Nominal Coreference Resolution using Semantics in Portuguese

Author: Evandro Fonseca

Advisor: Renata Vieira
Co-advisor: Aline Vanin

Pontifícia Universidade Católica do Rio Grande do Sul

September 26th, 2018

Summary

- Introduction
- Coreference Resolution
- Related Work
- Proposed Model
- Experiments and Results
- Error Analysis
- Final Remarks

Introduction

- Coreference Resolution is one of the great challenges in NLP area;
- In less resourceful languages this challenge is even greater, due to:
  - most of the approaches are focused in machine learning;
  - usually there are no rich corpora, containing coreference samples;
- When we deal with semantics, this lack increases.

Introduction

- Machine learning may not be the best solution for Portuguese;
- In this presentation we show that:
  - Linguistic rules may be a good solution to solve the problem of languages like Portuguese;
  - Semantics may provide significant improvements.

Coreference Resolution

Task

- Process that consists in identifying the different mentions made to a specific entity in a discourse:

A discussão sobre a biotecnologia nacional está enviesada, pois está sendo entendida como sinônimo de transgenia. A opinião é do agrônomo Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina). Guerra participou do debate "Biotecnologia para uma Agricultura Sustentável", realizado ontem durante a 52ª Reunião Anual da SBPC (Sociedade Brasileira para o Progresso da Ciência), sobre as biotecnologias apropriadas ao desenvolvimento do país. Guerra citou a micropropagação de vegetais (produção de mudas em laboratório, feita para evitar doenças e selecionar vegetais saudáveis) como exemplo de biotecnologia de baixo custo. Com ela, aumentou-se a produção de moranguinho, no sul do país, de 3,2 kg para 60 kg por hectare. Para o agrônomo, o Brasil deve...

[A discussão sobre [a biotecnologia nacional [1]] [0]] está enviesada , pois está sendo entendida como sinônimo de transgenia . A opinião é de [o agrônomo Miguel Guerra, de [a UFSC [5]] [6]] ([Universidade Federal de Santa Catarina [6]]) . [Guerra [5]] participou de [o debate [0]] "[Biotecnologia para uma Agricultura Sustentável [0]]" , realizado ontem durante a 52ª Reunião_Anual de [a SBPC [10]] ([Sociedade Brasileira para o Progresso da Ciência [10]]) , sobre [as biotecnologias [1]] apropriadas a o desenvolvimento de[o país [13]] . [Guerra [5]] citou a [micropropagação de vegetais [18]] ([produção de mudas em laboratório , feita para evitar doenças e selecionar vegetaissaudáveis [18]]) como exemplo de [biotecnologia [1]] de baixo custo . Com ela , aumentou -se a produção de moranguinho , em o sul de [o país [13]] , de 3,2 kg para 60 kg por hectare . Para [o agrônomo [5]] , [o Brasil [13]] deve...

Coreference Resolution

Coreference chain:

- By grouping these mentions we form sets of mentions, most known as coreference chains:

A discussão sobre a biotecnologia nacional está enviesada , pois está sendo entendida como sinônimo de transgenia . A opinião é de **[o agrônomo Miguel Guerra, de a UFSC [5]]** (Universidade Federal de Santa Catarina) . **[Guerra [5]]** participou de o debate “Biotecnologia para uma Agricultura Sustentável” , realizado ontem durante a 52ª Reunião_Anual de a SBPC (Sociedade Brasileira para o Progresso da Ciência) , sobre as biotecnologias apropriadas a o desenvolvimento de o país. **[Guerra [5]]** citou a micropropagação de vegetais (produção de mudas em laboratório , feita para evitar doenças e selecionar vegetaissaudáveis) como exemplo de biotecnologia de baixo custo . Com ela , aumentou -se a produção de moranguinho , em o sul de o país , de 3,2 kg para 60 kg por hectare . Para **[o agrônomo [5]]**, o Brasil deve...

Coreference Resolution

Issues

- How to provide computational level processing in order to recognize some linguistic patterns?

Coreference Resolution

Issues

- There are cases where the coreference relation is simple to grasp, such as in:
  - [Miguel **Guerra**], [**Guerra**]
- Both share some identical part.

