Coreference Resolution in FreeLing 4.0

Montserrat Marimon*, Lluís Padró†, Jordi Turmo†
*Universitat Pompeu Fabra. Barcelona, Spain.
montserrat.marimon@upf.edu

†TALP Research Center.
Universitat Politècnica de Catalunya. Barcelona, Spain.
{padro,turmo}@cs.upc.edu

Abstract

This paper presents the integration of RelaxCor into FreeLing. RelaxCor is a coreference resolution system based on constraint satisfaction that ranked second in the CoNLL-2011 shared task. FreeLing is an open-source library for NLP with more than fifteen years of existence and a widespread user community. We present the difficulties found in porting RelaxCor from a shared task scenario to a production enviroment, as well as the solutions devised. We present two strategies for this integration and a rough evaluation of the obtained results.

Keywords: FreeLing, Coreference Resolution, CoNLL-2011, relaxation labeling

1. Introduction

FreeLing[1] is an open-source multilingual language processing library providing a wide range of analysis functionalities for several languages.

The project is conceived as a library that can be called from a user application in need of analysis services. The software is open-source, distributed under an GNU Affero General Public License[2], and dual-licensed to companies that embed it in their commercial products or online services.

The open-source approach has been very fruitful during the fifteen years of life of the project (the first version was released on 2003). The amount of accumulated downloads during this time is over 200,000. Contributions from the user community combined with the increasing availability of open source language resources has made it possible to extend the number of supported languages from three (English, Spanish and Catalan) to fourteen (adding German, French, Italian, Portuguese, Russian, Norwegian, Asturian, Welsh, Galician, Croatian and Slovene).

FreeLing offers a wide variety of language processing modules, though not all modules are available for all languages. Most relevant modules are: Language identification, tokenization and sentence splitting, lemmatization, date/time detection, numbers detection, multiword expressions detection, physical magnitudes detection, named entity recognition and classification, PoS tagging, SAMPA phonetic encoding, word sense disambiguation, shallow parsing, constituency parsing, dependency parsing, semantic role labelling, coreference resolution, semantic graph extraction, and document summarization.

More details about the FreeLing project can be found in (Padró and Stanilovsky, 2012) and in the online documentation in the project website.

One remarkable extension in version 4.0 was the inclusion of a Coreference Resolution module, based on RelaxCor, the second-ranked system in CoNLL-2011 shared task (Sapena et al., 2011).

However, academic shared tasks have a very specific scenario, which does not necessarily match the real-world settings in which a system like FreeLing is required to operate. Thus, considerable efforts must be devoted to the integration of a module successful in the laboratory, such as RelaxCor, in a production pipeline.

In this paper we describe how this coreference resolution module was integrated in FreeLing (for English and Spanish), as well as the encountered obstacles and solutions devised. We also present an alternative configuration for RelaxCor using hand-written constraints instead of the automatically learnt constraints used in CoNLL shared task.

The next section briefly summarizes related work. Section 3. overviews the basic idea behind RelaxCor and the main difficulties presented by its integration in FreeLing. Sections 4. and 5. describe an alternative set of hand-written constraints and compare its performance with the machine-learned model. Finally, Section 6. concludes.

2. Related Work

There are several open-source suites other than FreeLing that offer state-of-the-art level NLP functionalities. The most remarkable, for being open-source, widely used and offering a set of functionalities comparable to FreeLing are: Stanford CoreNLP (Manning et al., 2014), Apache OpenNLP[3], NLTK (Bird et al., 2009) and IXA Pipes (Agerri et al., 2014).

There are also other systems of NLP-related software, such as GATE or UIMA, which are not language analysis pipelines themselves, but architectures or frameworks to integrate existing components.

The above mentioned suites largely differ with respect to the used programming language, offered APIs, processing speed, supported languages, customization or retraining capabilities, whether they are more developer-oriented or end-

[1] http://nlp.lsi.upc.edu/freeling
[2] http://www.gnu.org/copyleft/agpl.html
[3] https://opennlp.apache.org

user oriented, etc. Thus, a detailed comparison is out of the scope of this paper.

