sentences select from various regulatory texts in the field of building engineering (see Sect. 4.1 for more details about the corpus), the percentage of errors due to MWE is 46.15% using `CSD-IE`. To handle problems caused by MWE is a thus relevant way to improve result of IE. Our solution to improve the quality of OIE-tools when they face the MWE-problem is a three-step operation: (*i*) Detect MWE, (*ii*) compress each MWE and (*iii*) expand each MWE at the end of the extraction step. This process is illustrated by Fig. 1.

Step 1 - Detection of Multi-Word Expressions

For us, a MWE is every phrase which the meaning will be modified (even become meaningless) by the addition or the deletion of any of its word. Consequently, a *domain term*, an *idiomatic expression*, a *phrasal verb*, a *named entity*, a *formula*, a *quotation* etc. is a MWE. These examples of MWE make us foresee that MWE are more easily and *reliably* identifiable in a given domain. One can have also domain-independent terms that are not related to the field of study but are

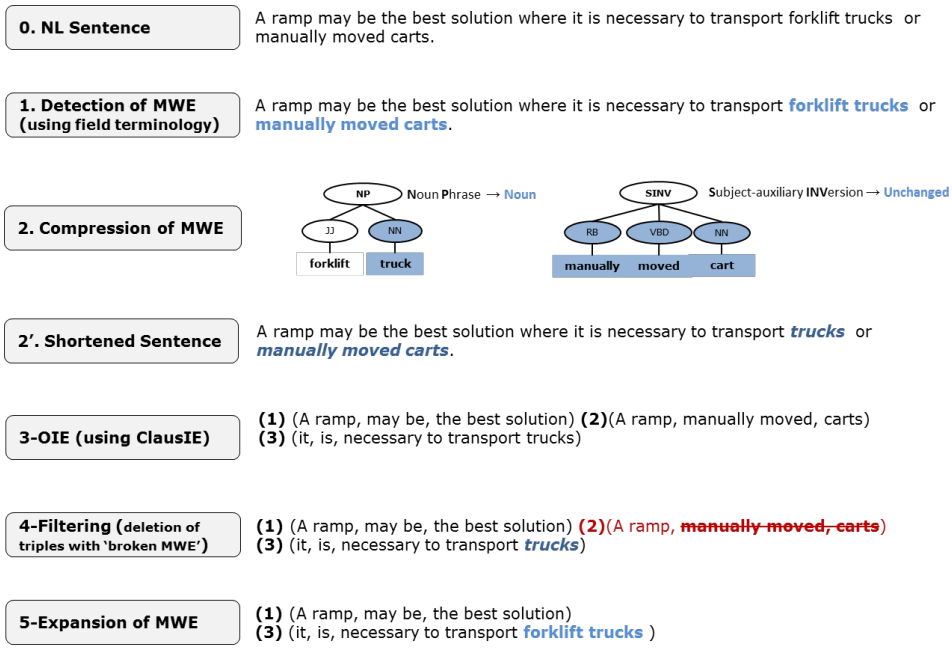

Fig. 1. An end-to-end example of Open Information Extraction when handling multi-word expressions.

frequently found in the corpus. It is the case of operators (example: less than, less than or equal to, as much as), idiomatic expressions (example: "Loose your head", "Jump in feet first"), units of measurements, etc. So, a set of MWE in a precise domain can be made up of the *terminology of the field* and *frequent terms*. This last category of terms can be obtain by means of existing statistical methods and the help of human experts. At this stage we identify the MWE present in the original sentence. We thus have a *list of possible MWE* in our corpus (see Sect. 4.1 for more details).

