# Summ-it++: an enriched version of the Summ-it corpus

**A. Antonitsch, A. Figueira, D. Amaral, E. Fonseca, R. Vieira, S. Collovini**
PUCRS University – Porto Alegre – Brazil
andre.antonitsch, anny.figueira, daniela.amaral,
evandro.fonseca, sandra.abreu (@acad.pucrs.br), renata.vieira@pucrs.br

**Abstract**

This paper presents Summ-it++, an enriched version the Summ-it corpus. In this new version, the corpus has received new semantic layers, named entity categories and relations between named entities, adding to the previous coreference annotation. In addition, we change the original Summ-it format to SemEval.

**Keywords:** Coreference, Named entities, Semantic relations

## 1. Introduction

Coreference resolution is an important challenge for language processing. Currently for Portuguese there are three main corpora with some kind of coreference annotation: HAREM (Freitas et al., 2010), Garcia's corpus (Garcia and Gamallo, 2014) and Summ-it (Collovini et al., 2007).

HAREM contains annotation of named entities and their identity relations, its main purpose was the evaluation of Named Entity Recognition (NER) systems. The corpus contains manually annotated named entities distributed in ten semantic categories. Relations between these named entities have also been annotated manually, in four types: identity, inclusion, placement and other.

Garcia's corpus contains coreference annotation for person entities.

Summ-it contains noun phrase coreference annotation, being thus the corpus with the most complete coreference chains. It was semi-automatically annotated with morphosyntactic information, and manually annotated with coreference. Besides coreference the texts were also manually annotated with rhetorical relations. Also, for each text, there are manual and automatically generated summaries.

In this paper we describe Summ-it++, an enriched Version of Summ-it. The proposed new version adds two new annotation layers: named entities and relations between named entities. In addition, the format was changed to the SemEval (Recasens et al., 2010), a well-known and widely used format.

Therefore we provide a corpus that integrates different annotation layers, in a format that can be evaluated according to usual evaluation metrics for coreference, making use of available tools that compute such metrics. So, with this resource we aim to contribute to further Portuguese NLP research.

The paper is organized as follows: Section 2 describes the new Summ-it++ corpus, as well as its annotation scheme; Section 3 presents the process of the corpus generation; Section 4 describes the conversion from corpus to SemEval form; and finally, Section 5 presents our conclusions and future work.

## 2. Summ-it++

Summ-it++ is an evolution of the corpus Summ-it (Collovini et al., 2007). The original Summ-it consists of fifty journalistic texts from the Science section of the Folha de São Paulo newspaper. The texts are annotated in many layers. Here we will consider mainly the morphosyntactic and coreference annotation, which we want to integrate with two new semantic layers. The corpus has a total of 560 coreference chains with an average of 3 members (noun phrases for each chain). The largest chain has 16 members (noun phrases). Summ-it has been used in previous coreference resolution research for Portuguese ((de Souza et al., 2008), (Coreixas, 2010), (Fonseca et al., 2014), (Da Silva et al., 2010)) and has had an important role in the training and the validation of classification models. Basically, in this new version, we added two semantic layers (named entities and their relations). In addition, the original format was changed to SemEval (Recasens et al., 2010). To provide the two new semantic layers, two resources based on CRF algorithm were used: the CRF classifier, proposed in (Collovini et al., 2014), for relation extraction and NERP-CRF (**?**) for named entities. These new layers were automatically annotated and manually revised by humans. For the morphosyntatic annotation we used CoGrOO (Silva, 2013). In the following subsections we describe each semantic layer provided by the new corpus version.

### 2.1. Morphosyntactic Annotation

For Summ-it++ morphosyntactic annotation we used CoGroo (Silva, 2013). CoGroo is a open-source grammar checker widely used for Portuguese. It is capable of identifying Portuguese mistakes such as pronoun placement, noun agreement, subject-verb agreement, usage of the accent stress marker, subject-verb agreement, and other common errors of Portuguese writing. Besides, CoGrOO has pos-tagging, chunking and morphosyntactic annotation.

### 2.2. Named Entities

NER is the identification and classification of expressions mostly composed of proper names, which refers to a specific entity in the text. NERP-CRF (Amaral and Vieira, 2014) is the system responsible for the extraction of such entities in Summ-it ++. In this work, we classified the NEs according to the HAREM Conference's guidelines, which have the following classes: Abstraction, Event, Organization, Other, Person, Place, Thing, Time, Value, and Work (Freitas et al., 2010). In sentence (a), for example, we have the following NE classes: Person, such as *"Miguel Guerra"*, Organization, such as *"University of Santa Catarina"*, and Place such as *"Santa Catarina"*.

**(a)** *"A opinião é do agrônomo Miguel Guerra, da UFSC (Universidade de Santa Catarina)."* (*The opinion is from the agronomist Miguel Guerra, of UFSC (University of Santa Catarina)*).

