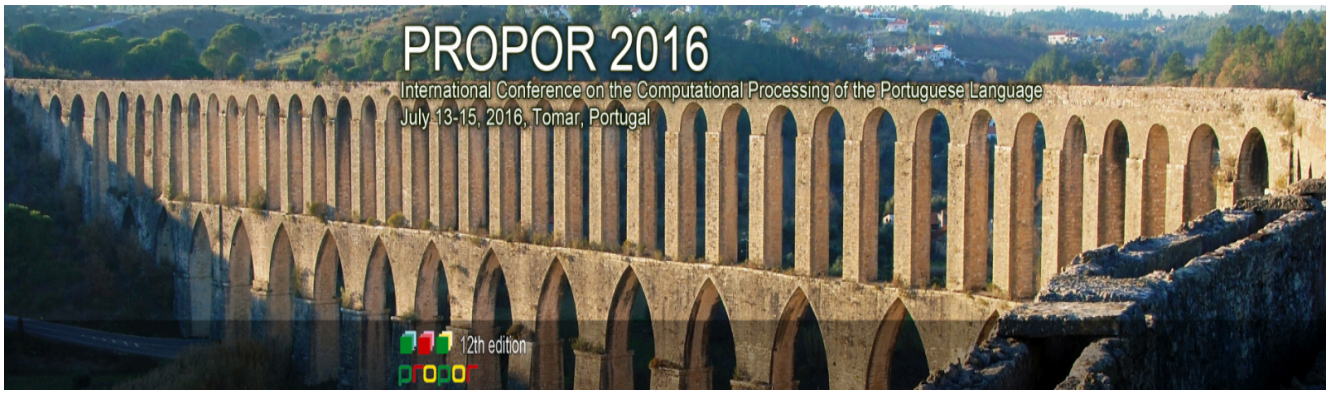


DEMONSTRATION SESSION

Following the spirit of the Demos Session of PROPOR'2010, 2012 and 2014, the PROPOR 2016 demonstration track aims at bringing together Academia and Industry and creating a forum where more than written or spoken descriptions of research are available. Thus, demos allow attendees to try and test them during their presentation in a dedicated session adopting a more informal setting. Products, systems or tools are examples of accepted demos, where both early-research prototypes and mature systems were considered.

July 13-15, 2016, Tomar - Portugal

Vládia Pinheiro (Unifor/Brazil)
Hugo Gonçalo Oliveira (CISUC, Universidade de Coimbra, Portugal)
Demos Chairs

ACCEPTED DEMOS

- **A Computational Tool for Automated Language Production Analysis Aimed at Dementia Diagnosis** – *Sandra Aluísio, Andre Cunha, Cintia Toledo and Carolina Scarton*
- **Annotating Portuguese Corpora with Word Senses Using LX-SenseAnnotator** – *Steven Neale and António Branco*
- **CORP: Coreference Resolution for Portuguese** – *Evandro Fonseca, Renata Vieira and Aline Vanin*
- **Hookit: natural language processing in a semantic based platform for social commerce** – *Sandro José Rigo, Vinicius Dambros Andrade and Denis Andrei Araújo*
- **LetsRead – Tool to Automatically Evaluate Children's Reading Aloud Performance** – *Jorge Proença, Dirce Celorico, Carla Lopes, Sara Candeias and Fernando Perdigão*
- **NILC-WISE: An Easy-to-use Web Interface for Summary Evaluation with the ROUGE Metric** – *Fernando Antônio Asevedo Nóbrega and Thiago Alexandre Salgueiro Pardo*
- **OpenWordnet-PT** – *Fabricio Chalub, Livy Real, Valeria de Paiva and Alexandre Rademaker*
- **Poe, now you can TryMe: Interacting with a Poetry Generation System** – *Hugo Gonçalo Oliveira*
- **Syntax Deep Explorer** – *José Correia, Jorge Baptista and Nuno Mamede*
- **VITHEA-Kids: Improving the Linguistic Skills of Children with Autism Spectrum Disorder** – *Vânia Mendonça, Cláudia Filipe, Luísa Coheur and Alberto Sardinha*
- **XCrimes: Information Extractor for the Public Safety and National Defense Areas** – *Daniel Sullivan, Vladia Pinheiro, Rafael Pontes and Vasco Furtado*

A Computational Tool for Automated Language Production Analysis Aimed at Dementia Diagnosis

Sandra Aluísio[1], Andre Cunha[1], Cintia Toledo[2], and Carolina Scarton[3]

[1] University of São Paulo, São Carlos SP 13566-590, BR,
Interinstitutional Center for Computational Linguistics,
sandra@icmc.usp.br, andre.lv.cunha@gmail.com
[2] Faculty of Medicine of the University of São Paulo, São Paulo SP, 01246-903
citoledo@hotmail.com
[3] University of Sheffield, Regent Court, 211 Portobello, Sheffield, S1 4DP, UK,
Department of Computer Science
carol.scarton@gmail.com

Abstract. Coh-Metrix-Dementia is a unified computational environment that allows for the automated analysis of language production in clinical (and similar) settings. In its current version, the tool is composed of an underlying Python library – built on top of several Natural Language Processing (NLP) tools for the Portuguese language – that can be used as a component inside other applications, and a web interface, suitable for use by health professionals. From a transcribed or written text sample produced by a subject, Coh-Metrix-Dementia can extract 73 textual metrics, comprising several levels of linguistic analysis from word counts to semantics and discourse, which can be used in language evaluation. The tool has been evaluated in automated classification tasks involving dementia patients, where it demonstrated to have a high classification accuracy, and is freely available under the GPLv3 license.

1 Background

Dementia is a socially relevant medical condition that is the result of a progressive and irreversible neurodegenerative disorder, which can manifest itself in several forms, and whose early diagnosis plays a pivotal role in successful clinical interventions. Language is an important source of diagnostic information, capable of revealing where the brain damage is located, and how severe it is. Qualitative language analysis can be performed manually at reasonable cost, but manual quantitative analysis of discourse production is severely time-consuming, hindering its application in clinical settings. In this scenario, automated computational tools that can provide textual metrics for both manual analysis and automatic classification can significantly aid early diagnosis. Automated discourse analysis aiming at the diagnosis of language impairing dementias already exist for the English language [1], but no such work had been done for Portuguese. For that reason, we developed a computational environment, entitled

Coh-Metrix-Dementia, capable of extracting several relevant attributes, called **metrics**, from a patient text. Such metrics can then be used in either manual or automatic language assessment tests.

2 Coh-Metrix-Dementia

The Coh-Metrix-Dementia environment[4] includes a base library, written in the Python programming language, that can be used to access the metric extraction system. The system's architecture is depicted in Figure 1. It receives as input the text that is to be analyzed, in a suitable format: the text is supposed to be separated in sentences, following the traditional rules of capitalization and punctuation. If the text is the transcription of a patient's speech sample, the text has to be manually segmented into sentences prior to being analyzed by Coh-Metrix-Dementia. After analyzing the text, the tool gives as output a set of 73 textual metrics, divided in 14 categories: Ambiguity, Anaphoras, Basic counts, Connectives, Constituents, Coreference, Disfluencies, Frequencies, Hypernyms, Logic operators, Latent Semantic Analysis, Semantic density, Syntactical complexity, and Tokens.

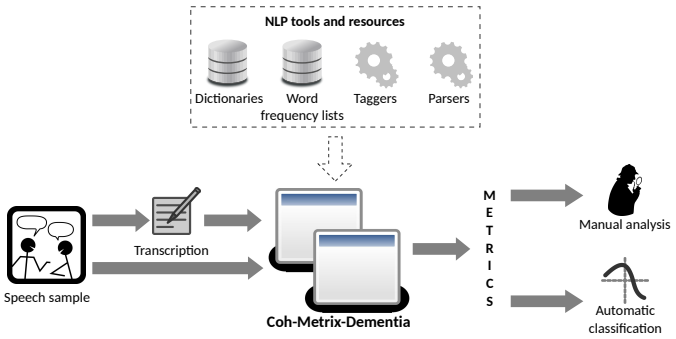


Fig. 1: Coh-Metrix-Dementia's general architecture.

On top of the basic library, a web interface was created that allows end users, such as physicians and other health professionals, to access the library's functionalities in an easy-to-use way. First, the user logs into the system; then, a home screen is shown, which displays a list of the already submitted texts (at first, this list is empty); if the user clicks on the *Submit text* button at the top right corner of this screen, a text submission screen is displayed, where the user can submit a text for analysis ; when analysis is done, the user is redirected back to the home screen, which now displays an entry for the newly submitted text; if the user clicks on this entry, a text overview screen is displayed, informing the value of each of the 73 metrics for that text.

[4] http://143.107.183.175:22380/

3 Evaluation and demonstration

Coh-Metrix-Dementia has been evaluated for use in the automated diagnosis of dementia [2]. In this evaluation, the metrics extracted by the tool, along with some metrics extracted manually from speech transcriptions of Alzheimer's Disease (AD) and Mild Cognitive Impairment (MCI) patients, were employed in automated diagnosis scenarios using classification and regression techniques. In our experiments with classification, it was possible to separate healthy control subjects, AD, and MCI with 0.817 F_1 score, and separate controls and MCI with 0.900 F_1 score. As for regression, the best results for the mean absolute error (MAE) were 0.238 and 0.120 for scenarios with three and two classes, respectively. We also employed Coh-Metrix-Dementia's metrics in order to observe those that better discriminate and provide features about the investigated groups (AD, MCI and Control Group); 27 were able to differentiate the groups. The results revealed some unexpected items which are not in accordance with the literature. Efforts to understand the unexpected results showed that all those metrics are related to segmentation of sentences. New methods for performing segmentation of the sentences will be studied to approach impaired speech. Statistical analysis allowed differentiating the characteristics of each group. Regarding macrostructural aspects the main metrics were empty emissions ($p < 0.001$), mean between adjacent sentences ($p < 0.033$), standard deviation givenness sentences ($p < 0.016$) and idea density ($p < 0.001$) which indicated that AD group has more repetitive discourse and without introducing new information compared to other groups [3]. Regarding microstructural aspects the main results were difficulties in verbal rescue ($p < 0.001$) characterized by the presence of breaks (long and short) in greater quantity in the speech of individuals with AD. In the demonstration session to be held in PROPOR 2016, attendees will be able to use Coh-Metrix-Dementia's web interface to extract metrics from any text, and see how the tool informs such values and how they can be downloaded for use in any intended application. Texts with disfluencies are submitted in two versions: raw content, with annotation of disfluencies, using XML tags, and revised content, without these annotations.

