

Performance Evaluation of Knowledge Extraction Methods

Juan M. Rodríguez^{1,2,3}, Hernán D. Merlino^{2,3}, Patricia Pesado⁴,
and Ramón García-Martínez³(✉)

¹ PhD Program on Computer Science,

National University of La Plata, La Plata, Argentina

jmrodriguez1982@gmail.com

² Intelligent Systems Group, University of Buenos Aires,

Buenos Aires, Argentina

jmrodriguez1982@gmail.com, hmerlino@gmail.com

³ Information Systems Research Group,

National University of Lanús, Lanús, Argentina

rgml960@yahoo.com

⁴ III-LIDI. Computer Science School,

National University of La Plata – CIC Bs As, La Plata, Argentina

ppesado@lidi.info.unlp.edu.ar

Abstract. This paper shows the precision, the recall and the F-measure for the knowledge extraction methods (under Open Information Extraction paradigm): ReVerb, OLLIE and ClausIE. For obtaining these three measures a subset of 55 newswires corpus was used. This subset was taken from the Reuters-21578 text categorization and test collection database. A handmade relation extraction was applied for each one of these newswires.

1 Introduction

The goal of this research is to decide which knowledge extraction method (for semantic relations) is the more accurate one for a given database. In this case, the chosen was Reuters-21578, a text categorization and test collection database (Lewis 1997). This collection was widely used in natural language process research projects; more specifically in text classification works (Joachims 1998). 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 research. The selected extraction methods were those that, according with the state of the art research made in (Rodríguez et al. 2015), proved to be among the top three in terms of quantity and quality of the extracted knowledge pieces.

Knowledge extraction is any technique which allows the analysis of unstructured sources of information, for instance: text in natural language, using an automated process to extract the embedded knowledge in order to show it in a structured form, capable of being manipulated for an automated reasoning process, for instance: a production rule or a sub graph in a semantic network. The output information for this kind of process is called: piece of knowledge (Rancan et al. 2007). If knowledge extraction is presented as an algebraic transformation, the formula could be formulated as follows:

$$\text{knowledge_extraction}(\text{information_structures}) = \text{piece_of_knowledge}. \quad (1)$$

Since (Banko et al. 2007) presented a method of knowledge extraction for the Web, many other knowledge extraction methods for the Web have been presented. The paradigm that encompasses this type of self-supervised methods is called Open Information Extraction.

Open Information Extraction (OIE) 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 (Banko et al. 2007).

2 Related Work

A state of the art research was made (Rodríguez et al. 2015) over a set of eight relevant semantic relation extraction methods; methods which work according Open Information Extraction paradigm. In our previous work the quality of each method's output has been compared, trying to understand which one performs a better extraction than other. The analyzed methods were: KnowItAll (Etzioni et al. 2005), TEXTRUNNER (Banko et al. 2007), WOE (Wu and Weld 2010), SRL-Lund (Christensen et al. 2011), ReVerb (Fader et al. 2011), OLLIE (Schmitz et al. 2012), ClausIE (Del Corro and Gemulla 2013), ReNoun (Yahya et al. 2014) y TRIPLEX (Mirrezaei et al. 2015). Our comparison work is summarized in Table 1, which is a double entry table, where each cell must be understood as a comparison made between two methods. The method indicated in the column against the method indicated in the row. The intersection cell shows the method that has achieved a higher quality and quantity of extracted pieces of knowledge (approximated), regardless of the measure used in the article. References from where comparison was taken, are also given.

Table 1. Summary of comparisons between methods

| Methods | TextRunner | WOE | SRL-Lund | ReVerb | OLLIE | ClausIE | ReNoun | TRIPLEX |
|------------|-------------------------|----------------------|-----------------------|-----------------------|-----------------------|----------------------|--------|---|
| KnowItAll | TextRunner ^a | | | | | | | |
| TextRunner | | WOE ^{b,e,i} | SRL-Lund ^d | ReVerb ^{e,i} | | ClausIE ⁱ | | |
| WOE | | | | ReVerb ^{e,i} | OLLIE ^{f,i} | ClausIE ⁱ | | |
| SRL-Lund | | | | | SRL-Lund ^f | | | |
| ReVerb | | | | | OLLIE ^{f,h} | ClausIE ⁱ | | TRIPLEX, TRIPLEX + ReVerb ^h |
| OLLIE | | | | | ReVerb ⁱ | ClausIE ⁱ | | OLLIE, TRIPLEX + OLLIE ^h |
| ClausIE | | | | | | | | |
| ReNoun | | | | | | | | |
| TRIPLEX | | | | | | | | |

