Evaluation of Open Information Extraction Methods Using Reuters-21578 Database

Juan M. Rodríguez
PhD Program on Computer Science
UNLP, La Plata & Intelligent Systems Group, University of Buenos Aires, ArgentinaArgentina.
Telephone number, incl. country code
jmrodriguez1982@gmail.com

Hernán D. Merlino
Information Systems Research Group
UNLa, Lanús& Intelligent Systems Group, University of Buenos Aires, Argentina.
Telephone number, incl. country code
hmerlino@gmail.com

Patricia Pesado
Patricia Pesado
III-LIDI. Computer Science School,
UNLP – CIC Bs As, La Plata, Argentina. Telephone number, incl. country code
ppesado@lidi.info.unlp.edu.ar

Ramón García-Martínez[†]
Information Systems Research Group,
UNLa, Lanús, Argentina
1st line of address
2nd line of address
Telephone number, incl. country code

ABSTRACT

The following article shows the precision, the recall and the F1-measure for three knowledge extraction methods under Open Information Extraction paradigm. These methods are: ReVerb, OLLIE and ClausIE. For the calculation of these three measures, a representative sample of Reuters-21578 was used; 103 newswire texts were taken randomly from that database. A big discrepancy was observed, after analyzing the obtained results, between the expected and the observed precision for ClausIE. In order to save the observed gap in ClausIE precision, a simple improvement is proposed for the method. Although the correction improved the precision of Clausie, ReVerb turned out to be the most precise method; however ClausIE is the one with the better F1-measure.

CCS Concepts

• **Computing methodologies→Artificial intelligence→Natural language processing→Information extraction.**

Keywords

Knowledge extraction, semantic relation extraction, self-supervised extraction, open information extraction, OIE, natural language processing, ReVerb, OLLIE, ClausIE.

1. INTRODUCTION

The main goal of this research work is to decide which knowledge extraction method (for semantic relations) is the most accurate for a given database. In this case, the chosen one was Reuters-21578, a text categorization and test collection database [1]. This collection was widely used in natural language process research projects, more specifically, in text classification works [2; 3; 4; 5].

As each newswire has a quite short text and being Reuters-21578 a well-known database, a subset of it has been chosen for this work. The selected extraction methods were those that, in accordance with the documentation research made in [6], proved to be among the top three in terms of quantity and quality of the extracted knowledge pieces.

Knowledge extraction is any technique that allows the analysis of unstructured sources of information, for instance: text in natural language, using an automated process to extract the embedded knowledge to show it in a structured form, capable of being manipulated in an automated reasoning process, for example: a production rule or a subgraph in a semantic network. Output information for this kind of process is called piece of knowledge [7; 8]. If knowledge extraction is presented as an algebraic transformation, the formula could be formulated as follows:

$$piece\_of\_knowledge = knowledge\_extraction(\mathrm{i}) \qquad (1)$$

Where i means any type of unstructured information.

Since Michele Bank, Oren Etzioni and others in [9] presented a method of knowledge extraction for the Web (or big corpuses) in 2007, many other knowledge extraction methods for the Web have been introduced. The paradigm that encompasses this type of self-supervised methods is called Open Information Extraction.

“Open Information Extraction is a paradigm that facilitates domain independent discovery of relations extracted from text and readily scales to the diversity and size of the Web corpus. The sole input to an OIE system is a corpus, and its output is a set of extracted relations. An OIE system makes a single pass over its corpus guaranteeing scalability with the size of the corpus” [9].

Semantic relation extraction methods that work in accordance with the OIE paradigm return a tuple for each semantic relation discovered. The tuple has the form (Entity 1, Relation, Entity 2), where entities are usually well-identified objects, persons, places, companies, dates, etc., and the relationship is the semantic

I CMLSC 2018, February 2–4, 2018, Phu Quoc Island, Viet Nam
© 2018 Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6336-5/18/02…$15.00
DOI: https://doi.org/10.1145/3184066.3184099

relationship between the two entities, often factual information, such as "Who did what to whom". To illustrate this, consider the following sentence:

Albert Einstein, who was born in Ulm, has won the Nobel Prize.