## 2.3. Semantic Relations between entities

Relation Extraction (RE) is the task of identifying and classifying semantic relations that occur between entities in a given text (Jurafsky and Martin, 2009). Relation extraction can be useful in many NLP tasks, in particular, for Coreference Resolution, the focus of which is to determine antecedent chains. The identification of these chains in a text can improve the process of relation extraction (Gabbard et al., 2011).

In the proposed corpus, we include relations of any type (as in open RE) occurring between named entities of the following categories: Organization, Person and Place. For that, we use the CRF classifier, proposed in (Collovini et al., 2014). The annotations were provided automatically, but manually revised. We define a relation descriptor as the text chunks that describe the explicit relation occurring between a pair of named entities in the sentence. For example: in sentence (a), we have the relation descriptor *"de"* (*of*) that occurs between the named entities *"Miguel Guerra"* and *"UFSC"*, in this sentence.

## 2.4. Coreference

Coreference basically consists of finding different references to a same entity in a text. In (a) the noun phrases *"o agrônomo " "the agronomist "* and *"Miguel Guerra" " Guerra"* are considered coreferent, in other words, they belong to the same coreference chain. The proposed corpus presents the manual annotation of coreference previously provided by Summ-it now in the SemEval format, which is more adequate for evaluation purposes, since available tools such as the CoNLL scorer[1] might be used. This tool generates widely used coreference metrics, as described in (Pradhan et al., 2011).

## 2.5. Annotation Scheme

The Summ-it++ new annotation format is SemEval: a single file, containing all Summ-it texts. Each text document is separated by "#begin document ID" and "#end document ID". The information of each sentence is organized vertically with one token per line, and a blank line after the last token of each sentence. The information associated with each token is available in columns (separated by "\t"). Besides the format, the novelty is the integration of coreference with named entities and their relations (seen in the last three columns in Table 2).

The annotation columns are:

**ID:** Token ID in sentence order;

**Token:** the word or multiword;

**Lemma:** the word lemma;

**POS:** Part-of-speech tagging of each word;

**Feat:** features (gender and number) of each word;

**Head:** denotes if the word is a head word (if yes, this field receives '0');

**NE:** represents the semantic category, as below:

| Semantic Class | Equivalence |
|---|---|
| Abstraction | ABS |
| Event | EVE |
| Organization | ORG |
| Other | OTH |
| Person | PER |
| Place | PLC |
| Thing | THI |
| Time | TIM |
| Value | VAL |
| Work | WOR |

Table 1: Semantic class equivalence scheme.

**Rel:** represents the relation descriptor which expresses a relation between a pair of named entities. When this relation exists, both named entities involved receive the token ID from the words that compose the relation descriptor. If the relation contains two or more descriptors, like in : [*"Cassius Vinicius Stevani"*] , [*"químico de"*] [*chemist of*] , [ *"USP"*], it's separated by a pipe.

**Coref:** each noun phrase starts using "( " followed by the chain ID. Note that the ") " just occurs in the last NP token. Basically: coreferent NPs receives the same chain ID.

The resulting new corpus has thus the integrated annotation of coreference, named entities and relations between named entities, making it an important resource for research in Portuguese NLP.

[1] http://conll.cemantix.org/2012/software.html

| ID | Token | Lemma | PoS | Feat | Head | NE | Rel | Coref |
|---|---|---|---|---|---|---|---|---|
| 1 | A | o | art | F=S | _ | _ | _ | _ |
| 2 | opinião | opinião | n | F=S | 0 | _ | _ | _ |
| 3 | é | ser | v-fin | PR=3S=IND | _ | _ | _ | _ |
| 4 | de | de | prp | _ | _ | _ | _ | _ |
| 5 | o | o | art | M=S | _ | _ | _ | (2 |
| 6 | agrônomo | agrônomo | n | M=S | 0 | _ | _ | _ |
| 7 | Miguel_Guerra | _ | prop | M=S | 0 | PES | (9) | _ |
| 8 |  |  | _ | _ | _ | _ | _ | _ |
| 9 | de | de | prp | _ | _ | _ | _ | _ |
| 10 | a | o | art | F=S | _ | _ | _ | _ |
| 11 | UFSC | _ | prop | F=S | 0 | ORG | (9) | (3) |
| 12 | ( | ( | ( | _ | _ | _ | _ | _ |
| 13 | Universidade_de_Santa_Catarina | _ | prop | F=S | 0 | ORG | _ | (3) ‖2) |
| 14 | ) | ) | ) | _ | _ | _ | _ | _ |
| 15 | . | . | . | _ | _ | _ | _ | _ |
|  |  |  |  |  |  |  |  |  |
| 1 | Guerra | _ | prop | M=S | 0 | PES | _ | (2) |
| 2 | participou | participar | v-fin | PS=3S=IND | _ | _ | _ | _ |
| ... |  |  |  |  |  |  |  |  |

Table 2: Annotation scheme

# 3. Corpus Generation

The generation of the new corpus had the goal of the addition of two new semantic layers of annotations to the original Summ-it, named entities and semantic relations. And converting this expanded corpus to the more common SemEval format.