References: a. (Banko et al. 2007), b. (Wu and Weld 2010), c. (Mesquita et al. 2010), d. (Christensen et al. 2011), e. (Fader et al. 2011), f. (Schmitz et al. 2012), g. (Yahya et al. 2014), h. (Mirrezaei et al. 2015), i. (Del Corro and Gemulla 2013)

We can draw the following preliminary conclusions:

- [i] The best studied method, in terms of quantity and quality of knowledge pieces extracted, is ClausIE.
- [ii] Since TRIPLEX in combination with OLLIE is only slightly better than OLLIE alone, we would expect that ClausIE exceeds TRIPLEX+OLLIE in precision.
- [iii] If we consider again quantity and quality of knowledge pieces extracted, after ClausIE, the next methods are: OLLIE, ReVerb and WOE, in that order.

3 Experiment

The goal of this experiment is to obtain a reliable estimation about which of these three methods: ReVerb, OLLIE o ClausIE (the top three methods according to our state of the art research), has the better precision, the better recall and the better F-measure for a given database. The precision, recall, and F-measure will be calculated using the following formulas:

$$\text{precision} = \frac{\text{amount of relevant extracted knowledge pieces}}{\text{amount of extracted knowledge pieces}} \quad (2)$$

$$\text{recall} = \frac{\text{amount of relevant extracted knowledge pieces}}{(\text{amount of handmade relation extraction} + \text{new extracted pieces})} \quad (3)$$

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

The “new extracted pieces” in formula (3), are the relevant extracted knowledge pieces which are not in the handmade set. In formula (4), the selected value for β is 1, for simplicity the 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 formula (see Hamburg 1979) for sample size determination will be used:

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

Where:

- N: total number newswire articles in Reuters-21578 (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 error

The research goal is to get the highest confidence level with a maximum margin error of 10 %. According to the formula (5), the current confidence level will be 86 %, to know which of the three evaluated method would be the preferred one to extract semantic relations of the Reuters-21578 database, with the established error margin.

3.1 Manual Extraction

The first part of our experiment was to develop a semantic relation extraction manually, for each selected newswire of the selected subset. During this part of the experiment we were helped with several senior students of Information Engineering Bachelor Degree level. The semantic relation extraction procedure was explained to them. Finally, the authors made a revision of the students extraction work. To show an example of these handmade extractions, let's see 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)

3.2 Verification

The next step was to run the methods over the same 55 Reuter's articles and made a validation by hand for each automatic extraction. A category of three values was used: right, invalid and more-or-less-right. This last value was used for extractions in a limit, when was difficult to see if the extraction was right or not. An extraction marked as more-or-less-right was not taken into consideration for obtaining the precision and recall, in this way a penalization for do a more-or-less-right extraction was avoided, or a double penalization if we think in the F1-measure. This value (more-or-less-right) was also used to avoid compute twice two right extractions very similar each other,

extractions where the only difference was in the second entity scope (typically in ClausIE). For the automatic extraction made by ClausIE over newswire with Id 44, the following two are of our interest:

- *(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 was marked as right and the second was marked as more-or-less-right. A second consideration we had made, before mark an automatic extraction as right, was to identify if there was a manual extraction to match with the automatic one, in other words we check that both extractions that made reference 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 though, in this case, there were differences within the two entities and differences with the relation too; the manual version and the automatic version were considered semantically equals. Then, for cases where a valid automatic extraction was identified, but there was not matching with any handmade extraction, it was marked as “new”. So at the moment of calculate the recall, the amount of valid relations was computed as all the handmade extractions plus all the automatic extraction marked as “new” (for a given method).

4 Results

The Table 2 shows (for each method) a summary of the total amount of automatic extractions made, the right ones and the total of the valid semantic relations, calculated in the way just described. The precision, recall and F1-measure were calculated using values in Table 2, with the formulas (2), (3) and (4). The obtained results are shown in Table 3.

