Incorporating Lexico-semantic Heuristics into Coreference Resolution Sieves for Named Entity Recognition at Document-level

Marcos Garcia
Grupo LyS, Dep. de Galego-Portugués, Francés e Lingüística
Fac. de Filoloxía, Universidade da Coruña
Campus da Coruña, 15701, Coruña, Galicia, Spain
marcos.garcia.gonzalez@udc.gal

Abstract

This paper explores the incorporation of lexico-semantic heuristics into a deterministic Coreference Resolution (CR) system for classifying named entities at document-level. The highest precise sieves of a CR tool are enriched with both a set of heuristics for merging named entities labeled with different classes and also with some constraints that avoid the incorrect merging of similar mentions. Several tests show that this strategy improves both NER labeling and CR. The CR tool can be applied in combination with any system for named entity recognition using the CoNLL format, and brings benefits to text analytics tasks such as Information Extraction. Experiments were carried out in Spanish, using three different NER tools.

Keywords: named entity recognition, coreference resolution, information extraction

1. Introduction

Most Named Entity Recognition (NER) systems label each instance of a Named Entity (NE) using only local information (i.e., from the immediate context of the analyzed token). This strategy involves the misclassification of some mentions of the same entity in a single document.[1] For instance, a common NER system may classify most mentions of the entity "Lionel Messi" ("Messi", "Leo Messi", etc.) in the same document as a *person*, but some of them might be wrongly labeled as an *organization* or a *location*. These misclassifications also harm subsequent processes such as Coreference Resolution (CR) or Information Extraction (IE), which heavily depend on the accuracy of NER.

Although some methods make use of global information (from the whole document) for NER (Finkel et al., 2005), these systems might also classify mentions of the same entity with different labels. Besides, many NERs depend on other NLP modules such as tokenizers, lemmatizers or PoS-taggers, whose adaptation to global information NERs require a big effort.

On the other hand, most common strategies for nominal CR include several matching heuristics which cluster different mentions sharing —among other information— parts of their surface form: e.g., mentions such as "Lennon$_{\text{-PER}}$" and "John Lennon$_{\text{-PER}}$" might be merged into the same entity (Recasens and Hovy, 2009; Lee et al., 2013). However, as these systems depend on the previous NER tools, it is not likely that they merge mentions belonging to different classes, such as "John Winston Lennon$_{\text{-PER}}$" and "Lennon$_{\text{-ORG}}$" (wrongly classified).

In order to improve both NER and CR, this paper explores the incorporation of lexico-semantic heuristics into the CR sieves with the highest precision. On one hand, this allows the CR tool to merge nominal mentions (which actually belong to the same entity) which had been misclassified with different NER labels. On the other hand, the heuristics also select the best named entity class for each entity, thus correcting previous NER errors.

Several experiments performed in Spanish show that the proposed strategy systematically improves NER and CR in different scenarios, and also tasks such as IE in novels.

[1] Following Recasens and Martí (2010), a *mention* is each instance of reference to an object, while an *entity* is the collection of mentions referring to the same object in a document.

2. Related Work

Although the best systems of the CoNLL 2002 and 2003 shared tasks on language-independent NER did not make use of global information (they use different combinations of classifiers with local features, (Carreras et al., 2002; Florian et al., 2003)), some other works successfully incorporated non-local features for this task.

The use of global information at document-level is important for preserving label consistency. Thus, strategies such as Malouf (2002) or Curran and Clark (2003) take into account, when analyzing a token, the previous assigned tags of tokens with the same form.

Instead of checking other tags of each token, Chieu and Ng (2002) incorporate global feature groups into a single maximum entropy classifier. Finkel et al. (2005) also make use of non-local features, adding them to sequence models using Gibbs sampling. This enforces label consistency on long-distance dependencies.

Previously, Mikheev et al. (1999) implemented a multi-pass approach which applies a maximum entropy classifier after a set of high-precision NER rules, thus incorporating global information and producing NER at document-level. Similarly, Raghunathan et al. (2010) address CR by means of a multi-pass sieve. After identifying the candidates, the first passes apply high-precision rules for merging the mentions belonging to the same entities. After that, further sieves (with lower precision) are used to increase recall using the global features obtained in the previous steps.

Inspired in the last two works, this paper explores the incorporation of NER correction heuristics into the highest pre-

cise sieves of a multi-pass system for coreference resolution. This combination potentially allows any implemented NER tool to improve its performance, also increasing the accuracy of CR due to the better NER labeling.