Extracting the relationships in the sentences and expressing them as a tuple in the form (Entity 1, Relation, Entity 2) should return the following:

- (Albert Einstein, has won, the Nobel Prize)
- (Albert Einstein, was born in, Ulm)

2. PREVIOUS WORK

A documentary investigation was carried out in [6] over a few semantic relation extraction methods, which work in accordance with the Open Information Extraction paradigm. In such work, the output quality of each method has been compared, trying to understand which one performs a better extraction.

The analyzed methods were: KnowItAll [10], TEXTRUNNER [9], WOE [11], SRL-Lund [12], ReVerb [13], OLLIE [14], ClausIE [15], ReNoun [16], TRIPLEX [17] and SONEX [18].

Such work can be summarized in table 1, which is a double entry table, where each cell must be understood as a comparison between two methods. The method indicated in the column against the method indicated in the row. The intersection cell shows the method that achieved a higher quality and quantity of extracted pieces of knowledge, regardless of the measure used in the article. References to articles, from where comparison was taken, are also given.

Table 1. Summary of comparisons between methods

Methods	B	C	D	E	F
A: KnowItAll	B[9]				
B: TEXTRUNNER		C[11,13,15]	D[12]	E[13,15]	
C: WOE				E[13,15]	F[14,15]
D: SRL – Lund					D[14]
E: ReVerb					F[14,17]
F: OLLIE					E[15]
G: ClausIE	G[15]	G[15]		G[15]	G[15]
I: TRIPLEX				I, I+E[17]	F, I+F[17]

The methods SONEX [18] and ReNoun [16] were analyzed in [6], but in the articles in which they were presented, there wasn't a strict comparison against other existing methods in order to determine their relative performance.

Some preliminary conclusions can be drawn:

- The best studied method, in terms of quantity and quality of knowledge pieces extracted is ClausIE.
- Since TRIPLEX in combination with OLLIE is only slightly better than OLLIE alone, we would expect that ClausIE exceeds it in precision.
- After ClausIE, the next methods are: OLLIE, ReVerb, and WOE, in that order (sorted by quality), but in accordance with the test case used, one method could outperform the other.

3. EXPERIMENT

3.1 Objective

The goal of this experiment is to obtain a reliable estimation about which of these three methods: ReVerb, OLLIE and ClausIE, the top three methods in accordance with documentation research [6], obtains a better precision, recall and F1-meassure for a given database. Precision, recall, and F-measure will be calculated using the following formulas:

$$\text{Precision} = \frac{\text{relevant extracted knowledge pieces}}{\text{extracted knowledge pieces}} \quad (2)$$

$$\text{Recall} = \frac{\text{relevant extracted knowledge pieces}}{\text{(handmade relation extractions + new extracted pieces)}} \quad (3)$$

$$F_{\beta} = \frac{(1+\beta^2) \cdot \text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}} \quad (4)$$

The new extracted pieces in 3 are the relevant extracted knowledge pieces that are not in the handmade set.

In formula 4 we select the parameterβ equal to 1, so that precision and recall have the same weight. For simplicity purposes, F-measure will be called F1-measure or just F1.

To calculate the confidence level and the associated margin of error for a given number of samples, the following formulafor sample size determination will be used[19]:

$$n = \frac{N \cdot Z^2 \cdot p \cdot (1 - p)}{(N-1) \cdot e^2 + Z^2 \cdot p \cdot (1-p)} \quad (5)$$

Where:

- N is the total number of newswire articles in Reuters-21578 (which is 21578)
- Z is the deviation from the mean accepted to achieve the desired level of confidence
- p is the ratio we hope to find (for an unknown sample, 50% is usually taken)
- e is the maximum permissible margin of error

The research goal is to obtain a confidence level of 95% with a maximum margin of error of 10%, meaning that 96 newswire articles need to be evaluated, taken randomly from the Reuters-21578 database. In this work, 103 newswire articles, taken randomly from Reuters-21578 were evaluated, so the confidence level is slightly over 95%.

A preliminary result of this research work was presented in [20].

