ISSN 2007-9737

Proposal for Named Entities Recognition and Classification (NERC) and the Automatic Generation of Rules on Mexican News

Orlando Ramos Flores, David Pinto

Benemérita Universidad Autónoma de Puebla,
Faculty of Computer Science,
Language and Knowledge Engineering Laboratory,
Mexico

orlandxrf@gmail.com, dpinto@cs.buap.mx

Abstract. In this paper, we introduce a proposal for extracting facts from news on Mexican online newspapers through their RSS (Really Simple Syndication). This problem will be addressed by using the task of automatic named entities recognition and classification (NERC), as well as the semantic relation extraction among entities, so that we can build a database of facts and rules from the obtained entities in an automatic manner. The final aim is to be able to infer new rules through the use of the knowledge databases constructed and an inference engine. In order to build the NER model, we perform a manual annotation of corpora with different tags that include the baseline tags (person names, organizations, locations, dates and numeral). The proposed idea is presented in this paper with an example scenario together with the procedure employed for solving the problem of automatic inference of new rules.

Keywords. NERC, semantic relations, facts-base, rules, Spanish news.

1 Introduction

Nowadays the Named Entity Recognition (NER) is a task widely addressed by researchers in different domains and languages such as English, Arab, Turk, Hindi, and Spanish. NER task is relevant because most of the time it must focus in a specific domain o language. The main application areas of Named Entity Recognition are: Information Extraction, Question-Answering, Machine Translation, Automatic Text Summarization, Text Clustering, Information Retrieval, Knowledge-Base or Ontology Population, Opinion Mining, and Semantic Search [10]. In this paper we introduce a proposal to recognize named entities and classify them in order to obtain semantic relations between two given entities, as well as to propose the automatic generation of rules in order to use them for validation and consistency from entities in the facts-base. The data employed for the experiments has been collected from the newspapers of states (at least one newspaper per state) of the Mexican Republic using a crawler system. In order to build the NER model we previously will manually annotate these news using baseline tags and new tags like “brand”, “event”, “age”, “measure”, “time”, “entertainment”, “laws”, “alias”, among other tags not defined yet.

The rest of this paper is organized as follows. Section 2 describes the related work about NER and Relation Extraction. Section 3 introduce our idea with an example for automatic recognition of named entities, classification of such entities and the relation extraction process proposed to transform facts in a facts-base and the automatic generation of rules. The conclusions are presented in the Section 4.

2 Related Work

The term Named Entity Recognition (NER) was coined in the Message Understanding Conference (6th edition); this is a task widely used in Information Extraction (IE) for identifying people names, organizations and geographical locations

in a raw text. It can be also employed for identifying numeric expressions such as currency and percentages [11]. Identifying references to these entities in raw texts was recognized as one of the most important sub-tasks of IE and was called “Named Entity Recognition and Classification (NERC)”[16].

2.1 Named Entity Recognition and Classification

Different learning methods have been proposed for NERC. While early studies were mostly based on handcrafted rules, most recent ones use supervised machine learning (SL) as a way to automatically induce rule-based systems or sequence labeling algorithms starting from a collection of training examples.

Nevertheless, when training examples are not available, handcrafted rules remain the preferred technique. The main shortcoming of SL is the requirement of a large annotated corpus. The unavailability of such resources and the prohibitive cost of creating them lead to two alternative learning methods: semi-supervised learning (SSL) and unsupervised learning (UL) [16]. Currently [10] classifies NERC in three main techniques: Rule-based approaches, Learning-based approaches, and Hybrid approaches.

2.1.1 Rule-based Approaches

These techniques are often based on handcrafted rules, including the use of information lists such as gazetteers, as well as rules based on syntactic-lexical patterns to identify and classify named entities. Rule-based NERC systems are considered highly efficient because they exploit the properties of language-related knowledge.

However, some limitations of these systems are that they are quite expensive, domain-specific and non-portable. Furthermore, these systems require human expertise with regard to knowledge of the domain and language along with programming skills [10].

