An Approach Based on Patterns for Synonymy Relations Detection

M Tovar1*, G Flores-Petlacalco1 and J A Reyes-Ortiz2

1Faculty of Computer Science, Benemérita Universidad Autónoma de Puebla, 14 sur y Av. San Claudio, C.U., Puebla, 72570, Mexico
2Universidad Autónoma Metropolitana, Av. San Pablo Xalpa 180, Azcapotzalco, 02200, Ciudad de México, Mexico

*Email: mireya.tovar@correo.buap.mx

Abstract. Thesaurus, taxonomies, and ontologies are created as a way to represent knowledge. In recent years, the efforts to create them are focused on automatic knowledge extraction. For this purpose, research in the extraction of concepts and semantic relations is needed. Concepts or entities extracted from unstructured data has a become a challenge for the research communities. Texts, specifically, scientific papers contain knowledge where experts express their contributions to the science, including: techniques, methods and claims. In order to describe them, experts use synonyms, which are relevant to consider whereas it is desirable to extract such type of contribution form scientific text. This paper describes a method based on patterns for semantic relations extraction from scientific texts. The semantic relations correspond to synonymy. We describe pattern acquisition and relations extraction as phases to data evaluation. The evaluation shows promising results with 75% precision, 58% recall and 65% F-measure.

1. Introduction

A large amount of information created daily requires of knowledge organization techniques that sort and represent the resources available. Thesaurus, taxonomies, and ontologies are examples of the effort to represent knowledge. For years, they were created manually. However, in recent years, the efforts are focused on creating them automatically, by using terms selection and recovery of semantic relations between terms [1].

In order to create thesaurus, taxonomies, and ontologies, concepts or entities need to be extracted from unstructured. This task is proposed from the Natural Language Processing field and it has a become a challenge for the research communities because the inherent complexity for the treatment of natural language texts. So that, researchers through scientific papers in natural texts, express their contributions in the science. In this manner, scientific texts describe the process, task, material, techniques, methods or claims related to the author's contributions. Such information is generalized as keyphrases or fundamental objects. But authors, also, use synonyms to refers the same keyphrases, which are relevant to consider whereas it is desirable to extract such type of contribution form scientific text. Identifying this information manually is a tedious and time-consuming task for analysts of scientific literature. Therefore, computational techniques based on NLP should be explored and proposed for this task. In addition, consider existing linguistic resources to support computational methods.
A semantic relation refers to the association between words and their meanings, as a way of giving sense and coherence to a text. There are several types of semantic relations. For example: Hyponymy-Hyperonymy, in this relation, the meaning of a word is included in other or relations of type. Meronymy-Holonymy, it is a relationship that establishes a connection between the meaning of a word that indicates a part with the meaning of another word that indicates the whole, that is, in the term that includes it. Other semantic relation is synonymy, in which one meaning can be related with several words, which can be used indistinctly depending on the context. The relation of type synonymy is a relation of identity between meanings. This relationship has two features: The meaning has the same semantic content and the two words can be freely exchanged in any situation. The synonym relation is a typical lexical semantic relation, which is included in most lexical databases and ontologies, for example WordNet. These synonym relations are useful in a number of NLP and text mining applications, such as: question-answering, information retrieval, text summarization, language generation and recommendation [2].

In recent years, the interest in semantic relations research has increased. Tasks such as identification, classification and automatic representation of semantic relationships have been addressed. In this way, the International Workshop on Semantic Evaluation (SemEval) was created. It constantly proposes tasks in the Natural Language Processing area. For the 2017 edition, it has presented a task related to an extraction of semantic relations between keyphrases. SemEval 2017 Task 10: "Extracting Keyphrases and Relations from Scientific Publications" (see https://scienceie.github.io/index.html) contains a total of three subtasks [3].

1) Identification of keyphrases.
2) Classification of identified keyphrases.
3) Extraction of relationships between two identified keyphrases.

