Bibliography

[1] **Conference on Computational Natural Language Learning (CoNLL-99)**. http://www.clips.ua.ac.be/conll99, 1999. [Online; accessed May 2nd, 2011].

[2] **Bíblia Florestal: Um Manual Lingüístico da Floresta Sintá(c)tica**. http://www.linguateca.pt/Floresta/BibliaFlorestal, 2008. [Online; accessed May 2nd, 2011].

[3] **OpenNLP Maxent - The Maximum Entropy Framework**. http://maxent.sourceforge.net/about.html, Aug. 2008. [Online; accessed May 2nd, 2011].

[4] **A Snapshot of Facebook in 2010**. http://www.facebook.com/notes/democracy-uk-on-facebook/a-snapshot-of-facebook-in-2010/172769082761603, Dec. 2010. [Online; accessed May 2nd, 2011].

[5] ABNEY, S. Parsing by chunks. **Principle-based parsing**, p. 257–278, 1991.

[6] ALPAYDIN, E. **Introduction to Machine Learning**. 2. ed., The MIT Press, 2010.

[7] BICK, E. **The Parsing System Palavras, Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. Aarhus University Press, 2000.

[8] BRILL, E. **Some Advances in Transformation-Based Part of Speech Tagging**. In: PROCEEDINGS OF THE TWELFTH NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI-94), Seattle, Washington, USA, 1994.

[9] BRILL, E. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. **Computational Linguistics**, v.21, p. 543–565, December 1995.

PUC-Rio - Certificação Digital Nº 0921334/CA

[10] BUCHHOLZ, S.; MARSI, E. **Conll-x shared task on multilingual dependency parsing**. In: PROCEEDINGS OF THE TENTH CONFERENCE ON COMPUTATIONAL NATURAL LANGUAGE LEARNING, CoNLL-X '06, p. 149–164, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[11] BUCHHOLZ, S.; VEENSTRA, J. ; DAELEMANS, W. **Cascaded Grammatical Relation Assignment**. In: PROCEEDINGS OF EMNLP/VLC-99, volume 99, p. 239–246, University of Maryland, USA, 1999.

[12] CARDIE, C.; PIERCE, D. **Error-driven pruning of treebank grammars for base noun phrase identification**. In: PROCEEDINGS OF THE 17TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS - VOLUME 1, COLING '98, p. 218–224, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.

[13] CARRERAS, X.; MÀRQUEZ, L. **Phrase Recognition by Filtering and Ranking with Perceptrons**. In: PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON RECENT ADVANCES IN NATURAL LANGUAGE PROCESSING, RANLP-2003, Borovets, Bulgaria, 2003.

[14] CARRERAS, X.; MÀRQUEZ, L. **Introduction to the conll-2004 shared task: Semantic role labeling**. In: PROCEEDINGS OF CONLL-2004, p. 89–97. Boston, MA, USA, 2004.

[15] CARRERAS, X.; MÀRQUEZ, L. **Introduction to the conll-2005 shared task: Semantic role labeling**. In: PROCEEDINGS OF CONLL-2005, p. 152–164. Ann Arbor, MI, USA, 2005.

[16] CARRERAS, X.; MÀRQUEZ, L. ; CASTRO, J. Filtering-ranking perceptron learning for partial parsing. **Machine Learning**, v.60, n.1, p. 41–71, June 2005.

[17] CRESTANA, C. E. M. **A token classification approach to dependency parsing**, March 2010. M.Sc. dissertation in Computer Science - Centro Técnico Científico, Pontifícia Universidade Católica do Rio de Janeiro.

[18] DÉJEAN, H. **Learning syntactic structures with xml**. In: PROCEEDINGS OF CONLL-2000 AND LLL-2000, p. 133–135. Lisbon, Portugal, 2000.

[19] DOS SANTOS, C.; MILIDIÚ, R.; CRESTANA, C. ; FERNANDES, E. **Etl ensembles for chunking, ner and srl**. In: Gelbukh, A., editor, COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING,

PUC-Rio - Certificação Digital Nº 0921334/CA

volume 6008 of **Lecture Notes in Computer Science**, p. 100–112. Springer Berlin / Heidelberg, 2010.

[20] DOS SANTOS, C. N. **Aprendizado de máquina na identificação de sintagmas nominais: O caso do português brasileiro**, February 2005. M.Sc. dissertation in Computer Science - Instituto Militar de Engenharia.

[21] DOS SANTOS, C. N. **Entropy Guided Transformation Learning**. 2009. PhD thesis - Pontifícia Universidade Católica do Rio de Janeiro.