References

1. Fraser, K.C., M.J., Rudzicz, F.: Linguistic features identify Alzheimer's disease in narrative speech. Journal of Alzheimer's Disease **49**(2) (2015) 407–422
2. Aluisio, S., Cunha, A., Scarton, C.: Evaluating progression of alzheimer's disease by regression and classification methods in a narrative language test in portuguese. In Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A., eds.: 12th International Conference, PROPOR 2016, Tomar, Portugal, July 13-15, 2016, Proceedings. (2016) To appear
3. Bryant, L., Spencer, E., Ferguson, A., Craig, H., Colyvas, K., Worrall, L.: Propositional Idea Density in aphasic discourse. Aphasiology **27**(8) (2013) 992–1009

Annotating Portuguese Corpora with Word Senses Using LX-SenseAnnotator

Steven Neale and António Branco

NLX - Natural Language and Speech Group
Faculty of Sciences, Department of Informatics
University of Lisbon, Portugal
{steven.neale, antonio.branco}@di.fc.ul.pt

Abstract. This paper describes LX-SenseAnnotator, an accessible and easy-to-use interface tool for manual annotating text with word senses. We demonstrate how the tool was used to manually annotate the CINTIL-WordSenses corpus, outlining the loading and browsing of Portuguese texts and how word senses themselves are selected and assigned. We also describe the potential for LX-SenseAnnotator to be adapted for other languages besides Portuguese.

Keywords: manual annotation, word senses, corpora, interface tools

1 Introduction

As part of our research on word sense disambiguation (WSD) in Portuguese, we have encountered the need for a simpler, more accessible way to annotate texts with word senses in order to quickly and easily create gold standard corpora for training and evaluation. Annotated corpora are hugely important, supporting both the analysis of large quantities of text [3] and the training and evaluation of natural language processing (NLP) tools, and there is an increasing interest in producing corpora containing "high-quality linguistic annotations at the semantic level" [6].

Many of the current unsupervised approaches to WSD assign the eight-digit 'synset identifiers' from ontologies and lexicons such as the Princeton WordNet – within which nouns, verbs, adjectives and adverbs are grouped into sets of synonyms or 'synsets' [2] – to unlabelled raw text. However, despite the obvious need for corpora manually-annotated with synset identifiers against which to evaluate these approaches, manual word sense annotation tools are either difficult to come by or seem intrinsically tied to specific corpora or lexica.

In this paper, we describe LX-SenseAnnotator, a user-interface tool designed specifically to offer a more open and flexible way to annotate texts with senses pulled from WordNets in the Princeton format. While LX-SenseAnnotator has the scope to become flexible enough to be adapted to other languages in future versions, its current implementation is formatted to handle Portuguese texts specifically and was recently used to manually annotate the first version of the gold-standard CINTIL-WordSenses corpus [5].

2 Using LX-SenseAnnotator

In its current Portuguese implementation, LX-SenseAnnotator is designed for importing text files that have already been part-of-speech (POS) tagged and lemmatized using the LX-Suite, an existing pipeline of shallow processing tools for Portuguese [1]. The POS tags are then used to organize each word in the imported text according to whether they are or are not sense-taggable. Nouns, verbs, adjectives and adverbs whose lemma is present in the WordNet being used (potential candidates for sense tagging) are marked in red – so that they can be easily distinguished from the rest of the text, which is marked in dark blue – unless they have already been assigned a synset identifier, in which case they are marked in green (Figure 1).

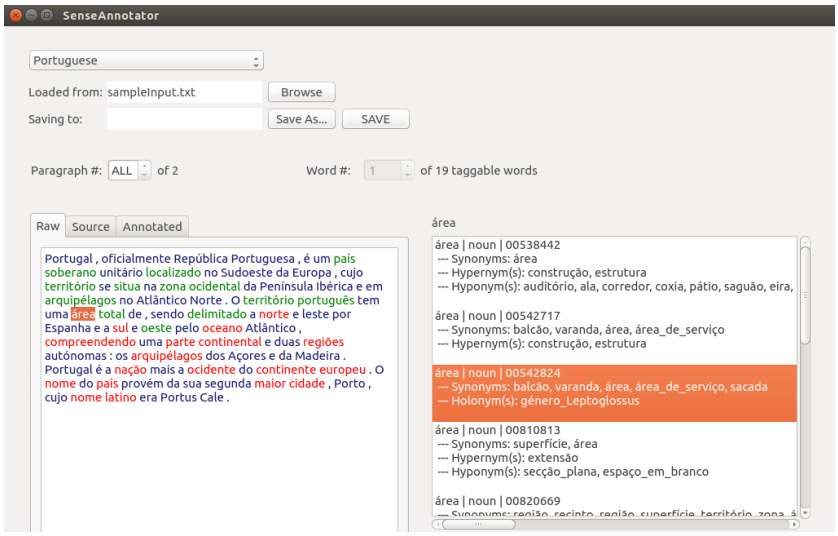


Fig. 1. Displaying a list of senses for the word 'área' (English 'area') using LX-SenseAnnotator.

Annotators can either click on red, sense-taggagable words or use the scroll box in the middle of the interface to browse through all currently sense-taggable words in the text, and when a sense-taggable word is selected it is highlighted and available senses from a pre-loaded WordNet are displayed in the results panel on the right-hand side of the interface. The lemmas and POS tags of specific words queried against the WordNet's index.sense file when they are highlighted by the annotator, and for every entry that the selected word has in the WordNet the appropriate eight-digit synset identifier – as well as additional contextual information from the synset such as synonyms, hypernyms, hyponyms, holonyms, antonyms and so on – is displayed in the list of available senses.

Once an annotator has decided which of the available senses is most appropriate for a given word – taking into account its discursive context – they simply double click on the sense to automatically assign it to the selected word, which now becomes green in the text display on the left-hand side of the interface. The word is removed from the list of sense-taggable words, although they can still be selected should annotators decide to remove the annotated sense or choose a more appropriate one later. This process was used to assign senses from the Portuguese MultiWordNet [4] to 45,502 words across 23,825 sentences for the first version of CINTIL-WordSenses.

3 Flexibility and Adaptability

For producing the first version of CINTIL-WordSenses the pre-loaded WordNet used was the Portuguese MultiWordNet, but because any WordNet in any language can be loaded into LX-SenseAnnotator providing that it adheres to the traditional Princeton format the tool is potentially extremely flexible. An important direction for future work is to take advantage of this by adapting the way texts are imported such that different pre-tagged text formats and POS tagsets are supported, which coupled with the flexibility of pre-loading WordNets would extend LX-SenseAnnotator for use with many more languages.

Acknowledgements

This work has been undertaken and funded as part of the EU project QTLeap (EC/FP7/610516) and the Portuguese project DP4LT (PTDC/EEI-SII/1940/2012).

References

1. Branco, A., Silva, J.R.: A Suite of Shallow Processing Tools for Portuguese: LX-suite. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics: Posters and Demonstrations. pp. 179–182. EACL '06, Association for Computational Linguistics, Trento, Italy (2006)
2. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press (1998)
3. Leech, G.: Adding Linguistic Annotations. In: Wynne, M. (ed.) Developing Linguistic Corpora: A Guide to Good Practice. AHDS Literature, Languages and Linguistics (2004)
4. MultiWordNet: The MultiWordNet project. `http://multiwordnet.fbk.eu/english/home.php` (nd), accessed: 2014-01-13
5. Neale, S., Pereira, R.V., Silva, J., Branco, A.: Lexical Semantics Annotation for Enriched Portuguese Corpora. In: Proceedings of the 12th International Conference on the Computational Processing of the Portuguese Language. PROPOR 2016, Tomar, Portugal (2016)
6. Passonneau, R.J., Baker, C., Fellbaum, C., Ide, N.: The MASC Word Sense Sentence Corpus. In: Proceedings of the 8th International Conference on Language Resources and Evaluation. European Language Resources Association, Istanbul, Turkey (2012)

CORP: Coreference Resolution for Portuguese

Evandro Fonseca[1], Renata Vieira[1] and Aline Vanin[2]

evandro.fonseca@acad.pucrs.br, renata.vieira@pucrs.br,
aline.vanin@ymail.com

[1]Pontifícia Universidade Católica do Rio Grande do Sul
[2]Universidade Federal de Ciências da Saúde de Porto Alegre

Abstract. This paper describes CORP, an open source, off-the shelf noun phrase coreference resolver for Portuguese with a web interface.

1 Introduction

We are building an open-source off-the-shelf system, which solves Portuguese noun phrase coreference, using plain texts as input. Our tool goes beyond basic syntactic heuristics, such as string matching, copular, juxtaposition. We consider semantics. In other words, string matching heuristics serve to deal with cases such as "Miguel Guerra" and "Guerra", in which both NPs share some identical part. Copular constructions are used to link two mentions as in "Miguel Guerra is the agronomist". Juxtaposition refers to cases of appositive constructions such as "Miguel Guerra, the agronomist". So far, it may seem a simple problem, but refined syntactic knowledge must be taken into account even in those cases. For instance, we do not want to find that "mushrooms found in Brazil" is coreferent with "mushrooms found in France". In other situations, establishing a coreference relation is even more difficult. In cases such as "the boy" and "the kid", there is a semantic relation which is usually part of the readers' common sense knowledge. We are currently dealing with this sort of problem. The current version of the system is available through a web interface and is detailed below.