In this paper, we are focused on partially solving the third subtask: "Extraction of relationships between two identified keyphrases". The main objective is to detect relations of type Hyponymy-Hyperonymy, as well as relations of synonymy between keyphrases. In this paper, a system for detecting relations of synonymy is proposed. The system finds equivalent concepts to help to the enrichment of language resources and prevent the unnecessary repetition of terms. In addition, it can serve as a starting point for future research in this area.

The approach uses a pattern-based approach for extracting synonymy relation. The patterns are detected by mapping identified keyphrases with a semantic relation confirmed between them from scientific texts. Later, the detected patterns are complemented with others, obtained from literature. The system is evaluated on a set of keyphrases previously identified from a test data set.

This paper is organized as follows. The next section, the related work about research in the extraction of semantic relations area is presented and compared. Section 2 presents the pattern-based proposed approach, which involves two main phases: a) pattern acquisition that consist of obtaining the confirmed relations, mapping keyphrases and obtaining the final patterns; b) relation extraction, which involves a preprocessing text, keyphrase matching and validation of tuples. In Section 3 the results and discussions of the approach are shown, also the dataset is explained. Finally, Section 4 presents the final conclusion and future work.

2. Related Work
Related works in the task of semantic relations extraction are shown in this section. There is a great variety of semantic relationships, such as: synonymy, hyponymy, hyperonymy, and meronymy. Extracting semantic relations in texts has been an object of study for several authors. Researches about them are presented as follows.

Reference [4] proposes a method for extracting relations of type Hyponymy-Hyperonymy from texts using patterns. At first, a set of base patterns is obtained through observation. These are later complemented by using the bootstrapping method on known lexical collections in the area of interest. This approach is a complement to statistic-based methods and gets encouraging results. Furthermore, this approach can be translated to obtain other type of semantic relations.
Reference [5] based on Hearts [4] and Ortega's work [6], create a system for the extraction of hyponymy in web texts for the Spanish language. They use pattern extraction to locate pairs or words with a relation in web text fragments. Each pair is assigned a trust value, so that only pairs above a minimum limit are considered, and pairs with wrong classification are deleted. To calculate a confidence value, the Ortega patterns are used as a reference, as well as a list of the seeds that created them. However, that value will change as more pairs of hyponyms are extracted from a pattern.

Reference [7] proposes a method for extracting patterns of synonymy in texts and sort them. They search and mark phrases that can contain references to synonym relation in Wikipedia articles, using the first paragraphs with bold type, redirect links in the body, occurrence of the title, and pages of disambiguation of terms. Later, from the marked text, a pattern extraction is made. It is nothing but the extraction of text fragments between terms, which are then evaluated to get a confidence value using the harmonic mean. Fragments of high value are considered as new patterns and added to a thesaurus. Finally, patterns that start with numbers and nouns are erased.

Reference [8] proposes a method for extracting patterns of synonymy in texts and sort them. They search and mark phrases that can contain references to synonym relation in Wikipedia articles, using the first paragraphs with bold type, redirect links in the body, occurrence of the title, and pages of disambiguation of terms. Later, from the marked text, a pattern extraction is made. It is nothing but the extraction of text fragments between terms, which are then evaluated to get a confidence value using the harmonic mean. Fragments of high value are considered as new patterns and added to a thesaurus. Finally, patterns that start with numbers and nouns are erased.

Reference [9] uses Convolutional Neural Networks (CNN) for extracting semantic relations between keyphrases. The networks are trained by using semantic relation features between confirmed keyphrases that are represented in a matrix. Such features are: their positions, and the words that comprise both the terms involved and the text between them. To make predictions, a pair of terms is represented in a vector, similar to the training matrix, that is created to determine the type of semantic relations between them. The semantic relation can be hyponymy-of, synonymy-of or not-related.

Reference [10] presents a method for automatic extract hyponym-hypernym relations from texts. The method uses a dependency parser and regular expressions, next it extracts patterns from the dependency graph of a preselected sentence.

As seen in the review of the state of the art, the semantic relation extraction (synonymy) from scientific text has been neglected. Therefore, the approach presented in this paper performs a pattern extraction on fragments of scientific texts, where the mapping a set of keyphrases with confirmed semantic relation is carried out on the training dataset. The candidate patterns are discriminated to get the most relevant and are complemented with others found in the literature. Finally, the set is evaluated to identify semantic relations of synonymy between identified key phrases on the testing dataset.