Step 2 - Compression of a Multi-Word Expression

The reason why precision of OIE-tools is affected by MWE is that the latter is considered by the former to be *non atomic*. Hence, to limit potential hazardous fragmentation of expressions, we propose to extract information from a new version of sentences where each MWE will have been replaced by a *shortened version*. So, now the question is: how do we get this short version of MWE? When trying to answer this question, we must have in mind that the shortened sentences must always be semantically and syntactically correct to be appropriately handled by OIE-tools. We propose the following steps for shortening a MWE (using its syntactic parse tree):

1. if the MWE is a *clause* (list of labels for clauses is available in [4]) or a *verb phrase*, *there is no shortening*;
2. else, if the MWE is a noun phrase, the first token *labelled noun* is considered to be the shortened version of the MWE;
3. else, we take the string provided by the smallest phrase[1] within the tree.

Let us note that some MWE will be short enough so that they will remain the same after the shortening. Although, such MWE (like any other MWE) is considered to be atomic. This is important to have in mind because, if an OIE-tool breaks a MWE, the resulting triple will be incorrect.

After this stage, we now perform OIE itself, which is the **third step**. This OIE is done by using existing OIE-systems. Consequently, the following steps come after OIE and take as input results of OIE, i.e triples.

Step 4 - Filtering of Open Information Extraction Results

Earlier in this work, we have pointed out a set of things which degrades precision of OIE-tools. We have focused on the problematic role caused by multi-word expressions. Now, we use the only characteristic of MWE to finalise our OIE-process. Indeed, a MWE is *unbreakable*. Consequently, when a triple contains only a fragment of a MWE, it is considered as incorrect. This filtering is done before the expansion stage, so the MWE are in their "shortened" form.

Step 5 - Expansion of a Multi-Word Expression

After the OIE has been done from the shortened version of the sentence, we now have to reconcile extracted facts with the original (long) sentence. We then look into the list of the extracted facts to replace shorten version of MWE by their initial long form. This is the aim of this step.

4 Preliminary Evaluation and Discussion

4.1 Evaluation

After making some statistics on the factors which lead to incorrect Information Extraction, we have decided to tackle the multi-word expressions-problem. The first step of the approach we propose is to identify them in the input sentence. Be able to perform such identification implies to have a list of possible MWE. That is why we hypothesize that the sentence describes the realities of a specific field of interest. Actually, to know that we are working in a specific domain implies to have a good idea of the terminology of this domain. For our evaluation, we have taken the list of terms (labels of the concepts in the field) as the set of our MWE. These terms have been obtained through a *key terms extraction process*.

[1] An exhaustive list of labels for phrases is available in the Penn Treebank [4].

We have taken advantage of existing tools (`Alchemy` [2] in our case) to carry out this extraction. For this preliminary evaluation, our corpus is made up of 50 random sentences from documents about *fire safety* [5], *energy efficiency* [6] and *accessibility* [1]. Our list of MWE consists of result provided by `Alchemy` without terms containing a proper noun. Moreover, we have added units of measurement.

To perform OIE itself after preprocessing tasks, we have used `ClausIE` of Del Corro and Gemulla [7]. In addition, we compare our results to `CSD-IE` of Bast and Haussmann [2] and to the "original" version of `ClausIE`. Results are presented by Tab. 1.

Table 1. Results of the primary evaluation of OIE using `ClausIE` (by handling MWE -thus called `ClausIE-MWE`) and `CSD-IE` (without any preprocessing) on a corpus built from law in building engineering construction.

Tools	**#extractions**	**#extractions-correct**	**#extractions-incorrect**	**#errors-due-to-MWE**
`CSD-IE`	218	127 (58,26%)	91 (41.74%)	42 (46.15%)
`ClausIE`	315	201 (63.80%)	114 (36.20%)	60 (52.63%)
`ClausIE-MWE`	165	135 (81.81%)	30 (18.89%)	9 (30%)

`CSD-IE` performance in *this domain-specific* corpus (58.26%) is less good than in "open datasets" like the Wikipedia (70.0%) and New York Times dataset (71.5%) [2]. The same remark can be made to `ClausIE`. But by handling the MWE-problem, we obtain 81.81% of correct extractions (18.19% of errors). We still have a certain number of errors. Some of these errors are caused by our handling of MWE as discussed in the next section and others errors come from OIE-tools we use at the extraction step itself.