2 CORP Architecture

In this research we are using what is currently available for pre processing tasks. As we are developing a system in Java, we have used Java based open source tools such as Cogroo [2] and OpenNLP[1] . OpenNLP provides POS tagging and named entities recognition, while Cogroo provides noun phrase chunks and shallow structure. For our studies, we use the correference annotated Summ-it corpus [1]. Our system is an adaptation of the model proposed in [5]. We adapted and implemented a set of modules. The first two correspond to noun phrase extraction and filtering. The other modules are used to link two mentions if the conditions established by linguistic rules are satisfied. These modules are described in detail in [3]. Recently, we added two sematic modules (Hyponymy and Synonymy) based on the relations provided by ONTO-PT [6]. An experiment using semantic knowledge is reported in [4].

[1] http://opennlp.apache.org/

3 CORP - Web Interface

A demonstration of the system is available at "http://ontolp.inf.pucrs.br/corref/". The interface is intuitive and contains: a upload button, to submit the text; "Limpar texto" to clear input text and the output and three example buttons, containing samples of previously processed texts (Figure1). When submitting an input text, the system returns its coreference chains.

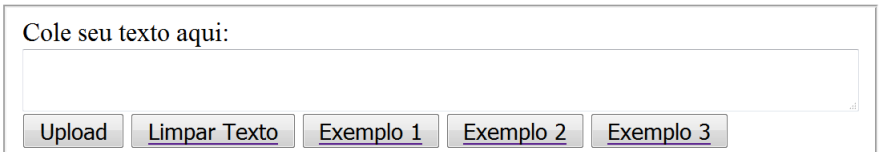


Fig. 1. CORP Web Interface

[O trabalho de [pesquisadores [166]] de [a USP [29]] [136]] está revelando [uma série de [novas espécies [68]] de [um tipo [68]] todo especial de [fungo [157]] [169]] : [pequenos cogumelos [157]] que emitem [uma misteriosa luminosidade verde [51]] em [o escuro [74]] . As criaturas , antes desconhecidas em [o Brasil [72]] , podem ajudar a elucidar [o mecanismo bioquímico [167]] que leva a [a produção [136]] de [luz [141]] em [fungos [157]] . Além_disso isso , com um pouco mais de [estudo [32]] , poderiam servir como [sensores vivos de [poluição [140]] ou mesmo fontes de moléculas úteis para [a biotecnologia [32]] [139]] . Segundo [Cassius_Vinicius_Stevani [190]] , [químico [77]] de [a USP [29]] que coordena [os estudos [32]] , é possível [que o material recolhido [32]] abranja por o menos_dez [espécies novas [68]] . Não é pouca coisa , já que em o mundo todo se conhecem só 42 espécies de [o fungo [157]] , quase todas restritas a o Sudeste_Asiático . " Já temos

Fig. 2. Correference chains indicated by indexes and colours

Figures 2 and 3 show the output generated by the system. It generates both a coloured version of the text and a corresponding table containining all coreference chains. Note that there are some embedded mentions, such as "USP" in "pesquisadores da USP". In those cases we present the larger expression in the same colour and use a different collor only for the brackets and ID of the inside mention. In the table, unique mentions (mentions that appear only once in the text) are also listed. We use the same colors in the text and table to represent the coreference chains. In this example we see that a semantic rule has been used when matching fungos and cogumelos (*funghi and mushroom*).

4 Conclusion

In this paper, we presented a rule-based coreference resolution system for Portuguese. We believe that this tool may help many researchers, due to fact that the coreference resolution task may help in several NLP tasks.As further work, we intend to enrich our semantic rules and develop other modules, such as pronominal coreference resolution. The CORP implementation is part of a PhD thesis entitled: "Resolução de Correferências em Língua Portuguesa" (*Coreference Resolution in Portuguese Language*).

	Tokens	Sintagma
CADEIA_68		
SnID: 4	12 ... 13	**novas espécies**
SnID: 5	15 ... 16	**um tipo**
SnID: 33	109 ... 110	**espécies novas**
SnID: 68	237 ... 238	**o tipo**
SnID: 71	251 ... 251	**espécies**
CADEIA_157		
SnID: 7	20 ... 20	**fungo**
SnID: 8	22 ... 23	**pequenos cogumelos**
SnID: 18	58 ... 58	**fungos**
SnID: 39	129 ... 130	**o fungo**
MençõesÚnicas:		
ID: 6	17 ... 18	**todo especial**
ID: 12	34 ... 36	**As criaturas**

Fig. 3. Coreference chains and unique mentions

Acknowledgments

The authors acknowledge the financial support of CNPq, CAPES and FAPERGS.

References

1. S. Collovini, T. I. Carbonel, J. T. Fuchs, J. C. Coelho, L. Rino, and R. Vieira. Summ-it: Um corpus anotado com informações discursivas visando a sumarização automática. In *V Workshop em Tecnologia da Informação e da Linguagem Humana*, 2007.
2. W. D. C. de Moura Silva. Aprimorando o corretor gramatical cogroo. Master's thesis, Universidade de São Paulo, 2013.
3. E. B. Fonseca, R. Vieira, and A. Vanin. Adapting an entity centric model for portuguese coreference resolution. In *Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016)*, In Press, 2016.
4. E. B. Fonseca, R. Vieira, and A. Vanin. Improving coreference resolution with semantic knowledge. In *Proceedings of the 12th International Conference on the Computational Processing of Portuguese (PROPOR 2016)*, In Press, 2016.
5. H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916, 2013.
6. H. G. Oliveira and P. Gomes. Eco and onto. pt: a flexible approach for creating a portuguese wordnet automatically. In *Proceedings of Language Resources and Evaluation Conference*, volume 48, pages 373–393. Springer, 2014.

Hookit: natural language processing in a semantic based platform for social commerce

Sandro José Rigo[1,2], Vinicius Dambros Andrade[2], Denis Andrei Araújo[1,2]
[1]Applied Computing Graduate Program (Pipca) - University of Vale do Rio dos Sinos (Unisinos), São Leopoldo, Brazil
[2]HOOKIT – Porto Alegre, Brazil

rigo@unisinos.br, vini@hookit.cc, denis@hookit.cc

General context

This proposal for the Demo Session at PROPOR 2016 is aimed at exhibiting the practical results of the application of Natural Language Processing resources in a social e-commerce platform, called "Hookit" (available at http://www.hookit.cc). The platform is being developed under support of the TECNOVA/FINEP[1] grant, which has the objective of stimulating startups in Brazil, integrating universities and companies.

The platform itself comprises several modules that allow, in first place, the integration of several products in a friendly and resourceful environment. The users interaction with this environment is processed in modules that implement machine learning algorithms to support personalized recommendation of products. A fashion ontology, called "Fashion Relations", was developed to support semantic interaction to be considered in the recommendations.

As main differential and contribution, the HOOKIT platform applies the semantic support provided by the Fashion Relations ontology to help in the support for natural language interaction with the users, along with the support for the recommendation module. The users can either browse the HOOKIT website with the traditional search and select operations, or they can interact with a chatbot, for a more friendly support in the process of choosing some product.

Prototype characteristics

The figure 1 shows some parts of the HOOKIT website, in which the user has access to normal operations to fulfil the needs in a regular session. In the HOOKIT platform the user can select either products or "Looks". The products are selected directly, from the name and attributes. The "Look" is a collection of several products,

[1] http://www.finep.gov.br/apoio-e-financiamento-externa/programas-e-linhas/descentralizacao/tecnova

that in general is suggested by stylists, and therefore adequate according to some style or occasion.

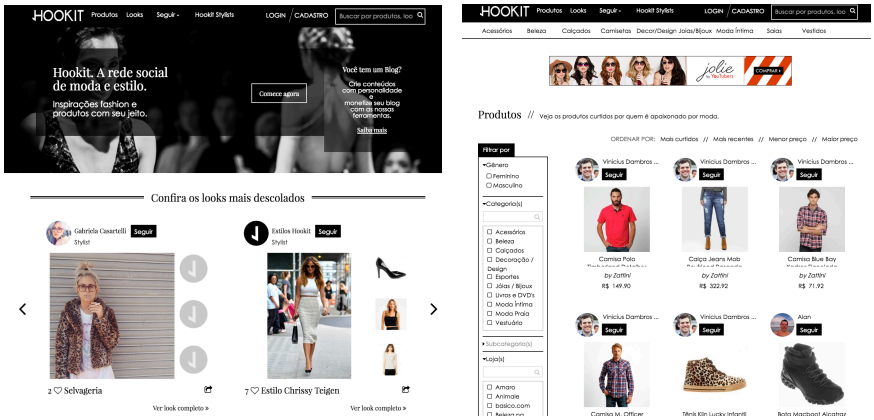


Figure 1: General vision of the HOOKIT website

The figure 2 shows some examples of the interface available for the user to interact with the chatbot. In this example a situation of introductory conversation is exemplified. In the left part of figure 2, the chatbot receives a greeting mention, which is replied with a greeting and a question about the name of the user. The right part of Figure 2 shows the user's answer and then the chatbot uses the informed name to continue the conversations asking for more information about the necessities of the user.