2.1.2 Supervised Learning

Supervised learning based approaches are based on the idea of providing labeled training data involving positive and negative examples This learning approaches typically consist of a system that reads a large annotated corpus, memorizes lists of entities, and creates disambiguation rules based on discriminative features. The labeled data are then used to train the learning model which is further used to recognize and classify named entities out of unannotated or test data, that is why it is normally named: training data.

A baseline method that is often proposed consists of tagging words of a test corpus when they are annotated as entities in the training corpus [10, 16]. The main techniques used are: Conditional Random Fields (CRF) [8, 13], Hidden Markov Models (HMM) [2, 4], Maximum Entropy Models (ME) [5, 21], and Support Vector Machines [3, 21], even though other technique may be found in literature.

2.1.3 Semi-Supervised Learning

Traditional classifiers require a considerable amount of annotated training data. For this reason SSL uses both labeled and unlabeled corpus to make their hypothesis. The main technique is called “*bootstrapping*” and involves a small degree of supervision, such as a set of seeds, for starting the learning process. The results are then used to re-train the system to generate more labeled examples. This process continues to several times to make the learning decisions refined [10, 16].

2.1.4 Unsupervised Learning

Unsupervised learning is a method that uses information which is neither classified nor labeled. The goal of unsupervised learning is to generate a model that considers the structural and distributional features of data in order to find more information about the data that allows to categorize it. The typical approach in unsupervised learning is clustering and association rules-based approach. Clustering based approach uses distributional statistics to extract named entities out of unlabeled data by making use of context similarity.

Association rules-based technique is concerned with finding associations among items within large databases [10, 16].

2.2 Relation Extraction

Relation Extraction (RE) is the task of detecting and classifying predefined relationships between entities identified in raw text. The main approaches reported in literature for RE are rule-based methods and statistic-based methods [6].

2.2.1 Rule-based Approach

Rule-based approaches need to predefine rules that describe the structure of entity mentions. These methods require the rules builder to have a deep understanding of the background and characteristics of the field.

Hence, the obvious drawbacks are the huge demand of human participation and poor portability. An example can be seen in [17], where they use an unsupervised method and rule-based for extracting semantic relations from entities in the music domain.

2.2.2 Statistic-based Approaches

These approaches are classified according to [6] as:

1) Unsupervised method: It extracts strings of words between entities in a large amount of text, and clusters and simplify these word strings to produce relation-strings. However, since there is no standard form of relations, the output resulting may not be easy to map to relations, which is necessary for a particular knowledge base [6, 7].

2) Semi-supervised method: It uses the bootstrapping technique, it is similar to that we mentioned above. These methods typically suffer of semantic change and poor precision. An example of this approach is Snowball [1].

3) Supervised method: It is the most common used method for relation extraction, and obtains relatively high performance, and it is considered as a classification task. The supervised method can be simply divided into two types: feature-based methods and kernel-based methods.

4) Distant supervision method: It automatically generates training examples, and learns features through aligning raw text with Knowledge Bases (KBs) such as Freebase or DBpedia, a large semantic database. Thus the method does not need any human intervention and can extract vast numbers of features from a large amount of data.

5) Neural Network method: It is one of the early methods, and depends on the quality of the extracted features derived from the existing NLP tools. In this way, the errors are inevitably produced during the processing. Hence the resurgence of neural network (NN) provides the new insight into such a problem. The neural network was first applied to relation classification by [19].

3 Example Scenario

In this section, we introduce an example scenario with the aim of showing the proposed idea for automatic recognition of entities in news genre, and to classify them, so that we can create a facts-base and identify rules in an automatic manner, and validate them employing logic inference using a tool such as SWI-Prolog. The dataset to be used in the experiments are news from digital newspapers written in Mexican Spanish.

3.1 Recognize Entities in News

Firstly, we must identify entities such as "*person names*", "*locations*", "*organizations*", and "*dates*". In Figure 1, we show a news excerpt about NAFTA (North American Free Trade Agreement) and we assume that we have recognized the entities which have been marked (annotated) with its corresponding tag. One of the main problems of NERC is disambiguation among entities [9, 14, 16], e.g., the entities of "*Guajardo*", "*Ildefonso Guajardo Villarreal*", "*Ildefonso Guajardo*", and "*Guajardo Villarreal*" are all the same person. Even if, as human we are able to identify them in an easy

Fig. 1. Spanish news about North American Free Trade Agreement (NAFTA)

way, computer machines face a very complicated challenge for completing the task [10, 12, 16].