3.2 Evaluation

The first part of this work focused on performing a semantic relation extraction manually for each selected newswire. During this part of the experiment, we received the help of several senior students of Computer Engineering. The semantic relation extraction procedure was explained to them, but minor details were left to the discretion of each one of them. Finally, there was a revision of output in order to unify certain criteria. To observe an example of these handmade extractions, let's see the text in of the newswire with id 44:

…McLean Industries Inc's United States Lines Inc subsidiary said it has agreed in principle to transfer its South American service by arranging for the transfer of certain charters and assets to Crowley Mariotime Corp's American Transport Lines Inc subsidiary. U.S. Lines said negotiations on the contract are expected to be completed within the next week. Terms and conditions of the contract would be subject to approval of various regulatory bodies, including the U.S. Bankruptcy Court…

The following semantic relations were obtained manually:

- (McLean Industries Inc; is subsidiary of; United States Lines Inc)
- (McLean Industries Inc; said; it has agreed in principle to transfer its South American service by arranging for the transfer of certain charters and assets to Crowley Mariotime Corp's American Transport Lines Inc subsidiary)
- (McLean Industries Inc; has agreed to transfer; its South American service by arranging for the transfer of certain charters and assets to Crowley Mariotime Corp's American Transport Lines Inc subsidiary)
- (U.S. Lines; said; negotiations on the contract are expected to be completed within the next week)
- (negotiations on the contract; are expected to be completed; within the next week)
- (Terms and conditions of the contract; would be; subject to approval of various regulatory bodies)

3.3 Verification

The next step was to run the methods over the same 103 Reuters articles and made a validation by hand for each automatic extraction. A category of three values was used: right, invalid and almost-right. This last value was used for extractions that where in the borderline, when it was difficult to see if the extraction was right or not. An extraction marked as almost-right was not taken into consideration for precision and recall calculations, in this way a penalization for doing an almost-right job was avoided, or a double penalization if we think in the F1-measure. This value (almost-right) was also used to avoid computing two right extractions very similar to each other twice, where the only difference between them was in the second entity (typically in ClausIE). For instance:

- (it; has agreed; to transfer its South American service)
- (it; has agreed; in principle to transfer its South American service)

Both extractions were correct and both made reference to the same sentence. In this particularly case the first one was marked as right and the second was marked as almost-right. A second consideration we had, before marking an automatic extraction as right, was to identify if there was a manual extraction to match with the automatic one, in other words, we verified that manual extraction and automatic extractionrefers to the same sentence and to the same relation, regardless of minor details. Continuing with the same example, the following manual extraction:

- (McLean Industries Inc; has agreed to transfer; its South American service by arranging for the transfer of certain charters and assets to Crowley Mariotime Corp's American Transport Lines Inc subsidiary)

Was considered equivalent to the following automatic extraction:

- (it; has agreed; to transfer its South American service)

Even when, there were differences between relations and entities, both of them referred to the same semantic relation. When a valid automatic extraction was identified, but there was no match with any handmade extraction, it was marked as *new*. So when calculating recall, the amount of valid relations was computed as all handmade extractions plus all automatic extraction marked as *new* (for a given method), these are the *new extracted pieces* in formula 3.

3.3.1 ClausIE features

Before measuring the results (output) of the methods, we have to look for ClausIE features in order to standardize our measurements. The creators of ClausIE [15] tested its method using different modalities:

- With processing of coordinated conjunctions
- Without processing of coordinated conjunctions
- Counting redundant extractions as correct
- Ignoring redundant extractions

In the experiment described in this article, ClausIE was executed without processing coordinated conjunctions, and redundant extractions were ignored (they were marked as almost-right).