Figure 2: Example of introductory conversation with the chatbot

In figure 2 the example shows some "small talk" capacity of the system. The figure 3 shows an example in which the user asks for a more complex question, such as how to combine two articles. The return from the chatbot is a suggestion of a "Look", which comprises with the available information in the user interaction and also with the personal information about the user, in situations where the user is identified. The question written by the user, in Portuguese, is "Como posso combinar

saia com bolsa de couro?" (or in english: "How can I combine skirt with a leather bag?"). The answer of the chatboot is the phrase "How about this look?" and the indication of a Look that integrated the products mentioned.

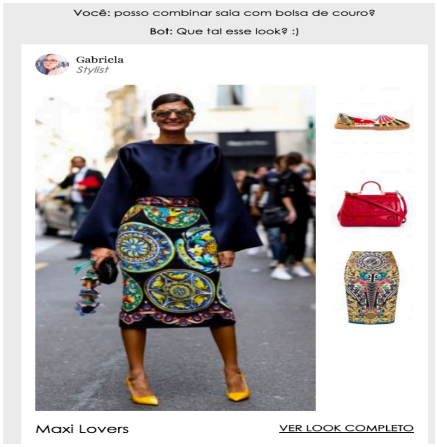


Figure 3: Look suggestion from a broad question

Our proposal for the Demo Session at PROPOR 2016 is to make available a terminal in which the users can experiment the system and to explore the available possibilities. Some worth mention aspects are the semantic support and the integration with machine learning and recommendation algorithms. Therefore the obtained results are differentiated from other known chatbots, resulting in a more robust approach, capable of dealing in a more effective way with the challenges of a natural language dialog.

LetsRead – Tool to Automatically Evaluate Children's Reading Aloud Performance

Jorge Proença[1,2], Dirce Celorico[1], Carla Lopes[1,3], Sara Candeias[4], Fernando Perdigão[1,2]

[1] Instituto de Telecomunicações, Coimbra, Portugal
{jproenca,direcelorico,calopes,fp}@co.it.pt
[2] Department of Electrical and Computer Engineering, University of Coimbra, Portugal
[3] Polytechninc Institute of Leiria, Leiria, Portugal
[4] Microsoft Language Development Centre, Lisbon, Portugal
t-sacand@microsoft.com

Abstract. This demo presents a web-based platform that analyzes speech of read utterances of children, aged 6-10 years old, from the 1st to 4th grades, to automatically evaluate their reading aloud performance. It operates by detecting and analyzing errors and disfluencies in speech. It provides some metrics that are used for computing a reading ability index and shows how close it is to the index given by expert evaluators for that child. Although this demo is not targeted to the participation of children, as pre-recorded utterances are used, the same methods will be applied to live reading tasks with microphone input. A fully developed application will be useful in aiding and complementing the current manual and subjective methods for evaluation of overall reading ability in schools.

Keywords: Reading Aloud Performance, Child Speech, Reading Disfluencies

1 Background

The LetsRead project [1] aims to develop a technological solution to automatically evaluate the reading performance of European Portuguese (EP) primary school children. It could become an important alternative or complement to the usual 1-on-1 evaluations done by teachers or tutors. The automatic evaluation can be performed through the completion of several reading tasks by the child and a live analysis of the utterances to extract performance metrics. Through these metrics, an overall reading ability index can be computed, that should be well correlated with the opinion of teachers [2], [3]. A final application would display sentences to be read by the child and take live microphone audio. The presented demonstration uses pre-recorded utterances instead, as an alternative to microphone input, but employs live server-side processing as well.

For this project, a corpus of young children reading aloud was collected. Children that attend primary school (1st cycle), aged 6 to 10 years old, were asked to read aloud a set of 20 sentences and 10 pseudowords (nonsense/nonexistent words). Further details on the annotation, disfluencies and state-of-the-art of this subject can be consulted in [4] and [5].

2 Demo Interface

At the client-side, this web demonstration allows a grade (1st-4th) and child from the LetsRead dataset to be selected and utterances of this child will be sequentially shown and played. In return, the system at the server-side computes and returns a set of important performance metrics. The interface is exemplified in Figure 1.

Fig. 1. Sample screen of the LetsRead demo web application.

After selecting the grade and child, the current sentence is presented in large letters, simulating an application where a child would have to read it live. The audio signal is presented bellow and played, also being processed by the server to extract performance measures. For this demo only, the results of the analysis are promptly presented by showing correctly and incorrectly pronounced words as well as extra content directly on the audio signal with different colored boxes. Also, the computed performance metrics are shown for the current sentence as well as the average values for the child, given that several sentences have already been sequentially processed. A final application may not necessarily show these results to the children, but only save them to teachers or tutors.

The overall reading ability index, computed from several time-based and pronunciation parameters, is shown in the bar in blue. Through a targeted crowdsourcing effort, information from a panel of experts (primary school teachers) regarding the reading ability of the children was gathered, resulting in a ground truth for scores. The mean

and standard deviation of this parameter is also presented in red and it can be seen how close the automatic index falls to it.

3 System Overview

The techniques employed at the server side take an utterance and detect correctly pronounced words and extra content in order to compute several related metrics such as reading speed and silence duration. Two techniques with similar goals are explored: alignment with word-level lattices detecting repetitions and false-starts [6] and forced alignment allowing optional silence and garbage of speech and noise. Both use phoneme posterior probabilities from a trained phonetic recognizer neural network to compute the likelihood of a word being correctly pronounced or not. The reading ability index is computed with a regression model which was trained with the ground truth scores given by teachers.

4 Future work

The application both at the client and server sides will keep being improved for the foreseeable future, with improved methods to analyze all types of reading disfluencies. The next steps are to acquire child speech in real time using a recording module and make the platform available online. A final application would also include management components for teachers, linked to the students they accompany.

References

1. "The LetsRead Project - Automatic assessment of reading ability of children." [Online]. Available: http://lsi.co.it.pt/spl/projects_letsread.html. [Accessed: 25-Mar-2016].
2. J. Duchateau, L. Cleuren, H. V. hamme, and P. Ghesquière, "Automatic assessment of children's reading level.," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 1210–1213.
3. M. P. Black, J. Tepperman, and S. S. Narayanan, "Automatic Prediction of Children's Reading Ability for High-Level Literacy Assessment," *Trans. Audio, Speech and Lang. Proc.*, vol. 19, no. 4, pp. 1015–1028, May 2011.
4. J. Proenca, D. Celorico, S. Candeias, C. Lopes, and F. Perdigão, "The LetsRead Corpus of Portuguese Children Reading Aloud for Performance Evaluation," in *Proc of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, Portorož, Slovenia, 2016.
5. J. Proença, D. Celorico, C. Lopes, M. Sales Dias, M. Tjalve, A. Stolcke, S. Candeias, and F. Perdigão, "Design and Analysis of a Database to Evaluate Children's Reading Aloud Performance," in *International Conf. on Computational Processing of Portuguese - PROPOR*, Tomar, Portugal, 2016.
6. J. Proença, D. Celorico, S. Candeias, C. Lopes, and F. Perdigão, "Children's Reading Aloud Performance: a Database and Automatic Detection of Disfluencies," in *ISCA - Conf. of the International Speech Communication Association - INTERSPEECH*, Dresden, Germany, 2015, pp. 1655–1659.

NILC-WISE: An Easy-to-use Web Interface for Summary Evaluation with the ROUGE Metric

Fernando Antônio Asevedo Nóbrega and Thiago Alexandre Salgueiro Pardo

Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
13560-970 – São Carlos-SP, Brazil
{fasevedo,taspardo}@icmc.usp.br
www.nilc.icmc.usp.br

Abstract. NILC-WISE is an easy-to-use web application for summary evaluation that provides resources and tools to evaluate summaries written in Portuguese, using the ROUGE metric. Its purpose is to be a default experiment environment, contributing to the summarization area.

1 Introduction

Automatic Summarization (AS) aims to produce a condensed version (or a summary) with the most important content from one or more related texts [7]. Usually, summaries produced by automatic systems are evaluated with the ROUGE framework [6], which analyses the n-gram overlapping among automatic and reference/model summaries.

The ROUGE framework is available for download in its official webpage[1]. However, this system is focused on English language and most of its resources and tools for text normalization are specific for this language. Given the investigations in summarization for the Portuguese language, as [5, 10, 11, 1, 3, 8, 2, 9], we consider that it is very important to have a standard and easy-to-use environment of ROUGE for evaluation of summaries in Portuguese.

It is important to say that it is possible to use the official ROUGE package for Portuguese if the user avoids some ROUGE parameters or if some internal resources in the system are changed. However, even if the user correctly sets the environment, the use of different resources and/or tools may produce different evaluation results (e.g., the use of distinct text normalization processes, as the algorithm of stemming and removal of stopwords). Such issues may harm the validity of comparative evaluations and be a problem for advancing the knowledge frontier in the area.

In this paper, we introduce NILC-WISE (Web Interface for Summary Evaluation developed at NILC), which is an easy-to-use web interface for applying ROUGE for summary evaluation for Portuguese language, aiming at dealing with the above issues and allowing for more systematic and uniform evaluations.

[1] http://www.berouge.com

2 The NILC-WISE interface

When accessing for the first time the NILC-WISE interface (see Fig. 1), the user must create a personal account[2] (providing e-mail, password and institution information). After performing login, the user may access any previous experiments he has eventually carried out and the available datasets on NILC-WISE, as well as to evaluate summaries based on a set of reference summaries. At the moment, as reference summaries, we have some different settings of the CST-News corpus [4]. However, in the future, we want to add other Portuguese and English datasets in the tool.