[22] DOS SANTOS, C. N.; MILIDIÚ, R. ; RENTERÍA, R. **Portuguese part-of-speech tagging using entropy guided transformation learning**. In: Teixeira, A.; de Lima, V.; de Oliveira, L. ; Quaresma, P., editors, COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, volume 5190 of **Lecture Notes in Computer Science**, p. 143–152. Springer Berlin / Heidelberg, 2008.

[23] FERNANDES, E.; DOS SANTOS, C. ; MILIDIÚ, R. **A machine learning approach to portuguese clause identification**. In: Pardo, T.; Branco, A.; Klautau, A.; Vieira, R. ; de Lima, V., editors, COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, volume 6001 of **Lecture Notes in Computer Science**, p. 55–64. Springer Berlin / Heidelberg, 2010.

[24] FERNANDES, E.; MILIDIÚ, R. ; DOS SANTOS, C. **Portuguese language processing service**. In: PROCEEDINGS OF THE WEB IN IBERO-AMERICA ALTERNATE TRACK OF THE 18TH WORLD WIDE WEB CONFERENCE, 2009.

[25] FINGER, M. **Técnicas de Otimização da Precisão Empregadas no Etiquetador Tycho Brahe**. In: PROCEEDINGS OF PROPOR, volume 9, p. 141–154, São Paulo, Brazil, 2000.

[26] FREITAS, M.; DUARTE, J.; SANTOS, C.; MILIDIÚ, R.; RENTERÍA, R. ; QUENTAL, V. **A machine learning approach to the identification of appositives**. In: Sichman, J.; Coelho, H. ; Rezende, S., editors, ADVANCES IN ARTIFICIAL INTELLIGENCE - IBERAMIA-SBIA 2006, volume 4140 of **Lecture Notes in Computer Science**, p. 309–318. Springer Berlin / Heidelberg, 2006.

[27] FREITAS, M.; UZEDA-GARRÃO, M.; OLIVEIRA, C.; DOS SANTOS, C. ; SILVEIRA, M. **A anotação de um corpus para o aprendizado**

PUC-Rio - Certificação Digital Nº 0921334/CA

supervisionado de um modelo de SN. In: PROCEEDINGS OF THE III TIL/XXV CONGRESSO DA SBC, p. 2178–2187, 2005.

[28] FREITAS, M. C.; ROCHA, P. ; BICK, E. **Floresta Sintá(c)tica: Bigger, thicker and easier**. In: Teixeira, A.; de Lima, V. L. S.; de Oliveira, L. C. ; Quaresma, P., editors, COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, volume 5190 of **Lecture Notes in Computer Science**, p. 216–219. Springer, 2008.

[29] GEE, J. P.; GROSJEAN, F. Performance structures: A psycholinguistic and linguistic appraisal. **Cognitive Psychology**, v.15, n.4, p. 411–458, 1983.

[30] HAMMERTON, J. Introduction to Special Issue on Machine Learning Approaches to Shallow Parsing. **Journal of Machine Learning Research**, v.19, n.2, p. 313–558, Nov. 2002.

[31] HIGGINS, D. **A transformation-based approach to argument labeling**. In: Ng, H. T.; Riloff, E., editors, HLT-NAACL 2004 WORKSHOP: EIGHTH CONFERENCE ON COMPUTATIONAL NATURAL LANGUAGE LEARNING (CONLL- 2004), p. 114–117, Boston, Massachusetts, USA, 2004. Association for Computational Linguistics.

[32] JOHANSSON, C. **A context sensitive maximum likelihood approach to chunking**. In: PROCEEDINGS OF CONLL-2000 AND LLL-2000, p. 136–138. Lisbon, Portugal, 2000.

[33] JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**. Prentice Hall, 2000.

[34] KINOSHITA, J.; SALVADOR, L. N.; MENEZES, C. E. D. ; SILVA, W. D. C. M. CoGrOO - An OpenOffice Grammar Checker. **Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007)**, p. 525–530, Oct. 2007.

[35] KNUTH, D. E. On the translation of languages from left to right. **Information and Control**, v.8, n.6, p. 607–639, 1965.

[36] KOELING, R. **Chunking with maximum entropy models**. In: PROCEEDINGS OF CONLL-2000 AND LLL-2000, p. 139–141. Lisbon, Portugal, 2000.

[37] KUDO, T.; MATSUMOTO, Y. **Use of support vector learning for chunk identification**. In: PROCEEDINGS OF CONLL-2000 AND LLL-2000, p. 142–144. Lisbon, Portugal, 2000.