4.2 Discussion

We have seen that handle MWE, in a given domain, helps to improve OIE on sentences of that domain. However we see that our method to handle MWE has to be improved. Indeed 30% of remaining errors after the shortening of MWE are due to that operation. Indeed:

– when we choose the first noun of a *noun phrase*-MWE to replace this MWE it is not always its suitable representative. Indeed some nouns can sometimes be *tagged as verb* and thus a *potential predicate* (e.g: "fire" in the expression *fire extinguisher*, "means" in the term *means of access*, etc.) and it can cause wrong extractions. Consequently, when we have more than a noun in a noun phrase, we must have more criteria to choose the representative.

[2] `http://www.alchemyapi.com/`

– In some sentences, parsers correctly identify the prepositional modifiers of all verbs, nouns, adverbs, etc. Consequently the *presence of MWE is a priori not a problem* for OIE-systems. Unfortunately, the deletion of prepositions (found for example in a *noun phrase*-MWE) during the shortening may lead to parsing errors. Indeed, parsers will try to identify new relations which may be wrong leading to incorrect extractions as illustrated below:
  1. `Original sentence :` "A stair is a fixed means of access."
  2. `Shortened version :` "A stair is a fixed means."
  3. `OIE : CSD-IE`→(A stair, means, is) & `ClausIE`→(A stair, a fixed means)

One of the possible solutions to avoid the shortening of MWE to escape from their multi-word problem is to replace a MWE by a *synonym*. Ideally, this synonym should have less words (why not a single word?) than the original MWE. Such synonyms could be found in Linked Open Data or in lexical databases like Wordnet, etc.

5 Conclusion

The goal of this work is to see how we can leverage domain knowledge (mainly terminology) to improve Open Information Extraction. We have focused on multi-word expressions which cause more than 45% of errors in existing OIE-tools. We have thus decided to reduce their length by proposing a shortening algorithm for multi-word terms. First results of our approach are very promising. In our goal to take advantage as much as possible of domain knowledge we can go further in facts filtering. For instance, simple *domain* and *range* constraints could help in detecting wrong facts. When we take the following fact (building, has width, door), we can state that it is incorrect by exploiting the fact that the range of the "predicate" *has width* is *xsd:float*.

References

1. American with Disabilities Act (ADA): 2010 ADA Standards for Accessible Design (sep 2010), `http://www.fire.tas.gov.au/userfiles/stuartp/file/Publications/FireSafetyInBuildings.pdf`
2. Bast, H., Haussmann, E.: Open information extraction via contextual sentence decomposition. In: Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on. pp. 154–159. IEEE Computer Society (2013)
3. Bast, H., Haussmann, E.: More informative open information extraction via simple inference. In: Advances in Information Retrieval. Lecture Notes in Computer Science, vol. 8416, pp. 585–590. Springer International Publishing (2014)
4. Bies, A., Ferguson, M., Katz, K., MacIntyre, R., Tredinnick, V., Kim, G., Marcinkiewicz, M.A., Schasberger, B.: Bracketing guidelines for treebank II Style Penn Treebank project. University of Pennsylvania 97 (1995)
5. Building Safety Unit Tasmania Fire Service: Fire Safety in Buildings (aug 2002), `http://www.fire.tas.gov.au/userfiles/stuartp/file/Publications/FireSafetyInBuildings.pdf`, obligaitions of owners and occupiers

6. California Energy Commission: 2008 Building Energy Efficiency Standards (2008), http://www.energy.ca.gov/2008publications/CEC-400-2008-001/CEC-400-2008-001-CMF.PDF, for residential and nonresidential buildings
7. Del Corro, L., Gemulla, R.: Clausie: clause-based open information extraction. In: Proceedings of the 22nd international conference on World Wide Web. pp. 355–366. WWW '13, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2013)
8. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1535–1545. EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011)
9. Mausam, Schmitz, M., Bart, R., Soderland, S., Etzioni, O.: Open language learning for information extraction. In: EMNLP-CoNLL. pp. 523–534. Association for Computational Linguistics (2012)