Regarding coreference resolution, Stanford CoreNLP includes an updated version of the first-ranked system in CoNLL 2011 shared tasks, supporting English and Chinese. Apache OpenNLP offers a basic support for English. IXA Pipes do not ship a coreference resolution module out-of-the-box, but third-party provided modules are available for Spanish and English. Finally, the latest FreeLing version offers coreferences for Spanish and English.

The first attempt to establish a common evaluation framework for coreference systems was carried out in SemEval 2010 (Recasens et al., 2009), which offered data sets for 6 languages (including English and Spanish). Later, CoNLL-2011 and CoNLL-2012 shared tasks (Pradhan et al., 2011; Pradhan et al., 2012) proposed similar tasks that have been a reference framework since then. The 2011 edition included only English, and was won by Stanford rule-based system (Lee et al., 2013). The 2012 edition included English, Arabic, and Chinese, and was won by a neural network based system (Fernandes et al., 2012), which has been the main trend in the state of the art since then.

3. Integration of RelaxCor into FreeLing

3.1. RelaxCor

RelaxCor is a coreference resolution system based on constraint satisfaction. The coreference resolution problem is represented as a graph with mentions in the vertices which are connected to each other by edges. Edges are assigned a weight that indicates the confidence whether the mention pair corefers or not. More specifically, an edge weight is the sum of the weights of the constraints that apply to that mention pair.

The knowledge used by the system is encoded in constraints, each of which has a confidence score. The larger the score absolute value, the more reliable the constraint is and the stronger effect when applied. The sign of the constraint confidence score indicates whether a pair or a group of mentions may corefer (positive) or not (negative). Only constraints over pairs of mentions are used in the current version of RelaxCor, though the model can handle higher-order constraints. Constraints and their confidence scores can be obtained from any source, including manual encoding or automatic acquisition from a training corpus.

Figure 3.1. shows (a simplified version of) the graph corresponding to the text:

> *FC Barcelona president Joan Laporta has warned Chelsea off star striker Leonel Messi. Aware of Chelsea owner Roman Abramovich's interest in the young Argentine, Laporta said last night: "I will answer as always, Messi is not for sale and we do not want to let him go."*

Constraints are applied to every pair of mentions in the text, and a compatibility score for that pair is computed (represented by edges in the graph). Many pairs remain unconnected if constraints find no evidence neither for nor against joining them. The algorithm will partition the graph, keeping together pairs with high compatibility and setting apart nodes with negative scores.

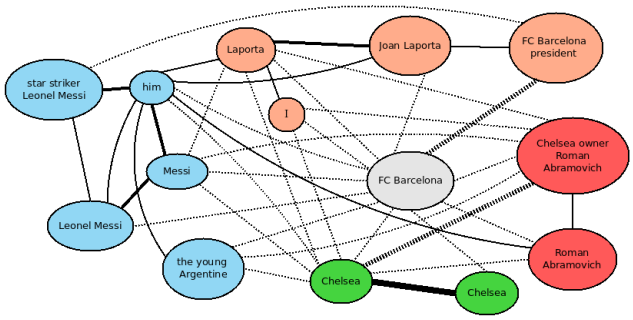

Figure 1: Graf produced from example text. Dotted lines represent negative compatibility scores. Solid lines represent positive scores. Line thickness is proportional to score absolute value. Colors represent the final groups created.

For instance, constraints assign large negative value to nodes corresponding to named entities of different classes (person and organization entities in this example), and large positive values to nodes of the same kind with similar names (e.g. Both mentions of *Chelsea*, or *Messi* and *Leonel Messi*). There are also nodes such as *him* that are compatible with many mentions, and the final decision depends both on the strength of their connections, and the conections among their neighbors (e.g. *him* not only has stronger connections with *Messi* and *star striker Leonel Messi*, but they both also share a strong connection with *Leonel Messi*, which reinforces all of them gluing together).