## 3.1. Morphosyntactic Annotation

The morphosyntactic annotation of the corpus was obtained through the CoGrOO (Silva, 2013) PoS-tagger and morphosyntactic annotator. Each text was split by the CoGrOO parser into tokens. It is worth noting, however, that CoGrOO concatenates composite proper nouns into a single token. As well as splitting preposition-article abbreviations that are common to Portuguese ('da', 'do' changes to 'de + a', 'de + o'). For each token produced, CoGrOO also supplies, lemma, Part-of-Speech tag, gender and number features. The CoGrOO chunker and shallow-parser was then used to generate the noun-phrases and subsequently annotate which tokens are head of a noun-phrase.

## 3.2. Named Entities

In this work the CRF-based classifier NERP-CRF (Amaral and Vieira, 2014) was applied to the Summ-it texts in order to extract and classify the named entities (NEs).

As a pre-processing phase, the POS tagging was provided through the use of the OpenNLP parser. With the texts properly tagged, the system is then able to extract and classify the NEs.

For the training of the CRF model, the Second HAREM's Golden Collection was utilized. Using this model, NERP-CRF was applied to the Summ-it texts and the output with the identified and classified ENs was generated. After this process, the output was manually revised and corrected by two annotators to be used as a reference to evaluate the system.

As a result, there were 1,086 NEs identified, distributed in the ten HAREM categories.

Precision (P), Recall (R) and F-Measure (F) obtained by the system are given for Person, Place and Organization classes, since the Relation Extraction task (Section 3.3) considers only relations between these three classes (see Tables 3 and 4). The corpus incorporated the manually revised entities.

| Classes | R | P | F |
|---|---|---|---|
| Person | 68.47% | 79.43% | 73.54% |
| Place | 86.98% | 100.00% | 93.03% |
| Organization | 72.63% | 52.47% | 60.92% |

Table 3: NERP-CRF NE Identification

| Classes | R | P | F |
|---|---|---|---|
| Person | 59.11% | 68.57% | 63.49% |
| Place | 56.25% | 69.23% | 62.06% |
| Organization | 71.05% | 51.33% | 59.60% |

Table 4: NERP-CRF NE Classification

## 3.3. Semantic Relation

For the extraction of semantic relations we applied a CRF classifier (Collovini et al., 2014) to the Summ-it texts. It identifies relation descriptors that express a explicit relation between pairs of named entities.

For the identification of the NE categories Person, Organization and Place, we use the NERP-CRF output

described in Section 3.2..
After, we identify the first pair of NEs in each sentence. Therefore we consider only one pair per sentence.
As result, the pair of NEs identified in the sentence is considered candidate for arguments of the relation instances as a triple (NE1, relation descriptor, NE2). For example, in the sentence (a) we have this triple: (Miguel Guerra, de, UFSC).
A sum of 101 relation candidates was extracted and given as input to the classifier. The classifier indicated the valid descriptors.
We evaluated the results considering the manual annotation of relation descriptors using two criteria (Collovini et al., 2015): exact matching (having all words in commnon) and partial matching (having at least one word in common). The results considering of number of correct (#C), Recall (R), Precision (P) and F-measure (F) for exact and partial matching are presented in Table 5, respectively.

| | #C | R | P | F |
|---|---|---|---|---|
| Exact matching | 28 | 0.43 | 0.70 | 0.53 |
| Partial matching | 35 | 0.54 | 0.87 | 0.67 |

Table 5: Results of the relation extraction of the subset from Summ-it

### 3.4. Coreference

The coreference information was extracted from Summ-it (Collovini et al., 2007). The original Summ-it corpus contains 560 coreference chains (annotated manually) with an average of 3 members (noun phrases for each chain). The largest chain has 16 mentions (noun phrases).

## 4. Conversion to SemEval

The conversion to the SemEval format began with the parsing of the original Summ-it texts. For that we used CoGrOO (Silva, 2013) to extract the tokens which form the base structure of the SemEval format.
Then morphosyntactic, named entities, entity relations and coreference were converted as follows.

### 4.1. Morphosyntactic Annotation

The morphosyntactic annotation is simply extracted by CoGrOO and displayed in the appropriate columns. CoGrOO (Silva, 2013) parses the natural language text from the original corpus and breaks it into tokens which are used to structure the SemEval format (Recasens et al., 2010). The morphosyntactic annotations (lemma, part-of-speech, gender and number features) are then displayed raw, as obtained from the parser, in the SemEval format in their respective columns.