3. NER at Document-level

In one document, different occurrences of mentions with the same surface form are likely to belong to the same semantic class and to the same discourse entity. This occurs in a similar way than the *one sense per discourse* hypothesis, which states that well-formed discourses tend to avoid multiple senses of a polysemous word (Gale et al., 1992). A common exception concerns some organizations which share the name of the city or region they belong to, such as sport teams.

However, it is worth noting that the single tokens which form complex mentions (i.e., "Lennon" in "John Lennon"; "Saramago" in "José Saramago") may occur in other mentions belonging to different entites ("Alfred Lennon$_{\text{-PER}}$", "Fundação José Saramago$_{\text{-ORG}}$"). Therefore, this should be taken into account when developing both NER and CR systems.

In this respect, some NER tools (namely those which do not use global information) often label different mentions sharing the surface form with different classes. These misclassifications usually involve wrong extractions in different text analytics tasks. As an illustration of this problen, Table 1 shows the 10 most frequent *person* and *location* mentions (and their frequency) in the Spanish book *Don Quijote de la Mancha*, using FreeLing (Padró and Stanilovsky, 2012) and Stanford CoreNLP (Finkel et al., 2005), respectively.[2] Mentions in italic are wrongly labeled, while bold indicates that these forms have also been labeled with other NE tag. Note that this is a complex and old book, so both systems produced several errors.

In a quick analysis of the results, both FreeLing and Stanford CoreNLP extracted reasonable well some of the main characters of the book. However, due to the document length ($\approx 380k$ tokens), some mentions belonging to *person* entities, with many occurrences in different contexts, were frequently labeled as *locations* and *organizations* (and even as *miscellaneous* entities, not shown in the table). This makes an automatic analysis of the main entities occurring in large documents more difficult

In some NLP pipelines, CR is the following task after NER, so adding correction heuristics in this step permits to improve both NER label consistency and CR without using additional tools.

4. Adding Correction Heuristics to Coreference Resolution Sieves

The CR strategy presented in Lee et al. (2013) is based on two main principles: (i) multi-pass: it applies a battery of sieves from high precision to high recall, linking the most confident pairs in the first passes (and learning features that are useful for the less precise ones). (ii) entity-centric: when analyzing a mention, it takes advantage of global features, learned from other mentions of the same entity.

[2]It was selected the most downloaded Spanish book at Project Gutenberg: `https://www.gutenberg.org/`

Class	Most frequent mentions
PER	don Quijote (1942), Sancho (1423), Dios (473), Sancho Panza (212), Dulcinea (116), don Fernando (110), Anselmo (104), Camila (104), Rocinante (103), Lotario (100)
LOC	***Rocinante*** (82), ***Sancho*** (73), España (43), *Altisidora* (39), ***Dulcinea*** (35), ***Sancho Panza*** (34), *Montesinos* (30), ***Camila*** (29), *Luscinda* (24), Zaragoza (20)
PER	Sancho (1338), Quijote (947), Sancho Panza (217), Rocinante (150), Fernando (126), Lotario (106), Cardenio (81), Don Quijote (78), Anselmo (65), Pedro (60)
LOC	***Camila*** (54), ***Quijote*** (38), **Quijote de la Mancha** (38), España (29), ***Sancho*** (26), *Montesinos* (22), *Luscinda* (14), Argel (14), ***Rocinante*** (12), Barcelona (12)

Table 1: Most frequent *person* mentions and locations (and their frequency) in the Spanish book using FreeLing (top rows) and Stanford CoreNLP (bottom).

Inspired by this strategy, this paper implements a partial CR tool with the two most precise passes, enriched with NER correction heuristics (the system is a simplified version of Garcia and Gamallo (2014b)). The input of the system is a tokenized text with NER labels in a IOB CoNLL-like format. This means that the system relies on the identification of the mention boundaries carried out by the NER tool.

In a first pass, the system extracts all the NE mentions (i.e., proper nouns) previously identified by the NER tool. At this step, each mention is a singleton, i.e., it belongs to a different entity. This pass also extracts, if possible, gender features of each mention by analyzing adjacent tokens.

Then, each pass traverses the text from the beginning, selects the mentions and, for each selected one, looks backwards for candidate mentions to link with. A mention is selected for analysis if (i) it is not the first mention of the text, and (ii) it is the first mention of the entity it belongs to. If a coreference link is performed, further passes which analyze other mentions of the same entity can use the features of the whole entity.