3.3.1.1 ClausIE coordinated conjunctions

The processing of coordinated conjunctions is a modality where ClausIE splits a sentence using its conjunctions to create different relations from a single sentence. For example, the sentence:

Bell makes and distributes electronic, computer and building products.

will produce the following semantic relations:

- (Bell, makes, electronic products)
- (Bell, makes, computer products)
- (Bell, makes, building products)
- (Bell, distributes, electronic products)
- (Bell, distributes, computer products)
- (Bell, distributes, building products)

But by default ClausIE doesn't use this modality. By default ClausIE will produce:

- (Bell, makes, electronic computer and building products)
- (Bell, distributes, electronic computer and building products)

The coordinated conjunction modality could be useful, but it increments the amount of correct extractions for a single sentence and the rest of the evaluated methods, OLLIE and ReVerb works without this modality. They produce extractions in the same way as ClausIE working by default. For these reasons, in this research, ClausIE was run in its default modality.

3.3.1.2 ClausIEredundant extractions

Redundant extractions are extractions contained in other extractions. ClausIE usually generates redundant extractions. Using an example given in [15], the following sentence:

Albert Einstein remained in Princeton until his death.

will generate the following redundant extractions:

- (Albert Einstein, remained, in Princeton)
- (Albert Einstein, remained, in Princeton until his death)

In these cases, only one extraction was marked as correct, the other was just ignored (marked as almost-right).

3.4 Expected results

In accordance with what was presented in section 2, we expected the best method to be ClausIE, followed by OLLIE and then by ReVerb. This preliminary conclusion was obtained from [15], which is the only article that made an evaluation of ClausIE (evaluation made by its authors). They compared ClausIE precision with the precision of others methods: OLLIE, ReVerb, WOE, and TextRunner. In that article these methods were tested against three different datasets:

- 200 sentences randomly extracted from the New York Times collection
- 200 sentences randomly extracted from Wikipedia pages
- 500 sentences randomly extracted from the service Yahoo's random link

The obtained precision is summarized in table 2

Table 2. Precision expected[1]

Datasets	Precision		
	ClausIE	**OLLIE**	**ReVerb**
ReVerb	0.615	0.440	0.534
Wikipedia	0.670	0.414	0.663
NYT	0.648	0.425	0.550
All datasets	**0.633**	0.431	0.563

3.5 Actual results

Precision, recall and F1-measure obtained for ClausIE, OLLIE, and ReVerb after the evaluation of these three methods against the subset of 103 newswire texts from Reuters-21578 is summarized in table 3.

As shown in such table, precision for OLLIE is within the expected order, precision for ReVerb is in a greater order of magnitude, and precision for ClausIE is lower than expected. The difference between the expected and the calculated precision for ClausIE is approximately 0.17, such discrepancy is not likely to be a measurement error. After some experiments, we started to think that the length of the input text affected the output in ClausIE. ClausIE was using every word in the input text to construct a dependency parser. This behavior produced a complex tree of dependencies; sometimes it worked fine and ClausIE found correct semantic relations. However, in many other cases, we suppose this kind of work was the main responsible for the errors encountered in the extracted semantic relationships.

Table 3. Calculated measures

Measure	Methods		
	ClausIE	**OLLIE**	**ReVerb**
Precision	0.467	0.456	**0.633**
Recall	**0.519**	0.416	0.319
F1-Measure	**0.492**	0.435	0.424

4. MAKING A CORRECTION IN CLAUSIE

In order to confirm the text length hypothesis, a new function was added to ClausIE to split each input text into independent sentences.

The main idea was to split the input text into independent sentences using the same parser (Stanford parser) that ClausIE used to build the dependency tree. But the parser didn't work as expected, particularly with abbreviations and acronyms (and Reuters texts have lots of these types of words). Some sentences were split in the middle because a dot at the end of an abbreviation was confused with a period. So, an English dictionary of abbreviations was added in order to detect possible abbreviations in the text. Also regular expressions were used to detect acronyms and numbers.

First, a regular expression search was performed using patterns for numbers and then other search was made using patterns for acronyms. For each text portion matched with a given pattern, all dots were replaced by a wildcard character. After extracting the sentence from the input text, the wildcard character was replaced by the original character.