### 4.2. Named Entities

The layer with the NEs was generated using the output from the NERP-CRF (**?**) classifier described in Section 3.2. The output consists of the identified and classified NEs extracted from the Summ-it texts. The NEs were then paired with the tokens included in the SemEval file. The matching was done through the criteria of exact matching of a token and a NE in the same sentence. The matched token in the SemEval format is then marked with the correct NE category.

### 4.3. Semantic Relation

The entity relation layer was obtained from the output of the CRF classifier (Collovini et al., 2014) described in Section 3.3. The data consists of the relation descriptors that were correctly extracted by the classifier (partial and exact matching) in the triple format (NE1, relation descriptor, NE2). Next, the elements in the triples were matched to tokens in the sentences of the SemEval file. A match of the three elements in a sentence signals a matched triple. In the matching process, some triples were disregarded due to error in the NE identification step.

### 4.4. Coreference

The coreference annotation was extracted from Summ-it (Collovini et al., 2007) and used to identify mentions and coreference links. The NP matching is based on the head nouns. The coreference chain information is then included in the SemEval file, following the pairing realized in the previous step. It is important to note the annotation in Summ-it++ considers noun phrases as captured by the CoGrOO parser, sometimes different from the original Summ-it annotation.

## 5. Conclusion

In this paper we presented a new version the Summ-it corpus. This new version was enriched with two additional layers, named entities and entity relations. These layers were obtained with the help of tools being developed in our research group. The output of the tools was analysed and corrected to be included in Summ-it++. As one main contribution we produced a unified corpus, which may contribute to the study of several NLP tasks, such as: Coreference Resolution, Relation Extraction, Named Entities Recognition, among others. The corpus is freely available[2]. As further work, we want to perform more detailed corrections of the resulting corpus, and increase the number of texts.

## Acknowledgments

The authors acknowledge the financial support of CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) and FAPERGS (Fundação de Amparo à Pesquisa do Rio Grande do Sul).

[2] http://www.inf.pucrs.br/linatural/summit_plus_plus.html

## 6. References

Amaral, D. O. F. d. and Vieira, R. (2014). Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. *Linguamática*, vol. 6, pages 41–49.

Collovini, S., Carbonel, T. I., Fuchs, J. T., Coelho, J. C., Rino, L., and Vieira, R. (2007). Summ-it: Um corpus anotado com informações discursivas visando a sumarização automática. In *Proceedings of V Workshop em Tecnologia da Informação e da Linguagem Humana , Rio de Janeiro, RJ, Brasil*, pages 1605–1614.

Collovini, S., Pugens, L., Vanin, A. A., and Vieira, R. (2014). Extraction of relation descriptors for portuguese using conditional random fields. In *Proceedings of Advances in Artificial Intelligence - IBERAMIA 2014 - 14th Ibero-American Conference on Artificial Intelligence*, pages 108–119, Santiago de Chile, Chile.

Collovini, S., de Bairros Filho, M., and Vieira, R. (2015). Analysing the role of representantion choices in portuguese relation extraction. In *Proceedings of Conference and Labs of the Evaluation Forum - CLEF 2015*, pages 91–102, Toulouse, France. Springer.

Coreixas, T. (2010). Resolução de correferência e categorias de entidades nomeadas. Dissertação de Mestrado, Pontifícia Universidade Católica Do Rio Grande Do Sul.

Da Silva, F. J. V., Carvalho, A. M. B. R., and Roman, N. T. (2010). A comparative analysis of centering-based algorithms for pronoun resolution in portuguese. In *Proceedings of Advances in Artificial Intelligence–IBERAMIA 2010*, pages 336–345. Springer.

de Souza, J. G. C., Gonçalves, P. N., and Vieira, R. (2008). Learning coreference resolution for portuguese texts. In *Computational Processing of the Portuguese Language - LNCS 5190*, pages 153–162. Springer.

Fonseca, E. B., Vieira, R., and Vanin, A. A. (2014). Coreference resolution in portuguese: Detecting person, location and organization. In *Journal of the Brazilian Computational Intelligence Society*, volume 12, pages 86–97.

Freitas, C., Mota, C., Santos, D., Oliveira, H. G., and Carvalho, P. (2010). Second harem: Advancing the state of the art of named entity recognition in portuguese. In *Proceedings of Language Resources and Evaluation Conference - LREC 2010*.

Gabbard, R., Freedman, M., and Weischedel, R. (2011). Coreference for learning to extract relations: yes, virginia, coreference matters. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Vol. 2*, pages 288–293. Association for Computational Linguistics.

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.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall series in Artificial Intelligence. Pearson Education Ltd., London, 2 edition.

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.

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010). Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8. Association for Computational Linguistics.

Silva, W. D. C. (2013). Aprimorando o corretor gramatical cogroo. Dissertação de Mestrado, Universidade de São Paulo.