PUC-Rio - Certificação Digital Nº 0921334/CA

[38] KUDO, T.; MATSUMOTO, Y. **Chunking with support vector machines**. In: PROCEEDINGS OF THE SECOND MEETING OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON LANGUAGE TECHNOLOGIES, p. 1–8, Pittsburgh, PA, USA, 2001. Association for Computational Linguistics.

[39] MANGU, L.; BRILL, E. **Automatic rule acquisition for spelling correction**. In: PROCEEDINGS OF THE FOURTEENTH ICML, Nashville, Tennessee, USA, 1997. Morgan Kaufmann Publishers Inc.

[40] MANNING, C. D.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**, volume 26. 1. ed., The MIT Press, June 1999.

[41] MARCUS, M. P.; MARCINKIEWICZ, M. A. ; SANTORINI, B. Building a large annotated corpus of english: the penn treebank. **Computational Linguistics**, v.19, p. 313–330, June 1993.

[42] MILIDIÚ, R.; DOS SANTOS, C. ; DUARTE, J. **Phrase chunking using entropy guided transformation learning**. In: PROCEEDINGS OF ACL 2008, number June, p. 647–655, Columbus, Ohio, 2008. Citeseer.

[43] MILIDIÚ, R.; SANTOS, C.; DUARTE, J. ; RENTERÍA, R. **Semi-supervised learning for portuguese noun phrase extraction**. In: Vieira, R.; Quaresma, P.; Nunes, M.; Mamede, N.; Oliveira, C. ; Dias, M., editors, COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, volume 3960 of **Lecture Notes in Computer Science**, p. 200–203. Springer Berlin / Heidelberg, 2006.

[44] MILIDIÚ, R. L.; DOS SANTOS, C. N. ; DUARTE, J. C. Portuguese corpus-based learning using ETL. **Journal of the Brazilian Computer Society**, v.14, p. 17–27, 12 2008.

[45] MILIDIÚ, R. L.; DUARTE, J. C. ; CAVALCANTE, R. Machine Learning Algorithms for Portuguese Named Entity Recognition. **Revista Iberoamericana de Inteligencia Artificial**, v.11, n.36, p. 67–75, 2007.

[46] MOLINA, A.; PLA, F. Shallow Parsing using Specialized HMMs. **Journal of Machine Learning Research**, v.2, n.4, p. 595–613, Nov. 2002.

[47] MUÑOZ, M.; PUNYAKANOK, V.; ROTH, D. ; ZIMAK, D. **A learning approach to shallow parsing**. In: PROCEEDINGS OF EMNLP/WVLC-99, University of Maryland, MD, USA, 1999.

PUC-Rio - Certificação Digital Nº 0921334/CA

[48] OSBORNE, M. **Shallow parsing as part-of-speech tagging**. In: PROCEEDINGS OF CONLL-2000 AND LLL-2000, p. 145–147. Lisbon, Portugal, 2000.

[49] PLA, F.; MOLINA, A. ; PRIETO, N. **Improving chunking by means of lexical-contextual information in statistical language models**. In: PROCEEDINGS OF CONLL-2000 AND LLL-2000, p. 148–150. Lisbon, Portugal, 2000.

[50] QUINLAN, J. R. **C4.5: Programs for Machine Learning**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[51] RAMSHAW, L.; MARCUS, M. **Text chunking using transformation-based learning**. In: PROCEEDINGS OF THE THIRD ACL WORKSHOP ON VERY LARGE CORPORA, p. 82–94. Cambridge MA, USA, 1995.

[52] RATNAPARKHI, A. **Maximum Entropy Models for Natural Language Ambiguity Resolution**. 1998. PhD thesis - University of Pennsylvania.

[53] SAGAE, K.; LAVIE, A. ; MACWHINNEY, B. **Automatic Measurement of Syntactic Development in Child Language**. In: PROCEEDINGS OF THE 43RD ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, number June, p. 197–204, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

[54] SHA, F.; PEREIRA, F. **Shallow parsing with conditional random fields**. In: PROCEEDINGS OF THE 2003 CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON HUMAN LANGUAGE TECHNOLOGY-VOLUME 1, number 1995, p. 134–141, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[55] SUN, W.; SUI, Z.; WANG, M. ; WANG, X. **Chinese semantic role labeling with shallow parsing**. In: PROCEEDINGS OF THE 2009 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, number August, p. 1475–1483, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[56] TJONG KIM SANG, E. F. **Text chunking by system combination**. In: PROCEEDINGS OF CONLL-2000 AND LLL-2000, p. 151–153. Lisbon, Portugal, 2000.