To detect possible abbreviations, two other regular expressions were used, one for searching double abbreviations, for example: "Nat. Hist.", and other for searching simple abbreviations. In each case, the matched portion of the text was verified against the dictionary of abbreviations, and if it matched with an abbreviation, all dots in that text were replaced by a wildcard character. Table 4 shows the regular expressions used:

Table 4. Used regular expressions

Regex use	Pattern
Find numbers	\d+\.\d+
Find acronyms	(?:[a-zA-Z]\.){2,}
Find double abbreviations	\w+\. \w+.
Find single abbreviations	\w+\.

5. NEW RESULTS FOR CLAUSIE

A new version of ClausIE was developed using the solution described in section 4 and it was executed against the 103 newswire texts from Reuters-21578. Precision and recall improved considerably. The new values obtained are shown in table 5.

Table 5. New calculated measures for ClausIE

Measure	Methods	
	ClausIE	**Modified ClausIE**
Precision	0.467	0.602
Recall	0.519	0.641
F1-Measure	0.492	0.621

Figure 1 shows differences in precision and recall from ClausIE original method and ClausIE with the split-text-into-sentences modification.

[1] Calculated in [15]

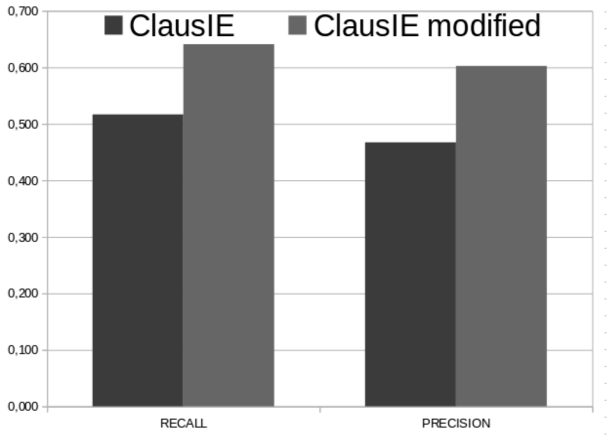

Figure 1.Precision and recall for original and modified ClausIE.

5.1 Process time improved

As a result of the new functionality added to split text into sentences, the time that ClausIE takes to process an input text has been improved considerably. Although it was not the main goal of this research (which was to improve precision), it is a positive improvement. In particular for the context where ClausIE is designed to run productively, those are Big Data corpuses, Web, etc.

It is difficult to measure the average time ClausIE takes to process an input text. It depends on the length of the text. The longer a text, the more complex the syntactic dependency tree that ClausIE constructs will be.

The longest text of the subset of 103 cases from Reuters-21578, which was a single paragraph of 248 words and 1620 characters, was processed by ClausIE in 6 minutes and 13.914 seconds. However, the new version of ClausIE (withthe modification for splitting text) took only 16.6 seconds to process all the text.

Time was measured with the time line command tool for Linux in the following way:

```
time ./clausie.sh -f  18745.txt
```

Computer specifications and Java version used in this experiment are summarized in table 6.

Table 6. Workstation specification

Specification	Values
CPU	AMDPhenom(tm)8450Triple-Core Processor
RAM	4 Gb RAM DIMM DDR2S ńcrono 333 MHz
OS version	Linux Mint 17.2 Rafaela
Linux kernel version	GNU/Linux 3.16.0-38-generic x86_64
Java version	Java version "1.7.0_80". Java HotSpot(TM) 64-Bit Server VM (build 24.80-b11, mixed mode)

6. CONCLUSIONS

Despite the considerable improvement in precision, which is closer to the value calculated in [15], ReVerb is still the method with the highest precision for the corpus of Reuters-21578. However, ClausIE is the method with the highest F1-measure because ReVerb has a poor recall. This result allows us to confirm the values obtained by the authors of ClausIE, also confirm that the method is still doing a good work even with news texts brought from a different source.

The problem encountered also reveals the weak point of ClausIE. As an input text grows, the dependency tree will be more likely to fail, which means it will be more likely to extract incorrect semantic relationships. If the text does not have its constituent sentences delimited or if it is a single but very long sentence, the chances of success for ClausIE are reduced.

Finally, keeping in mind that the introduced functionality improved both precision and processing time of ClausIE, the conclusion is that this functionality should be definitively incorporated into the code and it should be part of its default behavior.

