CSI Peru News: finding the culprit, victim and location in news articles

Gina Bustamante **Arturo Oncevay**
Artificial Intelligence Research Group
Pontificia Universidad Católica del Perú
gina.bustamante@pucp.edu.pe, arturo.oncevay@pucp.edu.pe

Abstract

Distant Supervision (DS) methods have become widely used in Relation Extraction (RE) and been applied successfully in several fields. However, complications arise when we want to employ DS in less-studied domains such as biomedical, crime-related, among others, which involve non-famous entities and unique relations not found in large databases such as Freebase and DBpedia. Thus, we introduce a shift on the DS method over the domain of crime-related news from Peru. We attempted to find the culprit, victim and location of a crime description from a RE perspective. Obtained results are highly promising and show that proposed modifications are effective in low-resourced domains.

1 Introduction and Background

Distant Supervision for Relation Extraction (Mintz et al., 2009) relies on the assumption that, given a triplet $(e1, r, e2)$, every sentence containing both entities $e1$ and $e2$ also includes the relation r. Nevertheless, this hypothesis has some drawbacks in particular cases. First, in challenging domains with highly diverse content where there is no guarantee that $e1$ or $e2$ will be found in another sentence even if the relationship r does. Second, some domains imply complex relations that are not found between two entities but within an entity and an action (verb).

Previous work in domain-specific Relation Extraction proposes the use of already created knowledge bases related to the field (Aljamel et al., 2015), or its construction through ontology learning (Dasgupta et al., 2017). However, we argue that such expensive work is not necessary and that improving NER alongside the use of words related to the field and dependency trees is enough.

By implementing the mentioned approach, we look for contributing to diminishing the problems aforementioned and propose an alternative for DS in low-resourced domains. Additionally, we improve our initial results with neural methods and prove their usefulness in generalising pattern-based IE.

2 Customised Distant Supervision

2.1 Dataset

Large databases such as Freebase (Bollacker et al., 2008) and DBpedia (Lehmann et al., 2015), cannot cover relations of specific domains as Peruvian crime-related news. Hence, we chose that field to develop a novel dataset by collecting five thousand news from different Peruvian digital newspapers websites. Using this collection, we attempt to learn relations that could lead to the identification of the culprit, victim and location given a crime description.

2.2 Pattern-based Extraction

In this section, we describe how we established patterns using dependency trees to classify relations. For every target class, we defined ten patterns connecting a person to a culprit or victim related term, and a location to a crime-related term.

To achieve the mentioned task, we first fine-tuned the Named Entity Recognition (NER) module of the SpaCy library [1] with 500 annotated sample news (10% of our entire dataset). This step was required

[1] Official site: https://spacy.io/usage/spacy-101

in our context because crime-related articles commonly avoid specifying the names of the culprit or victim. Besides, NER systems tend to fail with locations they have not seen before in the training scope. Annotations contained information about places in Peru (initially unrecognised by the module) and a new type of entity called *person-reference* for texts that do not mention names explicitly. Otherwise, culprit, victim and location relations cannot be found between two entities, but within an entity and an action (verb). Consequently, we constructed a Lexical Resource (LR) containing words and actions associated with the three kinds of relations we want to classify.

Having the fine-tuned NER and Lexical Resource, to obtain our patterns, we manually annotated the culprit, victim, location and LR terms of 5000 sentences contained in 500 news. Next, we used SpaCy dependencies parser to get the dependency tree between a culprit/victim/location and the LR term related to it. Finally, we identify the 10 dependency tree structures more frequent for each class and set them as patterns.