3.2 Facts Base and Rules

Knowing that we have recognized all entities in the text, the following step is to identify the semantic relations between two given entities using the verb [18, 20], the nominal phrase [15, 21], in statistic way [9] or with another approach shown in Section 2.2. Assuming that we have used some approach to RE, we would have some relations of the type <Entity, Relation, Entity>, as the ones listed below:

— <Guajardo_PER, *estima_reanudar*, TLCAN_ORG>

— <Robert_Lighthizer_PER, *representante_comercial*, EU_LOC>

— <Ildefonso_Guajardo_Villarreal_PER, *titular_de*, Secretaría_de_Economía_SE_ORG>

— <Guajardo_Villarreal_PER, *llegar_a_un_acuerdo*, Norteamérica_LOC>

The semantic relations above will become facts of the **facts-base** in the following way:

— *person*(Guajardo),

— *person*(Ildefonso_Guajardo_Villarreal),

— *person*(Guajardo_Villarreal),

— *person*(Robert_Lighthizer),

— *location*(EU),

— *location*(Norteamérica),

— *organization*(TLCAN),

— *organization*(Secretaría_de_Economía_SE).

Now, if we assume that we have applied the disambiguation process to entities (e.g. from the "*Guajardo*" entities we can set just select one entity that represents all of them in which "*Guajardo*" appears), we can create rules from the facts-base, identifying the same entities between different facts, so that we can assume that a relation among them exist (e.g. by using the entity-linking in order to validate and/or disambiguate the entity employing knowledge-bases such as DBpedia, Wikidata or GeoNames aid us to find relations between entities). Some possible **rules** could be:

- *estimaReanudar*(X,Y):-person(X),organization(Y).
- *representanteComercial*(X,Y):-person(X),location(Y).
- *titularDe*(X,Y):-person(X),organization(Y).
- *llegarAunAcuerdo*(X,Y):-person(X),location(Y).
- ***sonColaboradores***(X,Y):-person(X),person(Y),organization(W),location(Z), titularDe(X,W),representanteComercial(Y,Z).
- ***perteneceA***(X,Y):-location(X),location(Y).

From the rules shown above, the first four ones are equivalent to the semantic relations previous found. The last two rules are our main goal, i.e., the aim is to identify new possible rules in an automatic manner (e.g. the last two rules). The penultimate rule refers to the process of finding the relationship among people that work together for a specific purpose, and the last rule refers to the act of belonging to a country of a continent part. This is just an example of the main idea proposed in this paper, we neither consider the facts and rules validation nor to check its consistency and other relevant points about logic inference in this moment. We just focus in introducing our proposal for this research work.

4 Conclusions

In this paper we introduce a proposal to Named Entity Recognition and Classification (NERC). We also consider the identification of semantic relations between entities so as to transform them in a facts-base in order to be able to automatically generate rules from the entities recognized.

This is not a trivial task, since we have to consider different subtasks like obtaining a model for entities recognition from the news. Thus, we plan to use Stanford NER, and other tools and models in Spanish for this task. Additionally, we are considering to perform manual annotation of the corpus gathered.

Finally, entity disambiguation, entity linking, validation, and consistency must be a number of subtask that need to be addressed as well.

As a future work, we will implement this proposal. Up to now, we are in the process of collecting a huge number of news from digital newspapers from Mexico country. Thereafter, we will use this approach together with NERC and automatic generation of rules for the construction of a knowledge graph about Mexican news.