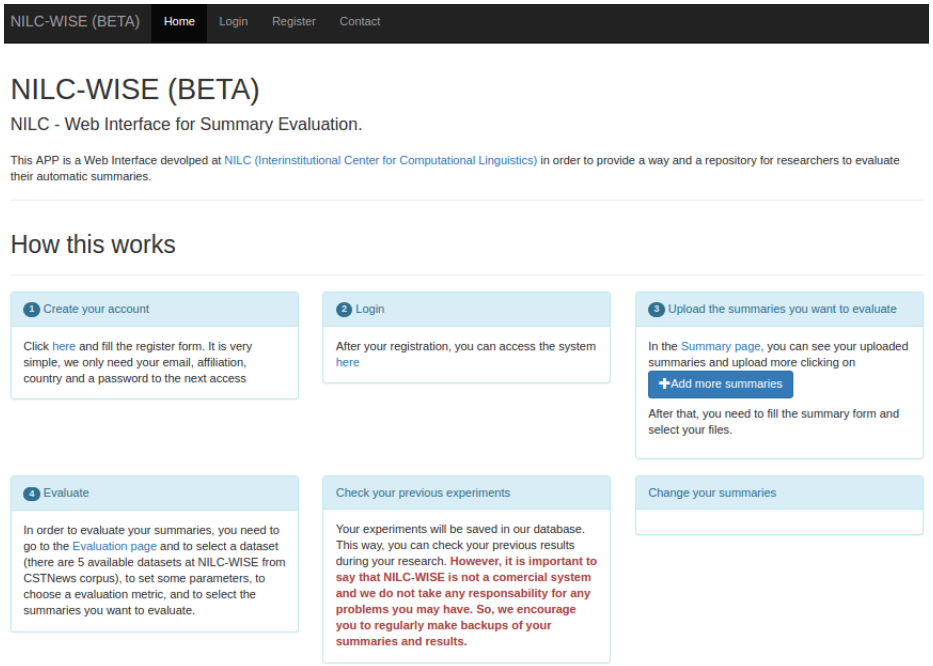


Fig. 1. NILC-WISE interface

Before performing the evaluation of his summaries, the user must create a summary directory in NILC-WISE and indicate a title for it, the filename pattern of the summary files[3] and the summary language (e.g., Portuguese or English). In this summary directory, the user may then upload the summary files (eg.: the summaries that were produced by the AS system of the user) to NILC-WISE.

During the evaluation process, the user must pick one available dataset on NILC-WISE, to select one of his summary directories and to configure the pa-

[2] We use this option in order to contact the researchers if it is necessary.
[3] This information is required for ROUGE.

rameters for text normalization and for ROUGE. For normalization, NILC-WISE uses a list of stopwords for Portuguese developed at NILC[4] and a stemming algorithm provided by the NLTK[5] package for Portuguese.

Acknowledgments

The authors are grateful to CAPES and FAPESP for supporting this work.

References

1. Camargo, R.T., Agostini, V., Di Felippo, A., Pardo, T.A.: Manual typification of source texts and multi-document summariesalignments. Procedia - Social and Behavioral Sciences **95**(0) (2013) 498 – 506
2. Cardoso, P., Pardo, T.A.S.: Multi-document summarization using semantic discourse models
3. Cardoso, P., Pardo, T.A.S.: Joint semantic discourse models for automatic multi-document summarization. In: Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology - STIL, Natal, RN, Brazil (2015) 81–90
4. Cardoso, P.C.F., Maziero, E.G., Castro Jorge, M.L.R., Seno, E.M.R., Di Felippo, A., Rino, L.H.M., Nunes, M.d.G.V., Pardo, T.A.S.: CSTNews – a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In: Anais do III Workshop "A RST e os Estudos do Texto", Cuiabá, MT, Brasil, Sociedade Brasileira de Computação (2011) 88–105
5. Castro Jorge, M.L., Pardo, T.A.S.: A generative approach for multi-document summarization using the noisy channel model. In: Proceedings of the 3rd RST Brazilian Meeting, Cuiabá/MT, Brazil (2011) 75–87
6. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. (2004) 74–81
7. Mani, I.: Automatic Summarization. Volume 3. John Benjamins Publishing Company (2001)
8. Muller, E., Granatyr, J., Lessing, O.: Comparativo entre o algoritmo de luhn e o algoritmo gistsumm para sumarização de documentos. Revista de Informática Teórica e Aplicada **22**(1) (75–94) 584–599
9. Nóbrega, F.A.A., Pardo, T.A.S.: Improving content selection for update summarization with subtopic-enriched sentence ranking functions
10. Ribaldo, R., Akabane, A.T., MachadoRino, L.H., Pardo, T.A.S.: Graph–based methods for multi-document summarization: Exploringrelationship maps, complex networks and discourse information. In: Proceedings of the 10th International Conference on ComputationalProcessing of Portuguese (LNAI 7243), Coimbra, Portugal (2012) 260–271
11. Silveira, S., Branco, A.H.: Enhancing multi-document summaries with sentence simplification. In: Proceedings of the 14th International Conference on Artificial Intelligence. (2012) 742–748

[4] http://www.nilc.icmc.usp.br/
[5] http://www.nltk.org/

Demo: OpenWordnet-PT

Fabricio Chalub[1], Livy Real[1], Valeria de Paiva[2], and Alexandre Rademaker[1,3]

[1] IBM Research Brazil
[2] Nuance Communications
[3] FGV-EMap

Abstract. This demo introduces OpenWordnet-PT (OWN-PT), an open wordnet for Portuguese. We will explore its web interface which offers an easy way for regular Internet users to utilise it. We also give a quick introduction to how to use RDF and SPARQL in the context of the OWN-PT.

1 Introduction

Wordnets are structured lexicons, usually implemented as databases of words linked by their semantic relationships. Their graph structure makes it possible to have automated processes for acquiring and using linguistic knowledge. Largely used by linguists and computer scientists, wordnets have several applications in Natural Language Processing (NLP), such as word sense disambiguation, synonyms finding and lexical simplification. This demonstration introduces OpenWordnet-PT (OWN-PT) [3], our open wordnet for Portuguese. Following Linked Open Data principles, OWN-PT is available in RDF/OWL. Both the data and the RDF template settings of the OpenWN-PT are freely available for download and the data can be retrieved via SPARQL at the endpoint[4].

The main point of this demo is to show how to use the OWN-PT web interface that offers a quick and easy way to search words and to suggest modifications to the database, such as adding/removing words, examples, glosses and adding comments to the entries. This interface [6] allows regular Internet users to interact with the OWN-PT content. However, an user able to pose queries in RDF and SPARQL will have a better understanding and will get more information out of OWN-PT. Many other open resources, such as DBpedia and GeoNames, are available in RDF. Thus the second point of this demo is to offer to the PROPOR community an introduction on how to use RDF and SPARQL in the context of the OWN-PT.

2 Wordnets and OpenWordnet-PT

The basic entry of a wordnet is a synset, a synonym set composed by one or more word forms that have a similar sense; a gloss, a small definition of the synset

[4] `http://wnpt.brlcloud.com/wn/`

sense; a numerical identity ID, that differentiates the senses; and (many of the synsets) also have examples of senses in context. The synsets are hierarchically structured through many semantic relations. Nouns, for example, are structured via hyperonym, hyponym, meronym and antonym relations.

The OWN-PT interface has several features that allow you to discover facts about the Portuguese language and also to contribute to the improvement of the lexical resource. Its search mechanism permits searching synsets by word forms and filter results (via facets) by different fields. For example, the user can filter the results by a specific part-of-speech tag; by using the lexicographer files (that group synsets by part-of-speech and semantic class. For example, the nouns that describe people or verbs that represent actions) and by the number of word forms each synsets has, both in English and in Portuguese. The search mechanism also allows one to use regular expressions when searching within the glosses, word forms or example fields, which highly increases the searcher's effectiveness.

Other important aspect of the OWN-PT interface is its collaborative way to correct and improve the actual state-of-the-art of the resource. Empty places or forms to suggest new words, glosses or examples are provided and one could insert a new suggestion with a single click. Those suggestions are displayed in the synsets to be voted for other users. The voting mechanism allows trusted users to vote for their desired modifications. Finally each synset has also a comment box, where it is possible to leave comments, to report bugs, to connect together a group of synsets or to draw the attention of other users. For this, the interface offers a mention device (@) and a hashtag device (#), inspired by Twitter.

Besides all the content that the OWN-PT web interface offers to end users, a richer structure and much more information on synsets and their relations can be obtained using the SPARQL endpoint that queries the RDF database. This demo will show the audience how to navigate through the OWN-PT RDF and understand its structure and how the OWN-PT is linked with other resources, such as OpenMultilingual Wordnet (OMW) [1] and the SUMO Ontology [5].

Most importantly, we will show a sizeable number of applications of OWN-PT to corpora[5]. Those prototypes are produced using the FreeLing 4.0 suite [2] for example, to analyse the Dicionário Histórico-bibliográfico Brasileiro[6], a dictionary of biographies of Brazilian politicians produced by historians at Fundação Getúlio Vargas [4].

This demo shows what the open and freely modifiable resource OWN-PT is and we hope that the community will be able to employ it for their own purposes, whether they are niche linguistics fieldwork or large commercial applications. Several NLP applications can be improved using such a resource, and while OWN-PT has been used by some big players in the international community (such as IBM, Google Translate, OpenMultiLingual WordNet and BabelNet), the Lusophone community has not, yet, paid much attention to it.

[5] `http://wnpt.brlcloud.com/wn/prototypes`
[6] `http://cpdoc.fgv.br/acervo/dhbb`

Bibliography

[1] Francis Bond and Kyonghee Paik. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue, 2012. 64–71.

[2] Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. Freeling: An open-source suite of language analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, 2004.

[3] Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. Openwordnet-pt: An open Brazilian Wordnet for reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India, dec 2012. The COLING 2012 Organizing Committee. Published also as Techreport http://hdl.handle.net/10438/10274.

[4] Valeria De Paiva, Dário Oliveira, Suemi Higuchi, Alexandre Rademaker, and Gerard De Melo. Exploratory information extraction from a historical dictionary. In *IEEE 10th International Conference on e-Science (e-Science)*, volume 2, pages 11–18. IEEE, oct 2014.