The first CR sieve, *String Match*, merges two mentions if they contain exactly the same text. *Relaxed String Match*, the second pass, links two mentions if the largest mention of the analyzed entity contains all the tokens of the shortest, in the same order (e.g., "Lennon" *vs* "John Lennon"). Note that comparing the largest mention of the entity (instead of just the analyzed and the candidate ones) allows the system to block links like "Lennon *vs* Alfred Lennon" if the "Lennon" was previously linked to mentions like "John Lennon" (since "John Lennon" and "Alfred Lennon" would not be compatible).

This second pass contains three constraints: (i) prepositional phrases: it blocks a CR link if the shortest mention appears inside a prepositional phrase on the largest one

(e.g., "Francia" *vs* "Tour de Francia"); (ii) trigger-words: this constraint uses NER-based trigger-words in order to forbid coreference links when one of the mentions contains a trigger-word of a different semantic class (e.g., "Fernando Pessoa" *vs* "Universidade Fernando Pessoa", where "universidade" belongs to *organization*); (iii) stop-words: prevents the merging if only one of the mentions contains words such as "Jr." (e.g., "Kennedy" *vs* "Kennedy Jr.").

In previous works, CR only links two mentions if they share the named entity label (Lee et al., 2013; Garcia and Gamallo, 2014b). In this paper, the correction heuristics merge them even if they belong to different classes, except for *location – organization* pairs which, by regular polysemy, often share the surface form even if they have different named entity types.

Thus, before performing the clustering of each mention pair, it is applied a label selection module which decides, when merging mentions with different NER labels, the most probable one. It works as follows: First, it verifies if any of the mentions is a singleton. If one of them is a compound singleton (with more than one token) and the other one has just one token, the NER label of the compound mention is selected ("Washington" *vs* "John Washington"). This is based on the claim that longer mentions are better analyzed by NER tools due to their larger number of lexical features.

If any of the two mentions is a singleton, the most frequent label of all the mentions of the same entity is selected. This heuristic relies on the accuracy of the NER tool. Then, if a singleton is compared to a mention which belongs to a larger entity, the label of the entity is preferred over the named entity class of the singleton. And finally, if both mentions are singletons, it is selected the label of the candidate entity (instead of the analyzed one). Note that the predicted NER labels are kept until the end of the process, so in each comparison they are used for computing the most probable one.

A simple gender checker is also applied when analyzing mentions that could be labeled as *person*: they are not linked if one of them belongs to a masculine entity and the other to a feminine one. Lists of the most common names in different languages are used for extracting the gender as well as for checking the link of *person* mentions.

5. Evaluation

This section contains several experiments aimed at knowing the impact of the correction heuristics in NER labeling, in CR and in IE.[3] For the two first evaluations, a multi-document corpus with coreferential annotation of *person* entities ($\approx 50k$ tokens) was used as gold-standard (Garcia and Gamallo, 2014). Three different NER tools were used: a knowledge-based one (K-b) (Garcia and Gamallo, 2015), FreeLing and Stanford NER. All of them were evaluated before and after the application of the CR tool. It also was evaluated the performance of CR in *person* entities, and the extraction of the main entities from the book *Don Quijote de la Mancha*, now obtained after the application of coreference resolution.

[3]Data and tools are available at `http://www.grupolys.org/~marcos/pub/lrec16.tar.bz2`

NER	Correct.	Prec	Recall	F1
K-based	No	73.98	73.98	73.98
	Yes	75.63	75.63	75.63
FreeLing	No	68.71	69.66	69.18
	Yes	69.45	70.41	69.93
Stanford	No	59.42	59.12	59.27
	Yes	61.30	60.93	61.11

Table 2: Impact of the correction in NER.

NER	Corr.	Metric	Prec	Rec	F1
Gold	—	MUC	96.6	93.6	95.1
		B^3	95.9	87.7	91.6
		BL	95.4	83.7	88.2
K-b	No	MUC	86.7	77.3	81.8
		B^3	80.8	72.9	76.6
		BL	88.5	66.9	70.1
	Yes	MUC	85.9	86.6	86.3
		B^3	81.4	80.5	80.9
		BL	88.6	67.7	71.0
FLing	No	MUC	71.7	84.8	77.7
		B^3	62.0	80.9	70.2
		BL	81.7	57.8	54.1
	Yes	MUC	72.1	87.5	79.1
		B^3	62.6	83.9	71.7
		BL	82.2	58.3	55.3

Table 3: Impact of the correction heuristics in CR.