7. REFERENCES

[1] Lewis, D. (1997). “Reuters-21578 text categorization test collection”. Accessed on 05/20/2017 http://goo.gl/NrOfu.

[2] Joachims, T. (1998). “Text categorization with support vector machines: learning with many relevant features”. Machine learning: ECML-98, 137-142.

[3] Steinbach, M., Karypis, G. and Kumar, V. (August 2000). “A comparison of document clustering techniques”. KDD workshop on text mining,Vol. 400, No. 1, 525-526.

[4] Yang, Y. and Liu, X. (August 1999). “A re-examination of text categorization methods”. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development information retrieval, ACM,42-49.

[5] Zhao, Y. and Karypis, G. (2001). “Criterion functions for document clustering: Experiments and analysis”, (Vol. 1, p. 40). Technical report.

[6] Rodr ǵuez, J. M., Garc á-Mart ńez, R.and Merlino, H. D. (2015). “Revisi ón sistem ática comparativa de evoluci ón de m étodos de extracci ón de conocimiento para la web”.Proceedings of XXI Congreso Argentino de Ciencias de la Computaci ón (CACIC 2015), Buenos Aires, Argentina.

[7] Rancan, C., Kogan, A., Pesado, P. and Garc á-Mart ńez, R. (2007).“Knowledge discovery for knowledge based systems”. Some experimental results. Res. Comput. Sci. J. 27, 3-13.

[8] G ómez, A., Juristo, N., Montes, C. and Pazos, J. (1997). “Ingenier á del conocimiento”. Editorial Centro de Estudios Ram ón Areces. ISBN, 84, 8004,269-9.

[9] Banko, M., Cafarella, M. J., Soderland, S., Broad-head, M. and Etzioni, O. (January 2007). “Open information extraction for the web”. IJCAI, Vol. 7, 2670-2676.

[10] Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., et al. (2005). “Unsupervised named-entity extraction from the web: An experimental study”. Artificial intelli-gence, 165(1), 91-134.

[11] Wu, F. and Weld, D. S. (July 2010). “Open information extraction using Wikipedia”. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 118-127.

[12] Christensen, J., Soderland, S. and Etzioni, O. (June 2011). “An analysis of open information extraction based on semantic role labeling”. Proceedings of the sixth international conference on Knowledge capture, ACM,113-120.

[13] Fader, A., Soderland, S. and Etzioni, O. (July 2011). “Identifying relations for open information extraction”. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Asso-ciation for Computational Linguistics, 1535-1545.

[14] Schmitz, M., Bart, R., Soderland, S. and Etzioni, O. (July 2012). “Open language learning for information extraction”. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computation-al Linguistics,523-534.

[15] Del Corro, L. and Gemulla, R. (May 2013).”ClausIE: clause-based open information extraction”. Proceedings of the 22nd international conference on World Wide Web. International World Wide Web Conferences Steering Committee, 355-366.

[16] Yahya, M., Whang, S. E., Gupta, R. and Halevy, A. (October 2014). “Renoun: fact extraction for nominal attributes”. Proc. 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar.

[17] Mirrezaei, S. I., Martins, B. and Cruz, I. F. (2015). “The triplex approach for recognizing semantic relations from noun phrases, ppositions, and adjectives”. The Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data (Know@LOD) co-located with Extended Semantic Web Conference (ESWC), Portoroz, Slovenia.

[18] Mesquita, F., Merhav, Y. and Barbosa, D. (2010). “Extracting information networks from the blogosphere: state-of-the-art and challenges”. Proceedings of the Fourth AAAI Conference on Weblogs and Social Media (ICWSM), Data Challenge Workshop.

[19] Lubov, A. and Hamburg, M. (1979). “Study guide to accompany: basic statistics: a modern approach”. Harcourt Brace Jovanovich.

[20] Rodr ǵuez, J. M., Merlino, H. D., Pesado, P. and Garc á-Mart ńez, R. (August 2016). “Performance evaluation of knowledge extraction methods”. International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer International Publishing, 16-22.