RelaxCor uses relaxation labeling for the resolution process. Relaxation labeling is an iterative algorithm that performs function optimization based on local information. It has been widely used to solve a variety of NLP problems. A vector of probability values is maintained for each vertex/mention. Each vector element corresponds to the probability that the mention belongs to a specific entity among the potential entities in the document. During the resolution process, the probability values are updated according to the edge weights and probability vectors of the neighboring vertices. The larger the edge weight, the stronger the influence exerted by the neighboring distributions. The process stops when there are no more changes in the probability vectors.

The relaxation labeling approach combines mention pair classification and linking in one step. Thus, decisions are taken considering the entire set of mentions, which ensures consistency and avoids local classification decisions.

Mentions are modeled as feature sets, and contain a variety of information (gender, number, person, PoS, sense, etc). Constraints use these features to establish a compatibility value among pairs of mentions (e.g. Two mentions with different gender will have a negative compatibility. Mentions that are pronouns and have the same person, gender, and number will have a positive compatibility, etc.). Note that semantic and syntactic information can also be used in the constraints.

RelaxCor was first proposed in SemEval-2010 (Sapena et al., 2010), and an improved version (Sapena et al., 2011) ranked second in CoNLL-2011. Extended details on how

RelaxCor works can be found in (Sapena et al., 2013).
We selected RelaxCor as the coreference resolution module to be integrated in FreeLing for the following reasons

- It was developed in UPC, so FreeLing team has a deep understanding of the algorithm and the code, which eases the portability. Also, the underlying algorithm is Relaxation Labelling which was already part of FreeLing, since it is used by one of the PoS taggers.
- RelaxCor uses a very general approach: The problem is modeled in terms of graph vertices encoded as feature sets and constraints among them, which makes it easy to adapt the code for new languages or even for new tasks.
- The constraint-based approach allows the addition of new languages with a relatively low cost: If no training data is available to use machine learning techniques, constraint can be hand-encoded as presented in this paper.

3.2. Integration Issues

Coreference Resolution is a complex NLP tasks, since it requires a lot of information from previous analysis steps: Tagging and parsing are required to identify candidate mentions and to discover syntactic relations among them that are relevant to the task (e.g. appositions, relative clauses, etc.). Named Entity detection and classification is also crucial to establish whether two names may refer to the same entity. Word sense disambiguation and Semantic Role Labelling also provide relevant information in some cases.
This large amount of dependences largely increases the difficulty of integrating a module developed in one specific scenario into a production pipeline, each with its own settings and dependences.
Some relevant issues that we had to take into account are:

- The input data in CoNLL-2011 shared task follows tokenization conventions that not necessarily match those used by FreeLing.
- The input data in CoNLL-2011 shared task uses a PoS tagset that has some differences with that used by FreeLing.
- The input data in CoNLL-2011 shared task contains a gold constituency parse which can be used to detect mentions. The module integrated in FreeLing will have to resort to the output of its own parser, which not only will contain errors, but also uses different labels and syntactic structures.
- The constituency parser in FreeLing is rule based and uses a simple strategy. Thus, it does not perform at the same accuracy level as a statistical parser.
- The input data in CoNLL-2011 shared task contains speaker information in some dialog documents, but FreeLing has no dialog or speaker detection module.

To tackle with these issues, we took the following integration decisions:

- To build two mention detectors:
  - Another one based on dependency trees, that would use the output of FreeLing statistical dependency parser, which offers more robust and accurate results. This detector follows FreeLing linguistic team criteria to establish what a candidate mention is, which are not necessarily the same followed in CoNLL shared tasks.[4]
- To adapt the mention feature extraction, as well as the syntactic checks required by some constraints, to FreeLing PoS tagset and syntactic labels and structures. Note that this had to be done twice: for constituency trees and for dependency trees.
- To taylor the train/development/test corpus to match FreeLing tokenization criteria. We ported the coreference annotations to the corpus retokenized according to FreeLing criteria, and we excluded from the corpus those documents where unsolvable retokenization clashes prevented the safe mapping of the gold annotations.
- To exclude dialog documents with speaker information from the corpus.