Both K-b and FreeLing were applied out-of-the-box. The Stanford NER was trained for building an IOB2 classifier with AnCora corpus (Taulé et al., 2008) (P: 80% / R: 79% / F1: 79%).[4] Note that this is not a fair comparison between the different NERs. FreeLing and Stanford NERs applied their own NE identification module, while K-b uses the NE boundaries of the gold-standard. Also, different labeling criteria (between the gold-standard and the training corpus/resources) might involve variations in the results.

Table 2 contains the results of the three systems, before and after the application of the NER correction heuristics (values were obtained using the CoNLL NER scorer). The results show that the correction heuristics improve both precision and recall in the three systems. Best scores using these heuristics increase the F1 in more than 1.7%.

The evaluation corpus only contains coreference annotation of *person* entities, so the next results show the impact of the correction heuristics in this class. Also, the partial CR tool only deals with NE mentions, so the annotation of other nominal mentions (without proper nouns) and the pronominal ones was removed for the evaluation.

Table 3 contains the results of the CR evaluation with and without the correction heuristics (MUC, B^3 and BLANC metrics). Stanford NER has not been evaluated due to alignment inconsistencies generated by tokenization. First rows are the values of the CR tool using the gold NER la-

[4]This new IOB2 model for NER is available at `http://gramatica.usc.es/~marcos/resources/ner-es-ancora-iob2.distsim.ser.gz`

Class	Most frequent entities
PER	Don Quijote de la Mancha (2155), Sancho Panza (2109), Dios (491), Dulcinea del Toboso (282), Camila (144), Lotario (138), Anselmo (135), don Fernando (135), Dorotea (110), Cardenio (100)
LOC	*Quiteria* (39), Francia (25), Sierra Morena (22), *Sanchica* (21), Zaragoza (20), Argel (20), Berbería (17), Sevilla (16), *Calla* (16), Salamanca (15)
PER	Sancho Panza (2116), Don Quijote (1523), Rocinante (203), Camila (143), Lotario (140), Anselmo (135), Fernando (135), Dorotea (110), Cardenio (98), Luscinda (95)
LOC	*Quijote de la Mancha* (76), Argel (20), Sierra Morena (16), Salamanca (15), Barcelona (15), Madrid (14), Roma (13), Candaya (10), Roncesvalles (9), Aragón (9)

Table 4: Most frequent *person* entities and locations (and their frequency) in the Spanish book combining the CR tool with FreeLing (top) and with Stanford NER (bottom).

bels, showing that the two CR passes have more than 95% precision when analyzing NE mentions of *person* entities. These results indicate that the NER correction process also allows the CR tool to improve its performance between 0.9 and 4.5 F1, depending on the metric and on the NER tool used.

Finally, Table 4 contains the results of the extraction of the 10 most frequent *person* and *location* entities of the book analyzed in Section 3., after applying the CR tool. Both FreeLing and Stanford NER *person* results do not have noticeable errors, while the FreeLing locations output contains 3 errors (two —rare— *person* entities and a verb form wrongly labeled as *location*). The locations extracted by the Stanford NER only contain one error, produced by a wrong merging of a location (*Mancha*) with the character "Don Quijote de la *Mancha*". The —different— number of mentions of each entity denotes incorrect mergings produced by the CR tool, as the next section will show.

Although this last test is just illustrative, the results show that the correction heuristics are useful for improving not only NER and CR, but also the NER label consistency of large and complex documents such as novels.

6. Error Analysis

As shown in Section 4., the heuristics presented in this paper partially rely on the semantic labels predicted by the NER. Therefore, some of the errors produced by the system derive from some frequent NER mislabelings.

This is the case of the *organization* results produced by FreeLing and the CR tool, which show lower performance due to a common error produced by the NER ("Cristina" was —incorrectly— labeled as *organization* 11 times, while "Cristina Fernández" only appeared 6 times as a *person*, the correct label). This involved an error propagation that produced many incorrect relabelings in this entity. The other three classes (*person*, *location* and *miscellaneous*), as well as the overall evaluation, have better results after the application of the heuristics.

The CR tool also relies on the boundary identification predicted by the NER, so errors in this step may dramatically harm the performance of CR. Incorrect NE identification produces two different kinds of errors: (i) merging of wrong NEs: "Altisidora$_{\text{-PER}}$" linked to "Viva Altisidora$_{\text{-PER}}$" (where "Viva" is an interjection wrongly inserted into the NE). And (ii) error propagation of compatible mentions: "A Sancho Panza y Rocinante" was analyzed as a single NE by FreeLing, so in the CR step, it was merged with some mentions of "Sancho Panza$_{\text{-PER}}$" and with others of "Rocinante$_{\text{-PER}}$".