Coreference Resolution

Issues

- However, we must consider more difficult cases:
  - [**o sul** do Brasil], [**o sul** dos Estados Unidos]
  - [**as regiões** mais úmidas], [**as regiões** mais secas]
  - [**Universidade** do Paraná], [**Universidade** de São Paulo]
  - [o copo de **vidro**], [**vidro**]

Coreference Resolution

Issues

- There are even more complex cases:
  - [Adalberto **Portugal**], [**Portugal**]
- Does Portugal refer to “Adalberto” or to Portugal, the “country”?

Coreference Resolution

Issues

- There are cases with no word matching:
  - [o menino], [o garoto]
  - [a França], [o país]
  - [o fóssil], [os ossos]
- Plus, note that gender and number may vary;
- For these and other cases, we must rely on semantic resources.

Related Work

- How does the literature handle to this task?

Related Work

Portuguese and other Languages

Model	Approach		Language				Semantics	
	Machine Learning	Rules	EN	PT	SPA	GL	NE Cat	Sem Rel
[Martschat and Strube2015]	√		√				√	
[Yang et al.2008]	√		√				√	
[Ng and Cardie2002]	√		√				√	
[Chang et al.2012]	√		√				√	
[Soon et al.2001]	√		√				√	
[Hou et al.2014]		√	√				√	√
[Lee et al.2013]		√	√				√	
[Rahman and Ng2011]	√		√				√	√
[Garcia and Gamallo2014a]		√		√	√	√	√	
[Silva2011]	√			√			√	√
Our [Fonseca et al.2017b]		√		√			√	√

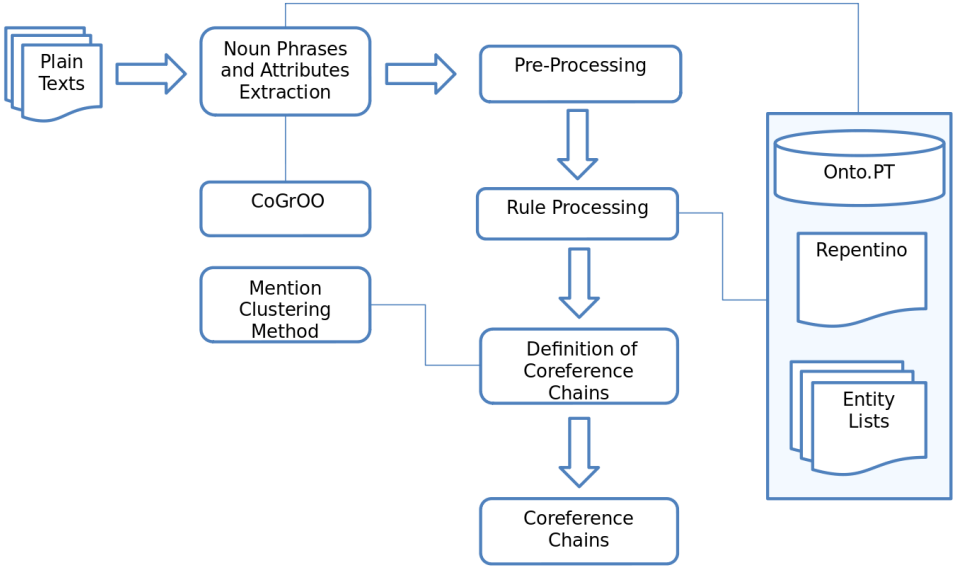
Coreference Resolution task and our scope

- Our proposal consists in a process for automatic nominal coreference resolution in Portuguese that incorporates semantic information;
- Eleven lexical and syntactic rules were adapted from [Lee et al.2013] and two new semantic rules were proposed:
  - Synonymy and Hyponymy[1];
- Our approach is domain independent.

[1] In natural language is most common we refer to an entity of more specific way and, after, most broad.

Proposed Model

Overview



Experiments and Results

Corpora

- We perform our experiments using three Portuguese corpora:
- Summ-it++ [Antonitsch et al.2016]
  - Latest version of the Summ-it corpus;
  - Gold chains;
  - Composed by 50 texts and 560 coreference chains.
- Corref-PT [Fonseca et al.2017a]
  - Semi-automatically annotated;
  - Composed by 182 texts and approximately 4000 coreference chains.
- Garcia et al.'s corpus
  - Composed by 97 texts (considering just Portuguese texts);
  - Gold chains (annotation just for Person entities).