3. Material and Methods
The proposed approach has two phases in order to identify semantic relations, specifically synonymy. In the first phase the patterns are acquired from the training dataset, and in the second phase the set of patterns is evaluated on a test dataset based on relations extraction task.

3.1. Patterns Acquisition
In this section, a method used to get the extraction patterns is described. It was carried out on a training dataset, containing a text extract where the relationships between keyphrases should be identified, and an annotation file with keyphrases identified and semantic relations between each phrase. The steps for obtaining the patterns are listed below.

1) Obtaining of confirmed relations. From the training dataset, the start and end position of each pair of keyphrases with a confirmed semantic relation of synonymy, is obtained, together with the own keyphrase.

2) Keyphrases mapping. Positions are sorted from lowest to highest and the portion of text between the lowest position to highest position is extracted.
3) Obtaining the pattern. From the text extracted in the previous step, the keyphrases are removed, to conserve only the words or characters between them.

After using our extraction method, several patterns are obtained, which are complemented with literature: [11], [2], [12] and [13]. Finally, 17 patterns are gathered as shown in Table 1.

| Pattern | Pattern | Pattern |
|---------|---------|---------|
| $S_1 \ (S_2)$ | $S_1$ alias $S_2$ | $S_1$ (is $\|$ are)$^+$ called $S_2$ |
| $S_1$ is often referred to as $S_2$ | $S_1$ aka $S_2$ | $S_1$ sometimes called $S_2$ |
| $S_1$ is referred to as $S_2$ | $S_1$ frequently abbreviated as $S_2$ | $S_1$ know as $S_2$ |
| $S_1$ are also called $S_2$ | $S_1$ called as $S_2$ | $S_1$ also referred to as $S_2$ |
| $S_1$ is given by $S_2$ | $S_1$ as know as $S_2$ | $S_1$ often described $S_2$ |
| $S_1$ abbreviated a $S_2$ | $S_1$ usually called $S_2$ | $S_1$ commonly know as $S_2$ |

3.2. Relations Extraction
The extraction of semantic relations of synonymy between keyphrases has been performed on a test dataset, with keyphrases identified previously, that were provided by the task organizers. The steps for determining whether a pair of keyphrases have a semantic relation of synonymy between them are listed below.

3.2.1. Preprocessing. The extracted text is read, and later split into sentences, using the function `sent_tokenize` (see http://www.nltk.org/api/nltk.tokenize.html) from the NLTK package. This way the starting and finishing positions of each sentence in the text are obtained.

3.2.2. Keyphrase Matching. The keyphrases and their positions in the text are read from the annotation file. Later, each keyphrase and its position get compared with the start and end position from each sentence. In other words, whether the keyphrase position is within the range of start and end position of a sentence, it is considered that this keyphrase belongs to the sentence and is added to a set of type: `{sentence, {Keyphrase_1, Keyphrase_2, ...}}`

3.2.3. Evaluation of Tuples. To extract the semantic relations between keyphrases, the next steps are followed. The keyphrases from each set obtained in the last step are combined in pairs using the function `itertools.combinations` (see https://docs.python.org/3/library/itertools.html) from Python. Later, for each pair of keyphrases, the next steps are performed, for determining whether a semantic relation of synonymy exists or not.

a) Keyphrase formatting. Each element from the pair is formatted as follows: at the beginning of the keyphrase NP_ is added, and each space is replaced by ".".

b) Replacing keyphrases. The formatted keyphrases from the previous step are replaced into the sentence.

c) Cleaning the sentence. The sentence with the keyphrases replaced is cleared of punctuation marks and character sets that match the regular expression ".*".

d) Matching patterns. Each sentence created in the previous step is matched with the patterns obtained previously. Coincidences are searched by the use of regular expressions. Whether a pattern matches within the text, it is determined that it is a pair of keyphrases with a semantic relation of type synonymy, and the pair is added to a set.

e) The set obtained, in before steps, that contained the list of semantic relations and their keyphrases are matched exactly with the results of the gold-standard. The precision, recall, and $F_1$-Measure metrics are used in this step.
4. Results and Discussion

The performance of our system is evaluated using a script, provided by the organizers of Task 10 of SemEval 2017, and a gold-standard, the metrics used are the well-known measures Precision, Recall and $F_1$-Measure [14].