4. Coreference Resolution in FreeLing

As mentioned in section 3.2. we integrated in FreeLing two versions of RelaxCor: One using constituency parsing to detect mentions and to extract syntactic information and the other using dependency parsing.

4.1. Constituency parsing version

The constituency parsing based module consists of a straightforward translation of the original RelaxCor from Perl to C++, adapting PoS tagset, constituent labels and syntactic structures. The goal was to have a FreeLing module that could use machine learning models learned on the CoNLL shared task, to avoid costly re-training and parameter-tuning procedures. Given differences between the criteria used in the training corpus and the output of FreeLing preprocessing stages, we expect this module to perform worse once integrated in FreeLing than it did in the shared task, so we will consider it as a baseline.

4.2. Dependency parsing version

The dependency parsing based module uses more accurate parsing trees, raising the quality of mention detector and syntactic information used in constraints. Since constituent information is not available here, originally trained models can not be used. One option would be retraining the module using a corpus adapted to the new criteria, which would require a costly re-annotation effort. Thus, we opted by developing a set of hand written constraints, inspired on the approach proposed in (Lee et al., 2013), the winner system in CoNLL-2011, that defines ten sieves of decreasing precision rules, applied in a cascaded schema.

[4]Our mention detector considers a mention the span of any subtree headed by a noun, a personal pronoun, or a relative pronoun, except those where the head is the word *what* or a temporal noun (*day*, *year*, *morning*, *minute*, etc.). No filtering of embedded mentions or pleonastic pronouns is performed.

In our case, relaxation labeling applies all constraints simultaneously, thus we need to emulate the cascading using compatibility scores of different ranges (e.g. more releable rules have scores around 50, while less precise constraints have scores about 5–10), so that higher-precision rules overweight any contradicting lower-precision rule. In addition, our model does not follow an entity-mention approach but a mention-pair approach, thus some sieves can be only approximated. Finally, we need to add some default rules that favor the creation of singletons in absence of strong enough coreference evidence, to avoid ending in trivial solutions where all mentions belong to the same group.

Next, we overview the sieves proposed by Stanford and present a few samples of the rules we encoded in each of them:

- **Sieve 1**: Mentions with the same speaker in dialog documents or when direct speech is used. We only deal with the later, since no speaker identification is available in FreeLing.
  Sample rules:
  +50 if both mentions are personal pronoun "I" and both are inside the object of the same reporting verb (*say*, *tell*, *ask*, etc.) or inside the same quotation.
  -25 if one mention is a personal pronoun and the other is not the same pronoun and both are inside the object of the same reporting verb, or inside the same quotation.

- **Sieves 2–3**: Mentions containing the same text (either the whole mention or up to the head word).
  Sample rules:
  +50 if both mentions are of the same type (named entity, pronoun, noun phrase), not nested in one another and their texts match completely.
  +25 if both mentions are of the same type and their texts match up to the head word.

- **Sieve 4**: Mentions appearing in specific constructions (e.g. relative clauses, appositions and predicative constructions).
  Sample rules:
  +50 if both mentions are in apposition.
  +50 if both mentions are in a predicative structure.
  +50 if one mention is a relative pronoun and has the other as syntactic antecedent.

- **Sieves 5–7**: Mentions that have the same head, or the head of one matches some word in the other.
  Sample rules:
  +25 if one mention is a named entity or noun phrase, the other is not a pronoun, they have the same head and their modifiers are compatible.

- **Sieves 8–9**: Mentions headed by the same proper name.
  Sample rules:
  +50 if both mentions are named entities and they have the same head and the same semantic class.