PUC-Rio - Certificação Digital Nº 0921334/CA

[57] TJONG KIM SANG, E. F. **Transforming a Chunker to a Parser**. In: COMPUTATIONAL LINGUISTICS IN THE NETHERLANDS, p. 177–188, 2000.

[58] TJONG KIM SANG, E. F. **Introduction to the conll-2002 shared task: Language-independent named entity recognition**. In: PROCEEDINGS OF CONLL-2002, p. 155–158. Taipei, Taiwan, 2002.

[59] TJONG KIM SANG, E. F. Memory-based shallow parsing. **Journal of Machine Learning Research**, v.2, p. 559–594, March 2002.

[60] TJONG KIM SANG, E. F.; BUCHHOLZ, S. **Introduction to the conll-2000 shared task: Chunking**. In: PROCEEDINGS OF CONLL-2000 AND LLL-2000, p. 127–132. Lisbon, Portugal, 2000.

[61] TJONG KIM SANG, E. F.; DE MEULDER, F. **Introduction to the conll-2003 shared task: Language-independent named entity recognition**. In: Daelemans, W.; Osborne, M., editors, PROCEEDINGS OF CONLL-2003, p. 142–147. Edmonton, Canada, 2003.

[62] TJONG KIM SANG, E. F.; DÉJEAN, H. **Introduction to the conll-2001 shared task: Clause identification**. In: Daelemans, W.; Zajac, R., editors, PROCEEDINGS OF CONLL-2001, p. 53–57. Toulouse, France, 2001.

[63] TJONG KIM SANG, E. F.; VEENSTRA, J. **Representing text chunks**. In: PROCEEDINGS OF THE NINTH CONFERENCE ON EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, EACL '99, p. 173–179, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

[64] VAN HALTEREN, H. **Chunking with wpdv models**. In: PROCEEDINGS OF CONLL-2000 AND LLL-2000, p. 154–156. Lisbon, Portugal, 2000.

[65] VEENSTRA, J. **Memory-based text chunking**. In: Fakotakis, N., editor, "MACHINE LEARNING IN HUMAN LANGUAGE TECHNOLOGY", WORKSHOP AT ACAI 99, volume 99, Chania, Greece, 1999.

[66] VEENSTRA, J.; BUCHHOLZ, S. **Fast NP chunking using memory-based learning techniques**. In: Verdenius, F.; van den Broek, W., editors, PROCEEDINGS OF BENELEARN'98, p. 71–78, Wageningen, The Netherlands, 1998. Citeseer.

PUC-Rio - Certificação Digital Nº 0921334/CA

[67] VEENSTRA, J.; VAN DEN BOSCH, A. **Single-classifier memory-based phrase chunking**. In: PROCEEDINGS OF CONLL-2000 AND LLL-2000, p. 157–159. Lisbon, Portugal, 2000.

[68] VILAIN, M.; DAY, D. **Phrase parsing with rule sequence processors: an application to the shared conll task**. In: PROCEEDINGS OF CONLL-2000 AND LLL-2000, p. 160–162. Lisbon, Portugal, 2000.

[69] WATANABE, T.; SUMITA, E. ; OKUNO, H. G. **Chunk-based statistical translation**. In: PROCEEDINGS OF THE 41ST ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS - ACL '03, number July, p. 303–310, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[70] WU, Y.; CHANG, C.-H. ; LEE, Y. A general and multi-lingual phrase chunking model based on masking method. **Computational Linguistics and Intelligent Text Processing**, p. 144–155, 2006.

[71] WU, Y.; YANG, J.; LEE, Y. ; YEN, S. Efficient and Robust Phrase Chunking using Support Vector Machines. **Information Retrieval Technology**, p. 350–361, 2006.

[72] ZHANG, T.; DAMERAU, F. ; JOHNSON, D. **Text chunking using regularized winnow**. In: PROCEEDINGS OF THE 39TH ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, p. 539–546, Toulouse, France, 2001. Association for Computational Linguistics.

[73] ZHANG, T.; DAMERAU, F. ; JOHNSON, D. Text Chunking based on a Generalization of Winnow. **Journal of Machine Learning Research**, v.2, p. 615–637, Nov. 2002.

[74] ZHOU, G.; SU, J. ; TEY, T. **Hybrid text chunking**. In: PROCEEDINGS OF CONLL-2000 AND LLL-2000, p. 163–165. Lisbon, Portugal, 2000.