The corpus provided by the organizer of Task 10 of SemEval 2017 includes 500 articles from the free collection of the ScienceDirect website (see http://www.sciencedirect.com). Each article has three files, an XML file with the complete article, to be used by the developers whereas they need it, a TXT file with an abstract of the article, to be used as a workspace where the keyphrases and semantic relations are extracted for evaluating, and finally, an ANN file with annotations of the identified keyphrases in the text extract, and the semantic relation of type synonymy or hyponymy-hyperonymy detected between them. The corpus is divided into three parts, 50 were provided as a development set, 350 as a training set, and finally, 100 as a testing dataset for evaluation. The domain considers three areas: Physics, Material Science, and Computer Science, and all articles are in the English language.

In the evaluation of the proposed approach, 400 scientific papers (development and training sets) are used for acquiring patterns and 100 scientific papers for evaluating them into synonyms identification task. All experiments are executed over a testing dataset. We analyse the results by each semantic relation: hyponym and synonym.

Table 2 shows the vocabulary information for the development, training and test datasets. Table 3 shows the synonymy relations total included in each dataset.

| Table 2. Datasets. |
|-------------------|
| Corpus            | Paragraph | Vocabulary |
| Development       | 50        | 2105       |
| Training          | 350       | 8079       |
| Test              | 100       | 4330       |

| Table 3. Semantic Relations. |
|-------------------------------|
| Relation | Development | Training | Test |
|----------|-------------|----------|------|
| Hyponymy | 123         | 418      | 95   |
| Synonymy | 45          | 253      | 112  |

4.1. Evaluation

The algorithm was evaluated using a gold-standard suggested by the organizers. First, the system was evaluated, without the cleaning sentence step, indicated in the section Relations Extraction. Later, an evaluation was done including this step, results are shown in Table 4.

| Table 4. Proposed approach results. |
|-------------------------------------|
| Relation   | P   | R   | $F_1$ |
|------------|-----|-----|-------|
| Not cleaning | 0.48 | 0.60 | 0.53  |
| Cleaning    | 0.75 | 0.58 | 0.65  |

An increment in the Precision is shown, the cleaning step has achieved better results. In Table 5 a comparison is shown compared with other participants in the task, only in the detection of semantic relations of synonymy.
Table 5. Comparison with other approaches.

| Team            | P    | R    | $F_1$  |
|-----------------|------|------|--------|
| MIT             | 0.81 | 0.81 | 0.81   |
| NTNU2           | 0.77 | 0.65 | 0.70   |
| Proposed approach | 0.75 | 0.58 | 0.65   |
| LaBDA           | 0.61 | 0.46 | 0.52   |
| TITI_COIN_rel   | 0.16 | 0.15 | 0.15   |

This approach obtained a good position in the evaluation, compared with other teams in the same task, that use approaches based on Neural Networks, or classification using Vector Support Machines, Naïve Bayes or Decision trees, according to those mentioned by [3] in the task description document. Our approach has achieved a 65% F-measure by considering both semantic relations (hyponym and synonym).

5. Conclusion

This paper has presented a system based on patterns for extracting semantic relations, specifically synonymy relations. First, the patterns were obtained by mapping keyphrases identified on scientific texts from ScienceDirect to extract the text between them. Later, the candidate patterns were evaluated for determining those that can be used in an extraction, and were complemented with others found in the state of the art. Finally, the system was evaluated on a test dataset, obtaining promising results only in detection of semantic relations of type synonymy in the texts.