References

1. **Agichtein, E. & Gravano, L. (2000).** Snowball: Extracting relations from large plain-text collections. *Proceedings of the Fifth ACM Conference on Digital Libraries*, DL '00, ACM, New York, NY, USA, pp. 85–94.
2. **Amarappa, S. & Sathyanarayana, S. (2013).** Named entity recognition and classification in kannada language. *International Journal of Electronics and Computer Science Engineering*, Vol. 2, No. 1, pp. 281–289.
3. **Asahara, M. & Matsumoto, Y. (2003).** Japanese named entity extraction with redundant morphological analysis. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 8–15.
4. **Bikel, D. M., Miller, S., Schwartz, R., & Weischedel, R. (1997).** Nymble: A high-performance learning name-finder. *Proceedings of the Fifth Conference on Applied Natural Language Processing*, ANLC '97, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 194–201.
5. **Borthwick, A., Sterling, J., Agichtein, E., & Grishman, R. (1998).** Nyu: Description of the mene named entity system as used in muc-7. *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*.

6. **Cui, M., Li, L., Wang, Z., & You, M. (2017).** A survey on relation extraction. **Li, J., Zhou, M., Qi, G., Lao, N., Ruan, T., & Du, J.**, editors, *Knowledge Graph and Semantic Computing. Language, Knowledge, and Intelligence*, Springer Singapore, Singapore, pp. 50–58.
7. **Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., & Yates, A. (2005).** Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, Vol. 165, No. 1, pp. 91–134.
8. **Finkel, J. R., Grenager, T., & Manning, C. (2005).** Incorporating non-local information into information extraction systems by gibbs sampling. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 363–370.
9. **Galicia-Haro, S. N., Gelbukh, A., & Bolshakov, I. A. (2004).** Recognition of named entities in Spanish texts. **Monroy, R., Arroyo-Figueroa, G., Sucar, L. E., & Sossa, H.**, editors, *MICAI 2004: Advances in Artificial Intelligence*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 420–429.
10. **Goyal, A., Gupta, V., & Kumar, M. (2018).** Recent named entity recognition and classification techniques: A systematic review. *Computer Science Review*, Vol. 29, pp. 21–43.
11. **Grishman, R. & Sundheim, B. (1996).** Message understanding conference-6: A brief history. *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, volume 1.
12. **Hoffart, J. (2013).** Discovering and disambiguating named entities in text. *Proceedings of the 2013 SIGMOD/PODS Ph.D. Symposium*, SIGMOD'13 PhD Symposium, ACM, New York, NY, USA, pp. 43–48.
13. **Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001).** Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 282–289.
14. **Meijer, K., Frasincar, F., & Hogenboom, F. (2014).** A semantic approach for extracting domain taxonomies from text. *Decision Support Systems*, Vol. 62, pp. 78–93.
15. **Mirrezaei, S. I., Martins, B., & Cruz, I. F. (2015).** The triplex approach for recognizing semantic relations from noun phrases, appositions, and adjectives. **Gandon, F., Guéret, C., Villata, S., Breslin, J., Faron-Zucker, C., & Zimmermann, A.**, editors, *The Semantic Web: ESWC 2015 Satellite Events*, Springer International Publishing, Cham, pp. 230–243.
16. **Nadeau, D. & Sekine, S. (2007).** A survey of named entity recognition and classification. *Lingvisticae Investigationes*, Vol. 30, No. 1, pp. 3–26.
17. **Oramas, S., Espinosa-Anke, L., Sordo, M., Saggion, H., & Serra, X. (2016).** Information extraction for knowledge base construction in the music domain. *Data & Knowledge Engineering*, Vol. 106, pp. 70–83.
18. **Punuru, J. & Chen, J. (2012).** Learning non-taxonomical semantic relations from domain texts. *Journal of Intelligent Information Systems*, Vol. 38, No. 1, pp. 191–207.
19. **Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012).** Semantic compositionality through recursive matrix-vector spaces. *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, pp. 1201–1211.
20. **Vasques, D. G., Zambon, A. C., Baioco, G. B., & Martins, P. S. (2016).** An approach to knowledge acquisition based on verbal semantics. *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pp. 4144–4153.
21. **Xuan, H. N. T., Le, A. C., & Nguyen, L. M. (2012).** Linguistic features for subjectivity classification. *2012 International Conference on Asian Language Processing*, pp. 17–20.

Article received on 29/10/2019; accepted on 07/03/2020.
Corresponding author is David Pinto.