- **Sieve 10**: Pronominal coreference.
  Sample rules:
  +15 if one mention is a 3rd-person, non-relative pronoun, the other is a named entity or noun phrase, they have morphological agreement and they belong to compatible semantic classes (person, man, woman, non-person, organization, location).

Since the used features and rules are very general, we also evaluated their performance on Spanish, using the annotated corpus provided by SemEval-2010 Task 1 (Recasens et al., 2009). After adapting the lexicon and syntactic criteria used by the features and rules, we obtained a coreference resolution system with a performance comparable to our version for English.

5. Experiments and Results

In this section we present the experimental setting we used to roughly evaluate the results of the integration.

We evaluated two systems: (1) the original RelaxCor ported to C++, using the output of FreeLing constituency parser and the original model trained on CoNLL-2011 data, and (2) the alternative version, using the output of FreeLing dependency parser and hand-written rules. Both systems were also applied to Spanish (after adapting configuration files with relevant lexica, PoS tags, syntactic labels, etc).

The used data was a subset of CoNLL-2012, excluding documents containing dialogs or severe tokenization mismatches (accounting for about 16-17% of the documents). The original CoNLL-2012 test and development sections were used as development corpus for the hand-written rules, and the train section of CoNLL-2012 was used as test. See Table 1 for a summary of corpus sizes.

		#doc orig.	#doc	#sent	#tok
English	Devel.	639	493	6,090	101,957
	Test	2,273	1,952	18,637	348,831
Spanish	Devel.	308	261	1,079	29,285
	Test	875	719	2,615	77,749

Table 1: Sizes of the used development and test corpus. Column *#doc orig* shows the number of documents in the original CoNLL corpus. The other columns show the sizes of the corpus after filtering dialog documents and tokenizaton mismatches.

Table 2 shows the results for both developed versions computed using the latest version (v8.01) of official CoNLL-2012 scorer. Rows marked *MD* present mention detection scores. Other rows present different performance metrics, including CoNLL-2011 and 2012 official metrics (average of MUC, B^3, and CEAF-e).

Several issues must be taken into account when interpreting the results:

- The development and test corpus partitions are not the same used in CoNLL shared tasks. Moreover, dialog documents and documents where tokenization could not be mapped where excluded.

		Constituency parsing, ML constraints						Dependency parsing, hand-written constraints					
		English			Spanish			English			Spanish		
Corpus	Metric	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1
Devel.	*MD*	*65.80*	*62.87*	*64.30*	*28.55*	*55.90*	*37.80*	*66.66*	*64.97*	*65.80*	*55.85*	*62.05*	*58.79*
	MUC	54.76	49.37	51.93	19.80	41.91	26.89	52.48	54.30	53.38	36.98	43.89	40.14
	B^3	41.12	28.06	33.36	15.60	42.53	22.83	36.88	45.58	40.77	34.81	44.72	39.15
	CEAF-m	39.18	37.45	38.29	24.12	47.22	31.93	46.60	45.42	46.00	45.07	50.07	47.44
	CEAF-e	28.54	33.89	30.98	22.80	38.58	28.66	43.42	35.62	39.13	46.73	45.81	46.27
	BLANC	42.69	28.20	32.01	10.34	31.65	15.59	40.44	48.36	42.63	27.40	36.30	31.06
	CoNLL	**41.47**	**37.10**	**38.75**	**19.40**	**41.00**	**26.12**	**44.26**	**45.16**	**44.42**	**39.50**	**44.80**	**41.85**
Test	*MD*	*57.57*	*54.25*	*55.86*	*28.07*	*52.73*	*36.64*	*59.81*	*57.35*	*58.56*	*59.08*	*60.64*	*59.85*
	MUC	43.46	38.48	40.82	19.46	39.89	26.16	45.67	46.64	46.15	38.99	42.57	40.70
	B^3	35.04	27.92	31.08	15.40	39.76	22.20	36.16	42.31	39.00	35.76	42.68	38.91
	CEAF-m	38.60	36.40	37.47	24.02	45.14	31.36	46.22	44.32	45.25	46.25	47.47	46.86
	CEAF-e	29.23	32.95	30.98	22.46	35.65	27.56	42.97	35.74	39.02	48.03	43.55	45.68
	BLANC	35.48	23.26	27.50	10.41	31.04	15.59	33.74	39.37	35.30	29.21	34.01	31.21
	CoNLL	**35.91**	**33.11**	**34.29**	**19.10**	**38.43**	**25.30**	**41.60**	**41.56**	**41.39**	**40.92**	**42.93**	**41.76**