[5] Adam Pease. *Ontology: a practical guide*. Articulate Software Press, 2011.

[6] Livy Real, Fabricio Chalub, Valeria de Paiva, Claudia Freitas, and Alexandre Rademaker. Seeing is correcting: curating lexical resources using social interfaces. In *Proceedings of 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference on Natural Language Processing of Asian Federation of Natural Language Processing - Fourth Workshop on Linked Data in Linguistics: Resources and Applications (LDL 2015)*, Beijing, China, jul 2015.

Poe, now you can TryMe: Interacting with a Poetry Generation System

Hugo Gonçalo Oliveira
CISUC, Dept. of Informatics Engineering, University of Coimbra, Portugal
hroliv@dei.uc.pt

Abstract. This demo presents two ways of interacting with PoeTryMe, a poetry generation system. The TryMe web interface communicates with a REST API for a simpler version of PoeTryMe that enables the generation of poems according to four parameters. A Portuguese instantiation of PoeTryMe is also continuously running as a Twitterbot.

Keywords: poetry generation, linguistic creativity, creative service

1 Introduction

Poetry generation is a popular task among the Computational Creativity [1] community. It is a kind of natural language generation where, besides following grammatical rules and evoking a meaning, produced text exhibits features such as a regular metre or rhymes.

PoeTryMe [2] is a poetry generation platform, originally developed for rendering the contents of Portuguese lexical-semantic networks extracted from text (e.g. [3]). Yet, given the flexibility of its modular architecture, it was later adapted to produce poetry of different kinds and in different languages [4, 5]. Since early 2016, PoeTryMe can be used through a web interface for its recently developed REST API, where users can try poem generation with a subset of available configurations. Its Portuguese version has also been running in the social network Twitter, as the user *@poetartificial*, where it regularly posts a poem inspired by current trends.

2 Architecture and Linguistic Resources

PoeTryMe has a modular architecture where different modules handle relevant aspects of its approach for poetry generation. Its adaptation is thus a matter of reimplementing some of the modules or changing the linguistic resources underlying them. Briefly, there are two core modules, one for producing natural language fragments and another for organising them into poetic forms. Text fragments are generated with the help of a semantic network – with relation instances represented as $triplets = (word_1, predicate, word_2)$ – and a generation grammar – with line templates for different relation types. Other modules include a morphology lexicon for inflecting nouns and adjectives; a tool for syllable-related operations; and a module for explaining the semantic relations between

the words of each produced line and seed words. Generation starts with a set of (seed) words that constrain the semantic network. Relevant relation instances are then rendered by a template for their type, after filling specific placeholders with the relation arguments. In a generate & test strategy, several lines are produced this way, but only a few are used in the poem, given their length and rhymes. A more detailed introduction to its architecture is found elsewhere [2].

The Portuguese instantiation of PoeTryMe uses CARTÃO as the semantic network [3], also used in the automatic extraction of a grammar from a collection of Portuguese poetry, LABEL-Lex[1] as the morphology lexicon, and SilabasPT[2] for splitting words into syllables and identifying the stress.

3 TryMe web interface

TryMe is a very simple web interface for interacting with a simpler version of PoeTryMe. It still enables the generation of a poem, given four parameters: (i) language (Portuguese, Spanish or English); (ii) poetry form, including poetic forms (e.g. block of four 10-syllable lines, sonnet) and the rhythm of well-known songs; (iii) seed words (open set); (iv) surprise factor, which sets the probability of using words that are undirectly connected to the seeds. Figure 1 shows this simple interface, accessible from `http://poetryme.dei.uc.pt/`, and an illustrative poem – a block of 4, generated in Portuguese, using the seeds *processar* and *português*, with a surprise factor of 0.001.

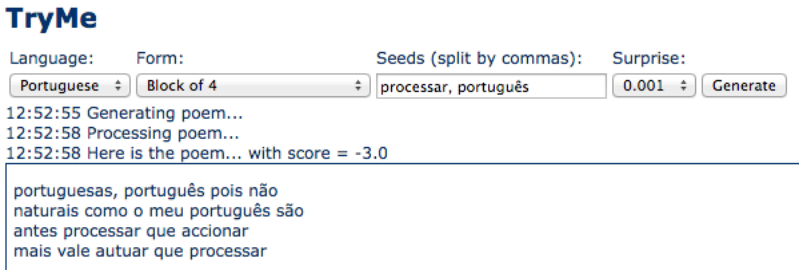


Fig. 1. The TryMe web interface

4 REST API

The TryMe interface interacts with PoeTryMe through a REST API, which provides the same features. The API endpoint is located at `http://concreteflows.ijs.si/poetry/rest/poetry` and generation requests can set

[1] `http://label.ist.utl.pt/en/labellex_en.php`
[2] `https://code.google.com/archive/p/silabaspt`

the language (`lang`), the form (`form`), the seeds (`seeds`) and the surprise factor (`surp`). Figure 2 illustrates the operation of the API with a request – block of two 10-syllable lines, in Portuguese, with seeds *poesia* and *portuguesa*, and a surprise factor of 0.0005 – and its response, in JSON.

- Request:
  `http://concreteflows.ijs.si/poetry/rest/poetry?lang=pt&form=10-2&seeds=poesia+portuguesa&surp=0.0005`
- Response:
  `{"form":"dueto.est", "language":"pt", "score":"-1.0", "seeds":["portuguesa","poesia"], "surprise":"5.0E-4", "text":"sem achar poesia, nem musa\né coisa portuguesa e lusa"}`

Fig. 2. Example of a request and a response of PoeTryMe's web API.

5 Twitterbot

The limits of a Portuguese instantiation of PoeTryMe are continuously being tested in a bot, currently running on the social network Twitter (`https://twitter.com/poetartificial`). From time to time, the bot selects one of the top trends for Portugal, analyses the content of the last tweets using the trend, and uses the most frequent words as seeds for PoeTryMe. The resulting poem is then posted.

Acknowledgements This work was supported by ConCreTe. The project ConCreTe acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 611733.

References

1. Colton, S., Wiggins, G.A.: Computational creativity: The final frontier? In: Procs of 20th European Conference on Artificial Intelligence (ECAI 2012), Montpellier, France, IOS Press (2012) 21–26
2. Gonçalo Oliveira, H., Cardoso, A.: Poetry generation with PoeTryMe. In Besold, T.R., Schorlemmer, M., Smaill, A., eds.: Computational Creativity Research: Towards Creative Machines. Atlantis Thinking Machines. Atlantis-Springer (2015) 243–266
3. Gonçalo Oliveira, H.: On the utility of Portuguese term-based lexical-semantic networks. In: Procs of 11th Intl. Conference on Computational Processing of the Portuguese Language (PROPOR 2014). Volume 8775 of LNCS., São Carlos, SP, Brazil, Springer (October 2014) 176–182
4. Gonçalo Oliveira, H.: Tra-la-lyrics 2.0: Automatic generation of song lyrics on a semantic domain. Journal of Artificial General Intelligence **6**(1) (December 2015) 87–110 Special Issue: Computational Creativity, Concept Invention, and General Intelligence.
5. Gonçalo Oliveira, H., Hervás, R., Díaz, A., Gervás, P.: Adapting a generic platform for poetry generation to produce Spanish poems. In: Procs of 5th Intl. Conference on Computational Creativity. ICCC 2014, Ljubljana, Slovenia (June 2014)

Syntax Deep Explorer

José Correia[1,2], Jorge Baptista[2,3], and Nuno Mamede[1,2]

[1] Instituto Superior Técnico, Universidade de Lisboa
[2] L^{2}F – Spoken Language Lab, INESC-ID Lisboa
Rua Alves Redol 9, P-1000-029 Lisboa, Portugal
jcorreia,Nuno.Mamede@inesc-id.pt
[3] Universidade do Algarve
Campus de Gambelas, P-8005-139 Faro, Portugal
jbaptis@ualg.pt

Abstract. *Syntax Deep Explorer* is a new tool that uses several association measures to quantify several co-occurrence types, defined on the syntactic dependencies (*e.g.* subject, complement, modifier) between a target word *lemma* and its co-locates. The resulting co-occurrence statistics is represented in *lex-grams*, that is, a synopsis of the syntactically-based co-occurrence patterns of a word distribution within a given *corpus*. These *lex-grams* are obtained from a large-sized Portuguese *corpus* processed by STRING and are presented in a user-friendly way through a graphical interface. The *Syntax Deep Explorer* will allow the development of finer lexical resources and the improvement of STRING processing in general, as well as providing public access to dependency-based, co-occurrence information derived from parsed *corpora*.

Keywords: Natural Language Processing (NLP), co-occurrence, collocation, association measures, graphic interface, lex-gram, Portuguese

The analysis of the co-occurrence patterns between words in texts shows the differences in use (and meaning), which are often associated with the different grammatical relations in which a word can participate [2,5]. The quantification of these patterns is a powerful tool in modern lexicography as well as in the construction of basic linguistic resources, like *thesauri*. The stakeholders on the study of those co-occurrence patterns are linguists, language students, translators, lexicographers and Corpus Linguistics' researchers, who study the behaviour of linguistic expressions in *corpora*. For all of these, co-occurrence data are essential for a better understanding of language use. Furthermore, comparing different association measures, each capturing different linguistic aspects of the co-location phenomena, enables the user to achieve a broader understanding of the distribution of a given word, hence its different meanings and uses. For a better analysis of co-occurrence patterns in a *corpus*, it is also important to provide the user with an interface that helps him/her to explore the patterns thus extracted, namely, by accessing concordances or moving on from an initial query to another interesting collocate.