The main contributions of this work are (1) an approach for patterns acquisition in order to characterize the synonymy relation between concepts in scientific domain; (2) a method for extracting synonym relations from scientific texts; (3) a set of patterns which are available to use and evaluate in other domains; (4) an evaluation based on precision, recall and $F_1$-measure, achieving promising results. The evaluation process has shown promising results, it has obtained a 65% F-measure by considering both semantic relations (hyponym and synonym). Even though the precision of our approach is not so high, it is very useful for analysts of scientific literature since they can reduce the time spent in locating such semantic relations manually.

The synonym relations discovered between concepts from scientific texts can be used to enhancing available resources like Wordnet. Also, they could be useful for constructing a knowledge base in order to support several tasks, such as: retrieval scientific information, semantic search engines or classification of scientific texts. Regarding to search engine, a question answering system could be constructed in order to answer typical questions for helping researchers and practitioner by using synonyms keyphrases: which papers have addressed a specific TASK?, which papers have utilized such MATERIALS? or which papers have addressed this TASK using variants of this PROCESS?

For future work, it is planned to research on other techniques, based on classifiers and characteristics with the objective of comparing the performance between different approaches.

Acknowledgments

This work is partially supported by the Sectorial Research Fund for Education with the CONACyT CB/257357 project and by the VIEP-BUAP 100409344-VIEP2019 project. The authors also would like to thank Universidad Autonoma Metropolitana, Azcapotzalco, with the research project SI001-18.

References

[1] Sánchez-Cuadrado S, Lloréns J, Morato J and Hurtado JA 2003 Extracción automática de relaciones semánticas In: 2da. Conferencia Iberoamericana en Sistemas, Cibernética e Informática. CISCI 2003 265-268
[2] Hu F, Shao Z and Ruan T 2015 Self-Supervised Synonym Extraction from the Web. *J Inf Sci Eng.* 31(3) 1133-1148

[3] Augenstein I, Das M, Riedel S, Vikraman L and McCallum A. SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017).* (Vancouver, Canada: Association for Computational Linguistics) 2017. pp 546–555

[4] Hearst M A 1992 Automatic Acquisition of Hyponyms from Large Text Corpora. In: *Proceedings of the 14th Conference on Computational Linguistics – Volume 2. COLING ’92* (Stroudsburg, PA, USA: Association for Computational Linguistics) pp 539-545

[5] Ortega R M, Aguilar C, Villaseñor L, Montes M and Sierra G 2011 Hacia la identificación de relaciones de hiponimia/hiperonimia en Internet. *Revista signos.* 44(75) 68-84

[6] Ortega R M 2007 Descubrimiento Automático de Hipónimos a partir de Texto no Estructurado. (Instituto Nacional de Astrofísica, Óptica y Electrónica, México)

[7] Simanovsky A, Ulanov A 2011 Mining text patterns for synonyms extraction. In: *Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on.* IEEE pp 473-477

[8] Muller P, Hathout N and Gaume B 2006 Synonym extraction using a semantic distance on a dictionary. In: *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing.* (Association for Computational Linguistics) pp 65-72

[9] Sheth A, Aleman-Meza B, Arpina I, Bertram C, S Warke Y, Ramakrishnan C, et. Al 2005 Semantic Association Identification and Knowledge Discovery for National Security Applications. *J Database Manag.* 16 33-53

[10] Uparate M G J, Mane P S U 2018 Automatic Information Extraction from Text. *Journal NX – A Multidisciplinary Peer Reviewed Journal.* 2018 132-137

[11] Lopez O L A 2013 Extracción automática de relaciones léxico-semánticas a partir de textos especializados [Tesis Doctoral]. (Universidad Nacional Autónoma de México, Posgrado en Ciencia e Ingeniería de la Computación)

[12] Nazar R, Renau I 2012 A co-occurrence taxonomy from a general language corpus. *J. Fjeld y Torjusen* 2012 367-375

[13] Yu H, Agichtein E 2003 Extracting Synonymous Gene and Protein Terms From Biological Literature. *Bioinformatics (Oxford, England).* 19 Suppl 1 340-9

[14] Tolosa G H, Bordignon F R 2008 *Introducción a la Recuperación de Información.* (Tolosa y Bordignon)