Table 2. Summary of relations and extractions for each method.

| Method | Total relations | Right extractions | Total extractions |
|---------|-----------------|-------------------|-------------------|
| ClausIE | 650 | 327 | 638 |
| ReVerb | 569 | 202 | 301 |
| OLLIE | 633 | 266 | 545 |

Table 3. Precision, Recall and F1-measure for each method.

| Method | Precision | Recall | F1-measure |
|---------|-----------|--------|------------|
| ClausIE | 0.513 | 0.503 | 0.508 |
| ReVerb | 0.671 | 0.355 | 0.464 |
| OLLIE | 0.488 | 0.420 | 0.451 |

5 Conclusions and Future Research

According with results summarized in the Table 1, ClausIE should have obtained the higher precision, followed by OLLIE and then by ReVerb but the obtained results contradict these assumptions. What we see is that ReVerb is the method with a higher precision, followed by ClausIE and finally by OLLIE. But if we see the obtained recall, this value is consistent with the expected results. ClausIE has a better recall than OLLIE, and OLLIE a better recall than ReVerb. ReVerb extracts less semantic relations than ClausIE or OLLIE (see in Table 2, 301 against 638 and 545), but the valid extraction percentage is bigger. To conclude, the F1-measure shows that ClausIE has a better F1-measure than ReVerb and OLLIE. The ReVerb F1-measure is a little better than OLLIE but they have almost the same value, the difference is only 0.013.

The next step in our research is to increase the evaluated newswires to 96 in order to get a confidence level of 95 % to establish which one of the three methods is the best to extract the semantic relations in Reuters-21578 database.

Acknowledgments. The research reported in this paper was partially funded by Projects UNLa-33A205 and UNLa-33B177 of National University of Lanus (Argentina). Authors wish to thank to senior students in our courses within Information Engineering Bachelor Degree at Engineering School - University of Buenos Aires for their help during the experiment.

References

- Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction for the web. In: IJCAI, vol. 7, pp. 2670–2676, January 2007
- Christensen, J., Soderland, S., Etzioni, O.: An analysis of open information extraction based on semantic role labeling. In: Proceedings of the Sixth International Conference on Knowledge Capture, pp. 113–120. ACM (2011)
- 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. International World Wide Web Conferences Steering Committee, May 2013
- Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Unsupervised named-entity extraction from the web: an experimental study. *Artif. Intell.* **165**(1), 91–134 (2005)
- 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. Association for Computational Linguistics, July 2011
- Hamburg, M.: Basic Statistics: A Modern Approach. Jovanovich, New York (1979)
- Joachims, T.: Text categorization with support vector machines. In: Nédellec, C., Rouveirol, C. (eds.) Learning with many relevant features, pp. 137–142. Springer, Heidelberg (1998)
- Lewis, D.D.: Reuters-21578 text categorization test collection, distribution 1.0. <http://www.research.att.com/~lewis/reuters21578.html>
- Mesquita, F., Merhav, Y., Barbosa, D.: Extracting information networks from the blogosphere: State-of-the-art and challenges. In: Proceedings of the Fourth AAAI Conference on Weblogs and Social Media (ICWSM), Data Challenge Workshop (2010)

- Mirrezaei, S.I., Martins, B., Cruz, I.F.: The triplex approach for recognizing semantic relations from noun phrases, appositions, and adjectives. In: *The Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data (Know@LOD) Co-located with Extended Semantic Web Conference (ESWC)*, Portoroz, Slovenia (2015)
- Rancan, C., Kogan, A., Pesado, P., García-Martínez, R.: Knowledge discovery for knowledge based systems. Some experimental results. *Res. Comput. Sci. J.* **27**, 3–13 (2007)
- Rodríguez, J.M., García-Martínez, R., Merlino, H.D.: *Revisión Sistemática Comparativa de Evolución de Métodos de Extracción de Conocimiento para la Web*. XXI Congreso Argentino de Ciencias de la Computación (CACIC 2015), Buenos Aires, Argentina (2015)
- Schmitz, M., Bart, R., Soderland, S., Etzioni, O.: 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*, pp. 523–534, July 2012
- Wu, F., Weld, D.S.: Open information extraction using Wikipedia. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 118–127. Association for Computational Linguistics, July 2010
- Yahya, M., Whang, S.E., Gupta, R., Halevy, A.: Renoun: fact extraction for nominal attributes. In: *Proceedings 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, October 2014

Trends in Applied Knowledge-Based Systems and Data
Science

29th International Conference on Industrial Engineering
and Other Applications of Applied Intelligent Systems,
IEA/AIE 2016, Morioka, Japan, August 2-4, 2016,
Proceedings

Fujita, H.; Ali, M.; Selamat, A.; Sasaki, J.; Kurematsu, M.
(Eds.)

2016, XVIII, 1021 p. 371 illus., Softcover

ISBN: 978-3-319-42006-6