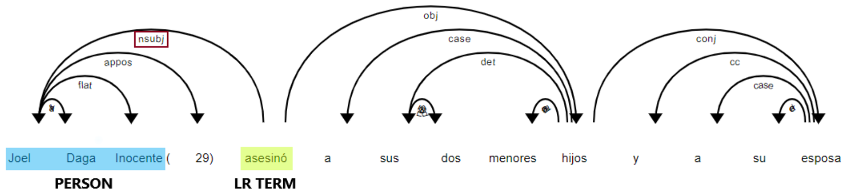

Figure 1: Example of a dependency tree structure in a sentence

In summary, we rely on the fact that if a sentence contains a person, a culprit-related word and the dependency tree between them match one of the established patterns, we can say that this person is the culprit of the described crime. An example of this is in Figure 1, where $Joel\ Daga\ Inocente$ is a named entity, and $asesino$ is a term in our LR related to the culprit. As between those two words the syntactic dependency is $nsubj$ (nominal subject), and it matches one of our patterns for the culprit, we output $Joel\ Daga\ Inocente$ as the responsible of the described crime.

2.3 Neural Networks for Sequence Labelling

In this experiment, we attempt to generalise RE (increase recall) by using Neural Networks. We hypothesise that a neural model can extract and learn more features apart from the dependency trees initially employed. Using the patterns defined in the previous step, we generated a lot of noisy samples from all our dataset and saved them in IOB format. With this data, we trained a CRF-tagger consisting of a bidirectional LSTM encoder and a CRF layer (Huang et al., 2015).

3 Results and Discussion

To evaluate our proposed approaches we built a test set with 500 documents from our data (different from the ones used in section 2.2), and for each sentence in the selected news, we manually annotated the culprit, victim and location. In our analysis, we start with a baseline that chooses randomly if the given person is the culprit or victim, and if a place was or not the location of a crime. With this method, we obtained poor precision and therefore a low F1-score of 33%.

Method	precision	recall	f1-score
Baseline	0.25	0.40	0.33
Pattern-based (general-domain NER)	0.68	0.20	0.44
Pattern-based (fine-tuned NER)	**0.75**	0.38	0.57
CRF-tagger	0.70	**0.50**	**0.60**

Table 1: Proposed methods results over test set.

In the pattern-based approach, we did two experiments comparing general-domain Spacy NER module against our fine-tuned NER. With the first one, we got an F1-score of 44% with a precision of 68% and

20% of recall. Otherwise, using the fine-tuned NER, we improve previous results to 57% of F1-score with 75% and 38% of precision and recall, respectively. Finally, using the neural CRF-tagger, we achieved a significant increase in recall and a final F1-score of 60%. Below in Figure 1, we can see a sample of accurate results from the CRF-Tagger.

Los agentes de La PNP capturaron a Seida Lucero Cosavalente Escalante quien mató a puñaladas a su pareja.

['O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-CULPRIT', 'I-CULPRIT', 'I-CULPRIT', 'I-CULPRIT', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

Figure 2: Example of CRF-tagger results.

4 Conclusions and future work

Described outcomes in the Pattern-based experiments evidence the need for a domain-specific NER module for this kind of fields since without it, recall metric is pretty low. On the other hand, it does not seem to affect precision significantly.

We conclude that our approach, using entity types and words from a lexical resource, combined with dependency trees as patterns, let us successfully classify relations in low-resourced domains. Besides, we see that pattern-based methods tend to have high precision and low recall, whereas neural methods improve these initial results, providing us with a more generalised model.

Finally, given the fact that we do not have a large amount of labelled data, as future work we can use a pre-trained encoder (Howard and Ruder, 2018) in the CRF-tagger to take advantage of large corpus with raw data, similar to the experiments done with pre-trained word embeddings (Mikolov et al., 2013).

References

Abduladem Aljamel, Taha Osman, and Giovanni Acampora. 2015. Domain-specific relation extraction: Using distant supervision machine learning. In *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, volume 1, pages 92–103. IEEE.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, New York, NY, USA. ACM.

Tirthankar Dasgupta, Abir Naskar, Rupsa Saha, and Lipika Dey. 2017. Crimeprofiler: crime information extraction and visualization from news media. In *Proceedings of the International Conference on Web Intelligence*, pages 541–549. ACM.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6:167–195.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore, August. Association for Computational Linguistics.