Experiments and Results

Evaluation – Non-Comparative Results

Model		Language	MUC			CoNLL
			P	R	F	F
[Martschat and Strube2015]		EN	76.8	68.1	72.2	62.5
[Fernandes et al.2014]		EN	75.9	65.8	70.5	63.4
		CH	71.5	59.2	64.8	62.9
		AR	49.7	43.6	46.5	54.2
[Lee et al.2013]		EN	60.9	59.6	60.3	59.4
[Garcia and Gamallo2014a]		ES	94.1	84.1	88.8	79.2
		GL	94.6	89.0	91.7	84.4
		PT	92.7	82.7	87.4	78.9
Our model (Summ-it++)	without semantics	PT	58.8	44.4	50.6	51.2
	with semantics		45.1 (-13.7)	52.1 (+7.7)	48.3 (-2.3)	48.6 (-2.6)
Our model (Corref-PT)	without semantics		64.2	47.8	54.8	51.4
	with semantics		54.9 (-9.3)	50.2 (+2.4)	52.5 (-2.3)	49.7 (-1.7)

Experiments and Results

Evaluation – Comparative Results

- We perform a comparative analysis between our and Garcia et al.'s model;
- Two texts from Garcia et al's corpus [Garcia and Gamallo2014b]

Model	MUC			CONLL
	P	R	F	F
[Garcia and Gamallo2014a]	97.9	96.0	97.0	90.5
Our model	80.0	16.0	26.7	21.6

- Our model has presented lower recalls (considering proper names only), we do not treat pronominal coreference;
- However we identify two coreference chains that are not considered by Garcia;
- We believe that to perform a fair evaluation, we should involve Summ-it++ and Corref-PT.

Error Analysis

O estado de **[São_Paulo [1]]** voltou a sofrer com **[os ataques [2]]** contra postos policiais , agências bancárias e ônibus , em a madrugada de esta segunda-feira . **[As ações [2]]** são atribuídas a a facção criminosa Primeiro_Comando de **[a Capital [9]]** (**[PCC [9]]**) , que já comandou outros ataques em duas ocasiões . Os ataques de esta madrugada , até agora , não deixaram mortos ou feridos . **[Uma bomba de fabricação caseira [15]]** explodiu em frente a o prédio de o Ministério_Público_Estadual e lojas vizinhas também foram atingidas por estilhaços . **[A rua [22]]** está interditada para a perícia e , a os poucos , os comerciantes são autorizados a entrar em seus estabelecimentos . A Secretaria_da_Fazenda também foi atingida por **[uma bomba [15]]** . Duas bases de **[a Guarda_Civil_Metropolitana [29]]** (**[GCM [29]]**) , sendo uma em o Capão_Redondo , Zona_Sul de **[São_Paulo [1]]** , foram alvo de os criminosos . Mais de dez agências bancárias , um posto de gasolina e um supermercado foram atacados . Calcula -se em 15 o número de ônibus incendiados , sendo dez em a região de o ABC e quatro em **[a capital [9]]** . Mesmo_assim , o sistema de transporte coletivo de a cidade está normal em esta manhã . **[A rua [22]]** onde fica **[o Departamento_de_Investigações_Sobre_o_Crime_Organizado ([47]] [Deic [47]]**) foi bloqueada e a passagem de veículos está proibida . . .

- [1] – [São Paulo](state) and [São Paulo] (city) were grouped;
- [9] – mention detection [~~Primeiro comando de~~ a Capital] and [PCC]; [a Capital] → [São Paulo] (city);
- [15] – [Uma bomba de fabricação caseira] and [uma bomba][2] were grouped.

[2] "bomb that hit Ministério Público and the bomb that hit Secretaria da Fazenda".

Final Remarks

- This is the first work to adapt and evaluate Lee's set of rules [Lee et al.2013] for Portuguese;
- We incorporated semantic knowledge to this process;
- Semantics may improve the task increasing the recall of our model;
- Semantic approaches are rare, even for English.

Main Contributions of this PhD Thesis

- A new process for Coreference Resolution in Portuguese [Fonseca et al.2017b]; (Linguamática)
- Our model solve coreferences using plain texts as input;
- Using semantics we found new coreference relations, impossible when we use lexical and syntactic processing;
- A new mention clustering process [Fonseca et al.2018]; (NLDB)

Main Contributions of this PhD Thesis

- Two new Portuguese corpora:
  - Summ-it++ [Antonitsch et al.2016] (LREC);
  - Corref-PT [Fonseca et al.2017a] (IBEREVAL).
- CorrefVisual [Sesti et al.2017]
  –tool to visualize and edit coreference chains (TILic);
- CORP [Fonseca et al.2016]
  –coreference resolution tool for Portuguese (PROPOR);
- All resources are available at PLN-PUCRS website[3]

[3] http://www.inf.pucrs.br/linatural/wordpress/index.php/recursos-e-ferramentas/

Thank you =)

Antonitsch, A., Figueira, A., Amaral, D., Fonseca, E., Vieira, R., and Collovini, S.
(2016).
Summ-it++: an enriched version of the summ-it corpus.
In *Proceedings of 10th edition of the Language Resources and Evaluation Conference*, Portorož, Slovenia.

Chang, K.-W., Samdani, R., Rozovskaya, A., Sammons, M., and Roth, D.
(2012).
Illinois-coref: The ui system in the conll-2012 shared task.
In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 113–117. Association for Computational Linguistics.

Fernandes, E. R., dos Santos, C. N., and Milidiú, R. L.
(2014).
Latent trees for coreference resolution.
Computational Linguistics.

Fonseca, E. B., Vieira, R., and Vanin, A.
(2016).
Corp: Coreference resolution for portuguese.
In *12th International Conference on the Computational Processing of Portuguese, Demo Session*.

Fonseca, E., Sesti, V., Collovini, S., Vieira, R., Leal, A. L., and Quaresma, P.
(2017a).
Collective elaboration of a coreference annotated corpus for portuguese texts.
In *Proceedings of II workshop on Evaluation of Human Language Technologies for Iberian Languages*, volume 1881, pages 68–82, Murcia, Spain.

Fonseca, E. B., Sesti, V., Antonitsch, A., Vanin, A. A., and Vieira, R.
(2017b).

Corp - uma abordagem baseada em regras e conhecimento semântico para a resolução de correferências.
Linguamatica, 9(1):3–18.

Fonseca, E., Vanin, A., and Vieira, R.
(2018).
Mention clustering to improve portuguese semantic coreference resolution.
In *Natural Language Processing and Information Systems*.

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

Garcia, M. and Gamallo, P.
(2014b).

Multilingual corpora with coreferential annotation of person entities. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference*, pages 3229–3233, Reykjavik, Iceland.

Hou, Y., Markert, K., and Strube, M. (2014). A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2082–2093, Doha, Qatar.

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.

Martschat, S. and Strube, M.

(2015).
Latent structures for coreference resolution.
Transactions of the Association for Computational Linguistics, 3:405–418.


Ng, V. and Cardie, C.
(2002).
Improving machine learning approaches to coreference resolution.
In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104–111. Association for Computational Linguistics.


Rahman, A. and Ng, V.
(2011).
Coreference resolution with world knowledge.
In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 814–824, Portland, Oregon, USA.


Sesti, V., Fonseca, E., and Vieira, R.

(2017).
Correfvisual: Ferramenta para a edição de correferências.
In *V Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana*, Uberlândia, Minas Gerais.

Silva, J. F. d.
(2011).
Resolução de correferência em múltiplos documentos utilizando aprendizado não supervisionado.

Soon, W. M., Ng, H. T., and Lim, C. Y.
(2001).
A machine learning approach to coreference resolution of noun phrases.

Computational Linguistics, 27(4):521–544.

Yang, X., Su, J., Lang, J., Tan, C. L., Liu, T., and Li, S.
(2008).

An entity-mention model for coreference resolution with inductive logic programming.
In *Proceeding of Association for Computational Linguistics*, pages 843–851.