PUC-Rio - Certificação Digital Nº 0921334/CA

A
Part-of-Speech Tags

The set of part-of-speech tags applied by the POS extractor used in this work is described here.

POS	Description
ADJ	Adjective
ADV	Adverb
ADV-KS	Subordinating adverb
ADV-KS-REL	Relative subordinating adverb
ART	Article
CUR	Currency symbol
IN	Interjection
KC	Coordinating conjunction
KS	Subordinating conjunction
N	Noun
NPROP	Proper noun
NUM	Number
PCP	Past participle
PDEN	Denotative word
PREP	Preposition
PROADJ	Adjective pronoun
PROPESS	Personal pronoun
PROSUB	Substantive pronoun
PRO-KS	Subordinating pronoun
PRO-KS-REL	Relative subordinating pronoun
V	Verb
VAUX	Auxiliary verb

Table A.1: Automatically extracted part-of-speech tags

PUC-Rio - Certificação Digital Nº 0921334/CA

B
ETL Baseline Systems

In this appendix, we present the list of associations between POS tags and target tags for each baseline system used by our ETL models.

The adopted tagger outputs a distinct POS tag for each punctuation mark. We omit these punctuation POS tags for brevity and because of their low occurrence compared to the other kinds.

	Baseline systems		
POS	**Direct**	**Chunk start**	**Chunk end**
ADJ	I-NP	X	X
ADV	O	X	X
ADV-KS	O	X	X
ADV-KS-REL	I-NP	X	NP
ART	B-NP	NP	X
CUR	B-NP	NP	NP
IN	O	X	X
KC	O	X	X
KS	O	X	X
N	I-NP	X	X
NPROP	I-NP	X	X
NUM	B-NP	NP	X
PCP	B-VP	X	VP
PDEN	O	X	X
PREP	I-NP	X	NP
PROADJ	B-NP	NP	X
PROPESS	B-NP	NP	NP
PROSUB	B-NP	NP	NP
PRO-KS	B-NP	NP	NP
PRO-KS-REL	B-NP	NP	NP
VAUX	B-VP	VP	X
V	B-VP	VP	VP

Table B.1: Tag associations for (NP, VP) baseline systems

PUC-Rio - Certificação Digital Nº 0921334/CA

	Baseline systems		
POS	**Direct**	**Chunk start**	**Chunk end**
ADJ	I-NP	X	NP
ADV	O	X	X
ADV-KS	O	X	X
ADV-KS-REL	I-NP	X	NP
ART	B-NP	NP	X
CUR	B-NP	NP	NP
IN	O	X	X
KC	O	X	X
KS	O	X	X
N	I-NP	X	NP
NPROP	I-NP	X	NP
NUM	B-NP	NP	X
PCP	B-VP	X	VP
PDEN	O	X	X
PREP	B-PP	PP	PP
PROADJ	B-NP	NP	X
PROPESS	B-NP	NP	NP
PROSUB	B-NP	NP	NP
PRO-KS	B-NP	NP	NP
PRO-KS-REL	B-NP	NP	NP
VAUX	B-VP	VP	X
V	B-VP	VP	VP

Table B.2: Tag associations for (NP, VP, PP) baseline systems

PUC-Rio - Certificação Digital Nº 0921334/CA

	Baseline systems		
POS	**Direct**	**Chunk start**	**Chunk end**
ADJ	B-ADJP	ADJP	ADJP
ADV	B-ADVP	ADVP	ADVP
ADV-KS	B-ADVP	ADVP	ADVP
ADV-KS-REL	B-ADVP	ADVP	ADVP
ART	B-NP	NP	X
CUR	B-NP	NP	NP
IN	B-ADVP	ADVP	ADVP
KC	O	X	X
KS	O	X	X
N	I-NP	X	NP
NPROP	I-NP	X	NP
NUM	B-NP	NP	X
PCP	B-VP	X	VP
PDEN	B-ADVP	ADVP	ADVP
PREP	B-PP	PP	PP
PROADJ	B-NP	NP	X
PROPESS	B-NP	NP	NP
PROSUB	B-NP	NP	NP
PRO-KS	B-NP	NP	NP
PRO-KS-REL	B-NP	NP	NP
VAUX	B-VP	VP	X
V	B-VP	VP	VP

Table B.3: Tag associations for ($NP, VP, PP, ADJP, ADVP$) baseline systems

PUC-Rio - Certificação Digital Nº 0921334/CA