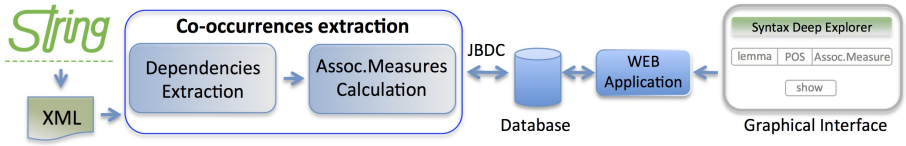


Fig. 1. *Syntax Deep Explorer* architecture.

We present *Syntax Deep Explorer*[4] (Fig. 1), a tool that allows the general public to explore syntactically-based collocations from Portuguese *corpora*, using a broad set of association measures, in view of an enhanced understanding of words' meaning and use [3]. The information on these collocation patterns is made available through a web interface by way of *lex-grams* (Fig. 2), a synopsis of the syntactically-based co-occurrence patterns of a word's distribution within a given *corpus*. This mode of presentation organizes the information for a simpler analysis by users. For now, only the four main lexical part-of-speech categories (adjectives, nouns, verbs and adverbs) are targeted by the *Deep Syntax Exporer*.

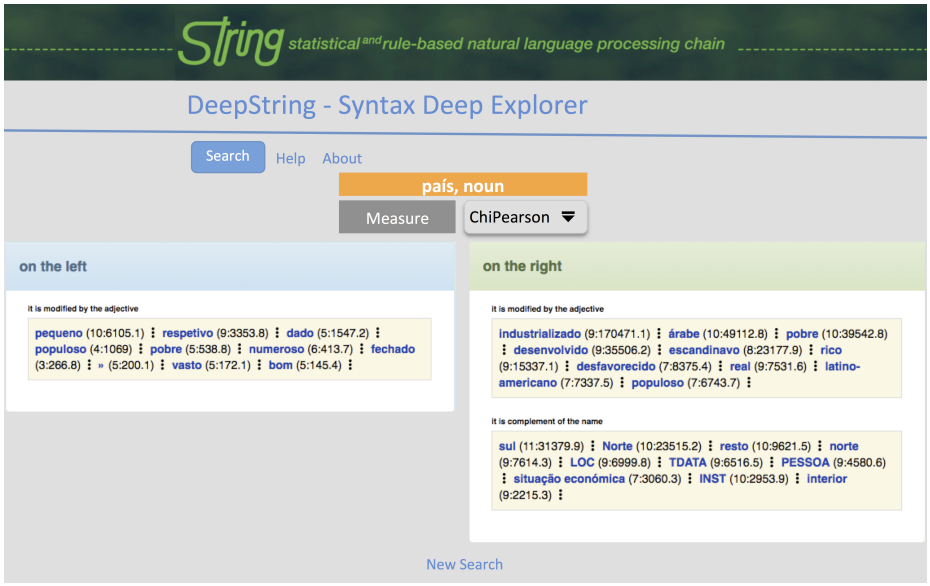


Fig. 2. *Lex-gram* of the noun `país` ordered by the Chi-Pearson measure.

The *Syntax Deep Explorer* is based on STRING [4], a hybrid, statistical and rule-based, natural language processing (NLP) chain for the Portuguese language, developed by Spoken Language Systems Lab (L^2F) from INESC-ID Lisboa[5]. STRING has a modular structure and performs all basic text NLP tasks: text segmentation (sentence splitting and tokenization) and part-of-speech (POS) tagging; rule-based and statistical POS disambiguation; parsing, named entity recognition, anaphora resolution, and temporal expressions normalization. The processing produces a XML output file.

[4] `string.l2f.inesc-id.pt/demo/deepExplorer` (last visit 31/05/2016).
[5] `www.l2f.inesc-id.pt` (last visit 31/05/2016).

The *Syntax Deep Explorer* uses the following (higher level) syntactic dependencies, produced by XIP [1], the STRING parsing module: (1) the `SUBJ` dependency, which links a verb and its subject; (2) the `CDIR` dependency, linking a verb and its direct complement; (3) the `CINDIR` dependency, which links the verb with a dative essencial complement; (4) the `COMPL` dependency, which links a predicate (verb, noun or adjective) to its essential `PP` complements; and (5) the `MOD` dependency that links a modifier with the element it modifies (*e.g.* adjetives as modifiers of nouns, or adverbs as modifiers of verbs); and (6) Named Entities, which also captured by the XIP parser by an unary `NE` dependency that delimits and assigns them to several general categories (`PERSON`, `ORGANIZATION`, `PLACE`, etc.); these categories are then used for co-occurrence statistics, instead of the individual entities themselves.

The *Syntax Deep Explorer* takes advantage of the rich lexical resources of STRING, as well as of its sophisticated syntactic and semantic analysis, and finds the syntactically-based co-occurrences patterns of a given word *lemma* storing that information in a database. Then, the tool calculates different association measures, producing a *lex-gram* with the co-occurrence statistical information, a snapshot representing the main distributional patterns of a given word. The *lex-gram* also makes it possible to move from any given pattern to another co-locate of interest found within. Furthermore, STRING rich lexical resources feature a large number of multiword expressions, which are processed in the same way as any simple (single-word) unit. Thus, it is also possible to analyse multiwords' distribution and their co-occurrence patterns. Results from evaluation show that the runtime of the extraction tool remains constant throughout the *corpus*, while the size of the database does not grow linearly, indicating that information is not being repeated. The web application response time also allows fast queries to the database. In the future, it is necessary to increase the number of *corpora* present in the database and to allow the comparison of different *lex-gram* for the same *lemma* across *corpora*. The automatic creation of *thesauri* from the stored distributional information is also envisaged.

Acknowledgment This work was supported by national funds through FCT–Fundação para a Ciência e a Tecnologia, ref.UID/CEC/50021/2013.

References

1. Aıt-Mokhtar, S., Chanod, J.P., Roux, C.: Robustness beyond shallowness: Incremental deep parsing. Natural Language Engineering 8, 121–144 (2002)
2. Church, K., Hanks, P.: Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics 16(1), 22–29 (1990)
3. Correia, J.: Syntax Deep Explorer. Master's thesis, Instituto Superior Técnico - Universidade de Lisboa, Lisboa, Portugal (September 2015)
4. Mamede, N., Baptista, J., Diniz, C., Cabarrão, V.: STRING: An Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. In: PROPOR 2012. vol. Demo Session (April 2012)
5. Sinclair, J.: Corpus, concordance, collocation. Oxford University Press (1991)

VITHEA-Kids: Improving the Linguistic Skills of Children with Autism Spectrum Disorder

Vânia Mendonça, Cláudia Filipe, Luísa Coheur, Alberto Sardinha
{vania.mendonca,claudia.patricia,luisa.coheur,jose.alberto.sardinha}@tecnico.ulisboa.pt

Instituto Superior Técnico, Porto Salvo, Portugal

Abstract. Each child with Autism Spectrum Disorder (ASD) has a unique set of abilities, symptoms and needs; hence, educational applications should allow to tailor exercises' content and options. However, most existing applications do not take this requirement into account. In this work, we present VITHEA-Kids, a platform that takes advantage of language and speech technologies tools to allow children with ASD to benefit from a customized learning experience.

Keywords: Computer-assisted language learning, Autism Spectrum Disorder

1 Introduction

Autism Spectrum Disorder (ASD) is characterized by persistent deficits in social communication and interaction, as well as restricted, repetitive behaviors or interests since an early age. It also often comprises difficulties in communication [2], but the challenges faced strongly vary across individuals. Given the interest that individuals with ASD display towards computers [6], educational applications could be useful to teach them new skills – an hypothesis experimented by several authors [7], which, along with the increasing popularity of mobile devices, has led to a great variety of applications targeting children with impairments (for a detailed review of related studies and applications, please refer to Mendonça [4]). However, there is a lack of applications in Portuguese that take into account each child's needs. In this context, we present VITHEA-Kids: a software platform in Portuguese where children with ASD can solve exercises to develop linguistic skills, having their needs accounted for. This platform makes use of in-house language and speech technologies to provide a customized experience.

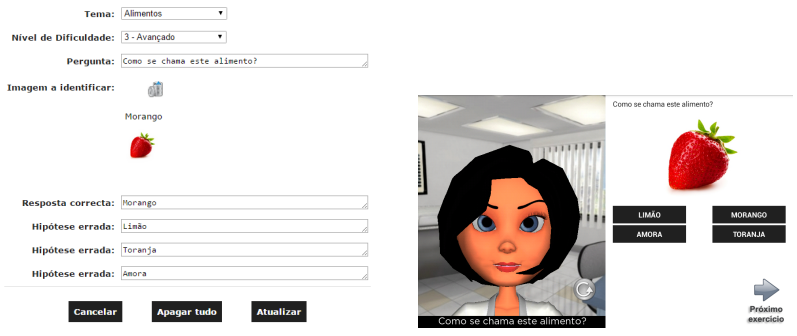
2 VITHEA-Kids

VITHEA-Kids is a platform where children can solve exercises created by their caregivers. It was build using the infrastructure of an in-house award-winning platform: Virtual Therapist for Aphasia Treatment (VITHEA) [1], in which patients with aphasia could solve oral word naming exercises prepared by their

therapists and presented by a talking animated character using a Text-To-Speech (TTS) synthesizer [5].

VITHEA-Kids allows to create multiple choice exercises, often used for children with ASD, allowing to work on skills such as vocabulary acquisition, word-picture association, and generalization. Each exercise is composed of a question (e.g, "What is the name of this object?"), an optional stimulus (e.g, the picture of a fork), and a set of possible answers, respectively ("Fork", "Spoon", "Cup", "Bowl"), in which only one of the answers is correct. Wrong answers can go from zero to three, allowing to create different exercises with small variations.