Table 2: Results of both RelaxCor integration strategies for English and Spanish.

- The machine-learning version was trained on CoNLL-2011 data, and is being evaluated on CoNLL-2012.
- Criteria used in the hand-written rule version to define which mentions are considered singletons are different from those used in CoNLL data.
- Our hand-written rules mark some coreferences (e.g. relative pronouns or predicative noun phrases) that are not marked in CoNLL data.

For all this, results presented here can not be compared to the state-of-the-art in CoNLL shared tasks, but only taken as a self-contained evaluation.

From this point of view, results show that both English models get a similar score on mention detection. Regarding the final coreference score, the distance between the training and test scenarios severely hampers the accuracy of the machine-learned models, while the hand-written model, being developed and tested in the target corpus, obtains significantly higher scores.

When applying the models to Spanish, the machine-learned model suffers a big drop in both measures, as one would expect. But the manual model –more easily tunable–, obtains results for Spanish in the same range than English.

6. Conclusions and Further Work

We have presented two strategies to integrate RelaxCor –a coreference resolution system developed for CoNLL-2011 shared task– into FreeLing. We have discussed encountered problems and solutions undertaken. We used pre-trained machine-learned models, as well as hand-written constraints, and evaluated the results. Even the evaluation is not comparable to the state of the art given the differences in the used corpus and criteria, we believe that the provided modules will be useful to FreeLing users.

There are still open lines to pursue: The machine learning version could be retrained and tested on an adapted version of CoNLL-2012. Constraint compatibility scores for hand-written rules could be automatically assigned using a training corpus. The utility of the modules could be assessed via indirect evaluation. Finally, the hand-written model could be adapted to other languages in FreeLing.

7. Acknowledgements

This research was partially funded by the Spanish Government via project Graph-MED (TIN2016-77820-C3-3-R).

8. Bibliographical References

Agerri, R., Bermudez, J., and Rigau, G. (2014). IXA pipeline: Efficient and ready to use multilingual NLP tools. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, May.

Bird, S., Loper, E., and Ewan, K. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Fernandes, E., dos Santos, C., and Milidiú, R. (2012). Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 41–48. Association for Computational Linguistics.

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*, 4(39):885–916.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Padró, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In

Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, pages 1–27. Association for Computational Linguistics.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, CoNLL '12, pages 1–40, Stroudsburg, PA, USA. Association for Computational Linguistics.

Recasens, M., Martí, T., Taulé, M., Màrquez, L., and Sapena, E. (2009). Semeval-2010 task 1: Coreference resolution in multiple languages. *SEW-2009 Semantic Evaluations: Recent Achievements and Future Directions*.

Sapena, E., Padró, L., and Turmo, J. (2010). Relaxcor: A global relaxation labeling approach to coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 88–91. Association for Computational Linguistics.

Sapena, E., Padró, L., and Turmo, J. (2011). Relaxcor participation in conll shared task on coreference resolution. pages 35–39. Association for Computational Linguistics.

Sapena, E., Padró, L., and Turmo, J. (2013). A constraint-based hypergraph partitioning approach to coreference resolution. *Computational Linguistics*, 39(4):847–884, December.