Concerning information extraction, the results on the Spanish book show that the proposed approach helps the NE labeling at document-level, since it reduces several NER errors. However, CR produced some incorrect mergings (mainly derived from NER mislabelings) that might involve errors in other tasks such as the combination of coreference resolution with relation extraction or open information extraction (Garcia and Gamallo, 2014a).

7. Conclusions and Further Work

This paper explored the addition of NER correction heuristics into the highest precise passes of a deterministic coreference resolution tool.

The results of several experiments showed that the proposed combination improves both the NER labeling of preexisting tools and the performance of nominal CR of *person* entities.

It was also performed an illustrative evaluation of this technique in a large Spanish book, showing that the document-level approach is useful for named entity extraction.

Current work is focused on the evaluation of this strategy in English and Portuguese, while further work will address the addition of different correction heuristics in other passes of the CR tool, as well as in a deeper error analysis aimed at avoiding error propagation.

8. Acknowledgements

This work has been partially suported by the Spanish Ministry of Economy and Competitiveness through the project FFI2014-51978-C2-1-R, and by a *Juan de la Cierva formación* grant, reference FJCI-2014-22853.

9. Bibliographical References

Carreras, X., Màrques, L., and Padró, L. (2002). Named entity extraction using adaboost. In *Proceedings of the 6th Conference on Computational Natural Language Learning (CoNLL 2002): Shared Task*, pages 167–170. Taipei, Taiwan.

Chieu, H. L. and Ng, H. T. (2002). Named entity recognition: A maximum entropy approach using global information. In *Proceedings of the 19th International Conference on Computational linguistics (COLING 2002), Volume 1*, pages 1–7. Association for Computational Linguistics.

Curran, J. R. and Clark, S. (2003). Language independent ner using a maximum entropy tagger. In Walter Daelemans et al., editors, *Proceedings of the 7th Conference on Computational Natural Language Learning (CoNLL 2003): Shared Task*, pages 164–167. Edmonton, Canada.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 2005)*, pages 363–370. Association for Computational Linguistics.

Florian, R., Ittycheriah, A., Jing, H., and Zhang, T. (2003). Named entity recognition through classifier combination. In Walter Daelemans et al., editors, *Proceedings of the 7th Conference on Natural Language Learning (CoNLL 2003)*, pages 168–171. Edmonton, Canada.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992). One sense per discourse. In *Proceedings of the Workshop on Speech and Natural Language*, pages 233–237. Association for Computational Linguistics.

Garcia, M. and Gamallo, P. (2014a). Entity-centric coreference resolution of person entities for open information extraction. *Procesamiento del Lenguaje Natural*, 53:25–32.

Garcia, M. and Gamallo, P. (2014b). An entity-centric coreference resolution system for person entities with rich linguistic information. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 741–752, Dublin.

Garcia, M. and Gamallo, P. (2015). Yet Another Suite of Multilingual NLP Tools. In José-Luis Sierra-Rodríguez, et al., editors, *Languages, Applications and Technologies*, volume 563 of *Communications in Computer and Information Science*, pages 65–75. Springer. Revised Selected Papers of the 4th International Symposium on Languages, Applications and Technologies (SLATE 2015), Madrid.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Malouf, R. (2002). Markov models for language-independent named entity recognition. In *Proceedings of the 6th Conference on Computational Natural Language Learning (CoNLL 2002): Shared Task*, pages 187–190. Taipei, Taiwan.

Mikheev, A., Moens, M., and Grover, C. (1999). Named entity recognition without gazetteers. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL 1999)*, pages 1–8. Association for Computational Linguistics.

Padró, L. and Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Turkey. European Language and Resources Association.

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., and Manning, C. (2010). A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 492–501. Association for Computational Linguistics.

Recasens, M. and Hovy, E. (2009). A deeper look into features for coreference resolution. In *Anaphora Processing and Applications*, pages 29–42. Springer-Verlag.

Recasens, M. and Martí, M. A. (2010). AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44.4:315–345.

10. Language Resource References

Garcia, M. and Gamallo, P. (2014). Multilingual corpora with coreferential annotation of person entities. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*, pages 3229–3233. European Language and Resources Association.

Taulé, M., Martí, M. A., and Recasens, M. (2008). Ancora: Multilevel annotated corpora for catalan and spanish. In *Proceedings of 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 96–101, Marrakesh. European Language and Resources Association.