VITHEA-Kids is composed of two applications, which will be live demonstrated during the conference (considering the availability of Internet connection). The caregiver's application allows to create and manage the exercises described above (see Figure 1a), group them in ordered sequences associated with a set of children, upload and manage image files to use in the exercises, and manage the users of both applications. It also allows for the caregiver to customize the child's application, namely the messages uttered by the animated character, through a TTS synthesizer [5], in certain situations (e.g, when the child logs in, or when they select a correct answer) and the images to display when the child correctly solves an exercise (as a way of reinforcing the correct choice). As for the child's application, it presents the list of sequences associated with the child logged in. Upon choosing a sequence, each exercise is presented in the order defined by the caregiver. The exercise's question is uttered by the animated character, and it is also displayed on the screen, along with the stimulus (when existing) and the possible answers in a random order (see Figure 1b). Tapping over the correct answer will lead to a reinforcement image and a customized feedback message uttered by the animated character. Selecting any other answer will activate a set of helping cues to prompt the child to select the correct answer: the selected answers disappears, the correct answer is highlighted and the remaining answers are uttered by the animated character. When the exercise session ends, information about child's performance is shown.



(a) Caregiver's application (b) Child's application

Fig. 1: VITHEA-Kids's applications

To ease the caregiver's task of creating new exercises, we also developed a module for the generation of multiple choice exercises based on a question template and a topic [4], although it is not integrated with VITHEA-Kids yet.

3 Conclusions and Future Work

In this work, we took the first steps into addressing the issues presented by the currently available software for individuals with ASD. Our platform makes use of language and speech technologies to support the development of Portuguese linguistic skills, allowing caregivers to create content and perform several customizations according to each child's needs. These applications were evaluated with several caregivers and also with a child, as detailed in Mendonça [4]. Currently, we are performing several improvements to the platform, as well as extending the possibilities of customization and the variety of exercises (including ones that could use an Automatic Speech Recognition (ASR) module [3] to validate oral answers).

Acknowledgments. This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013, and under project CMUP-ERI/HCI/0051/2013.

References

1. A. Abad, A. Pompili, A. Costa, I. Trancoso, J. Fonseca, G. Leal, L. Farrajota, and I. P. Martins. Automatic word naming recognition for an on-line aphasia treatment system. *Computer Speech & Language*, 27(6):1235–1248, Sept. 2013.
2. American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5*. American Psychiatric Association Arlington, VA, 5th edition, 2013.
3. H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso. *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003 Faro, Portugal, June 26–27, 2003 Proceedings*, chapter AUDIMUS.MEDIA: A Broadcast News Speech Recognition System for the European Portuguese Language, pages 9–17. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.
4. V. Mendonça. Extending VITHEA in order to improve children's linguistic skills. Master's thesis, Instituto Superior Técnico, 2015.
5. S. Paulo, L. C. Oliveira, C. Mendes, L. Figueira, R. Cassaca, C. Viana, and H. Moniz. Dixi — a generic text-to-speech system for european portuguese. In *Proceedings of the 8th International Conference on Computational Processing of the Portuguese Language – PROPOR '08*, pages 91–100, 2008.
6. C. Putnam and L. Chong. Software and technologies designed for people with autism: What do users want? In *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '08, pages 3–10, New York, NY, USA, 2008. ACM.
7. S. Ramdoss, A. Mulloy, R. Lang, M. O'Reilly, J. Sigafoos, G. Lancioni, R. Didden, and F. El Zein. Use of computer-based interventions to improve literacy skills in students with autism spectrum disorders: A systematic review. *Research in Autism Spectrum Disorders*, 5(4):1306–1318, Oct. 2011.

XCrimes: Information Extractor for the Public Safety and National Defense Areas

Daniel Sullivan[1], Vladia Pinheiro[1], Rafael Pontes[1], Vasco Furtado[1]

[1]Graduate Program in Applied Informatics – University of Fortaleza
Av. Washington Soares, 1321, Fortaleza, Ceará, Brazil

daniel.sullivancl@gmail.com , rafaellpontes@gmail.com, {vasco, vladiacelia}@unifor.br

Abstract. The increased volume of notifications about crimes or attempted crimes is a vast textual material to support public safety policies, and the reading, mining and analysis of all the textual volume of police reports are very time consuming tasks. In this scenario, subsidized by research and innovation project of the Department of Science and Technology of the State of Ceará (SECITECE), we are developing the XCRIMES tool - which allows automatic extraction of information about crimes from textual reports of public safety. XCRIMES uses the semantic knowledge base in Portuguese, InferenceNet, and the Semantic-Inferentialist Analyzer - SIA - to reason about the text and draw conclusions in order to leverage the human expertise, automating the information extraction process about the characteristics of crimes. In this software demonstration, it will be presented how XCrimes extracts information about the type of crime and crime scene.

Keywords: Information Extraction; Semantic Annotation; Information Retrieval.

1 Introduction

Currently, one of the biggest problems of Brazilian society is the lack of public safety. The increase in the number of crimes and the use of tools for online events notification, resulted in increased volume of notifications about crimes or attempted crimes. Therefore, there is vast textual material to support public safety policies, and the reading, mining and analysis of all the textual volume of police reports are very time consuming tasks. In this scenario, subsidized by research and innovation project of the Department of Science and Technology of the State of Ceará (SECITECE), we are developing the XCRIMES tool - which allows automatic extraction of information about crimes from textual reports of public safety, such as: localizations involved (the crime scene, place of residence of victims and suspects etc.); information on the profile of people (police, victims, suspects etc.); used weapons and vehicles involved; type and cause of crime.

2 XCrimes – Architecture and Prototype

XCRIMES uses the semantic knowledge base in Portuguese, InferenceNet [1], and the Semantic-Inferentialist Analyzer - SIA [2] - to reason about the text and draw conclusions in order to leverage the human expertise, automating the information extraction process about the characteristics of crimes [3]. Nowadays, Xcrimes is in the early stage of development, which can already be drawn from the following information: type of crime and crime scene.

Figure 1 presents the general architecture of XCrimes. The first layer (Client System) retrieves the user's input text, selects the necessary contextual data and sends them to the next layer. The layer "Xcrimes PLN" performs: (1) pre-processing of text (sentence detector, *tokenizer*, *lemmatizer*, etc) and Pos Tagging, using the parser Freeling [4]; (2) the SIA performs the word sense disambiguation, semantic annotation text by associating words and expressions to concepts of InferenceNet and Wikipedia bases, and generates a network of inferences, premises and conclusions, of the input text; (3) the Information Extractor module performs the match of the inferences generated by SIA with goals set by the client. The knowledge bases that support PLN activities are included in the layer "XCrimes BD" and "External Databases". In "Xcrimes Interface" layer contains the system interfaces to access to the system, record the objectives of widgets configuration, validation and export of information extracted, and widgets components that are embedded in the information systems of public security departments.

Figure 1. General Architecture of XCrimes

Figures 2 and 3 present an example and the prototype of the XCrimes system. The text "*Marcos Antonio da Silva Santana foi executado, a tiros, na Rua Marilia Dutra, no bairro da Maraponga. O crime ocorreu sábado à tarde. A polícia suspeita de um caso de vingança. O fato agora vai ser apurado pela equipe do 5º DP.*" Is informed by the user that selects the information objectives – crime scene and type of crime. After semantic analysis of the

input text, XCrimes gives the answer that the crime scene was "*Rua Marília Dutra*", the sentence of the text that contains the answer is "*Marcos Antônio da Silva Santana foi executado na Rua Marília Dutra*" e a inference that matched with the objective was is "*Marcos Antônio da Silva Santana é executado em um local*". For the second objective (type of crime), the system checks the inferences and makes a match with the response alternatives (murder, robbery, domestic violence, theft, etc.). The inference that obtains more relevance will be the answer given. In this example, the sentence with the answer is "*Marcos Antonio da Silva Santana was run to death*.", and the answer is that the type of crime is "homicide.

Figure 2. Example of the initial interface of XCrimes.

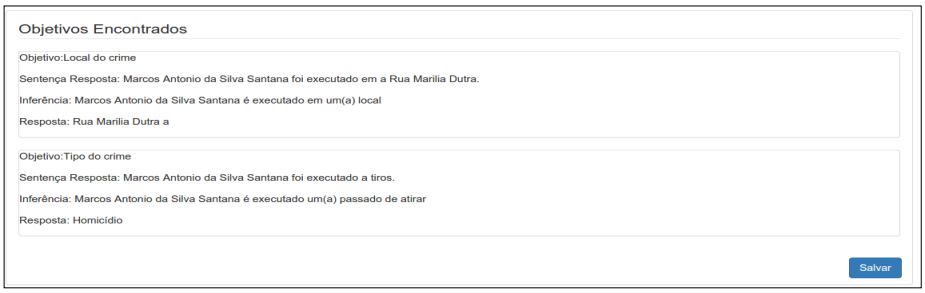


Figure 3. Information about type and crime scene, extracted by XCrimes.

References

1. V. Pinheiro, T. Pequeno, V. Furtado and W. Franco. InferenceNet.Br: Expression of Inferentialist Semantic Content of the Portuguese Language. In: T.A.S. Pardo et al. (eds.): PROPOR 2010, LNAI 6001(90–99). Springer, Heidelberg, 2010.
2. Pinheiro, V., Pequeno, T., Furtado, V. 2010. Um Analisador Semântico Inferencialista de Sentenças em Linguagem Natural. Linguamática. ISSN: 1647-0818. Vol.2. Num.1, pp. 111-130.
3. V. Pinheiro, V. Furtado, T. Pequeno and D. Nogueira. Natural language processing based on semantic inferentialism for extracting crime information from text. In: IEEE international conference on intelligence and security informatics (ISI). IEEE (2010), pp. 19–24.
4. Padró, Lluís e Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality.