Investigating Lexical NP-Chunking with Universal Dependencies for Portuguese

Aleksander Tomaz de Souza[1,2]
Evandro Eduardo Seron Ruiz[1,2]

[1] Center for Artificial Intelligence (C4AI) – INOVA.USP
Butantã, São Paulo, SP – Brasil

[2]Departamento de Computação e Matemática
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP)
Universidade de São Paulo – USP
Ribeirão Preto, SP – Brasil

aleksander@usp.br, evandro@usp.br

Abstract. *The task of shallow parsing consists of retrieving a limited amount of syntactic information from sentences written in natural language. This work aims to identify and extract a particular kind of noun phrase called lexical noun phrase (lNP). This paper's initial studies show the possibility of deducing lNP chunks from Universal Dependency annotated sentences written in Portuguese. We also demonstrate how the task of shallow parsing can benefit PoS-tagging using a committee of machine learning algorithms.*

Resumo. *A tarefa de análise superficial consiste em recuperar uma quantidade limitada de informações sintáticas de frases escritas em linguagem natural. Este trabalho tem como objetivo identificar e extrair um tipo particular de sintagma nominal denominado sintagma nominal lexical (SN_L). Os estudos iniciais mostrados neste artigo demostram, em primeira mão, a possibilidade de identificar e extrair SN_L a partir de sentenças escritas em português e anotadas pelo formalismo da* Universal Dependency. *Também demonstramos como a tarefa de análise sintática superficial pode se beneficiar das marcações de PoS usando um comitê de algoritmos de aprendizado de máquina.*

1. Introdução

A tarefa de análise sintática parcial, também conhecida por *shallow parsing* (SP), consiste na recuperação de uma quantidade limitada de informações sintáticas de frases escritas em linguagem natural [Hammerton et al. 2002]. Nesta tarefa destacamos a recuperação dos sintagmas nominais (SN). Dentre os sintagmas existentes, os SN são os sintagmas compostos por termos de teor substantivo, sendo este o elemento fundamental de sua estrutura. Tradicionalmente, nomes e pronomes assumem esse papel. Dado o alto poder discriminatório e o potencial informativo dos SN [Oliveira and Freitas 2006], a sua identificação é muito importante para diversas aplicações computacionais, tais como a Recuperação e a Extração de Informação (RI).

Considerando os modelos generalistas que classificam os SN, Oliveira e Freitas [Oliveira and Freitas 2006] propuseram um tipo específico de SN para ser utilizado

em sistemas de RI denominado Sintagmas Nominais Lexicais (SN_L). Estes sintagmas privilegiam expressões substantivas autônomas que têm a faculdade de exercer o papel de termos de indexação e, portanto, são estruturas sintagmáticas apontadas como elementos fundamentais para a síntese textual.

A identificação automatizada dos SN_L é uma atividade complexa, pois eles permitem, nas suas diferentes estruturas de composição, os sintagmas adjetivais, os sintagmas preposicionais, além de outros critérios. Abaixo vemos três períodos em que os termos em negrito correspondem à SN_L.

1. **A caneta** é esferográfica.
2. **Caneta** e **papel** para escrever.
3. **Caneta esferográfica *Montblanc*** para escrever em **papel apergaminhado de cor sépia**.

Nos exemplos acima apresentados, percebemos ocorrências tais como: i) segmentos nominais coordenados que devem ser compreendidos como SN_L independentes (Exemplo 2) e, ainda, ii) em sentenças de maior extensão, em que para determinar os limites de seu domínio, deve-se compreender segmentos sintagmáticos adjetivais e preposicionais (Exemplo 3).

Neste contexto, este artigo aborda a identificação automatizada de SN_L, em frases escritas em português e anotadas no formalismo da *Universal Dependency*, UD, de McDonald, Nivre e colaboradores [McDonald et al. 2013], utilizando técnicas de aprendizado de máquina (AM). Na seção seguinte, Seção 2, apresentamos brevemente os principais trabalhos relacionados à análise sintática superficial usando técnicas de AM. Na Seção 3 apresentamos os dados e a metodologia usados nos experimentos de identificação de SN_L e, subsequentemente, na Seção 4 relatamos os resultados dos experimentos por nós realizados. Algumas considerações sobre os resultados do experimento estão detalhadas na Seção 5, bem como algumas considerações gerais sobre o tema.

2. Trabalhos relacionados

Os SN_L são destacados entre os sintagmas nominais por tomarem apenas substantivos como elemento principal. Essa fundamentação teórica pode ser vista no trabalho de Andrew Radford [Radford 1981], ainda na década de 1980. Somente quase 20 anos depois Ramshaw e Marcus [Ramshaw and Marcus 1999] criaram um sistema computacional de fragmentação textual para a divisão da sentença em segmentos não-sobrepostos, feito esse que formou a base para os futuros estudos de análise sintática parcial.

Ramshaw e Marcus também inauguraram a aplicação de AM para a tarefa de identificação de sintagmas. Eles usaram a metodologia de *Transformation-Based Learning*, TBL, para essa função de *shallow parsing*. Este método obteve precisão e revogação na ordem de 92% para SN e 89% para outros sintagmas. Na metodologia TBL, ou seja, a de aprendizagem baseada em transformação, a ideia da aprendizagem é começar com alguma solução simples (regras iniciais) que identifique os sintagmas e aplicar transformações (novas regras) que melhorem o desempenho anterior de marcação dos sintagmas. O TBL é uma abordagem de classificação linear, assim como o é o algoritmo de Winnow [Littlestone 1988], também muito utilizado para *parsing* parcial.

A abordagem de *memory-based learning* foi usada por Erik Sang [Sang 2002], em 2002 para resolver o problema de SP. Sang usou essa abordagem para identificar sintagmas nominais e também para o *parsing* completo. Neste artigo Sang relatou que obteve precisão e revogação de $\approx 93\%$ para os sintagmas nominais. Praticamente na mesma época, Molina e Pla [Molina and Pla 2002] foram pioneiros na aplicação de cadeias escondidas de Markov (HMM) para a mesma tarefa. Estes autores conseguiram resultados como F-score $= 93.25\%$, ou seja, resultados equiparados ao estado da arte para a tarefa de SP naquela época.

Choi e colaboradores [Choi et al. 2005] também se depararam com este problema de recuperar informação sintática eficientemente sem recorrer a uma análise sintática completa. A ideia dos autores foi encontrar regras gramaticais delimitadoras de sintagmas, fazendo do problema de SP um problema de classificação que eles resolveram usando árvores de decisão. O melhor resultado obtido pelo grupo foi F-score $= 91.7\%$.

Outras línguas, além do inglês, também se beneficiaram de métodos de aprendizado de máquina para realizar SP, tal como a língua turca [Topsakal et al. 2017] e o hindi-inglês [Sharma et al. 2016], uma fusão de duas línguas, muito popular na Índia. De modo análogo, foram também construídos SP para máquinas de tradução, entre elas para a tradução português-inglês, como a desenvolvida na Universidade de Alicante, pelo grupo de Garrido Alenda [Garrido Alenda et al. 2004]. Pelas nossas pesquisas, João Ricardo da Silva [da Silva 2007], em sua tese de doutoramento, inaugura a segmentação frasal de textos escritos em Português usando um SP construído sobre autômatos de estados finitos. O resultado deste trabalho foi relatado pouco tempo antes [Branco and Silva 2006] conferindo uma precisão de 99,92% e revocação de 99.95% sobre um corpus anotado manualmente de 280 mil *tokens* composto por artigos de jornais e novelas, conhecido como LX-Corpus [Branco and Silva 2004]. De acordo com as nossas pesquisas bibliográficas, o trabalho de Ophélie Lacroix [Lacroix 2018] inaugurou a identificação de marcações sintagmáticas nominais sobre textos escritos em Inglês e anotados no formalismo *Universal Dependency*.

Percebemos que muitos destes métodos de SP foram desenvolvidos na primeira década dos anos 2000. A próxima década, a década inciada em 2011, marca o período de ascensão das abordagens baseadas em *deep learning* as quais não são tratadas neste texto.

3. Metodologia

Dados

A definição dos SN_L têm sua natureza nos modelos de anotação sintagmática (constituintes) existente. Assim, usando seleção aleatória, extraímos do *corpus* CETENFolha, versão 1.0 [Santos et al. 2004], um conjunto de 102 sentenças. Os dados sintáticos contidos nesse *corpus* permitem a análise gráfica de suas estruturas hierárquicas que nos nortearam para identificação dos sintagmas noutro *corpus*. O interesse nesse *corpus* deu-se devido suas sentenças serem: i) estruturadas segundo a gramática de constituintes; ii) serem escritas em português brasileiro e iii) por estarem entre as sentenças anotadas e revisadas manualmente pelo formalismo *Universal Dependencies* no *córpus* Bosque, versão 2.0 [Rademaker et al. 2017]. Ou seja, para cada sentença temos a marcação dos SN em dois formalismos: constituintes e UD.

Um exemplo dessas perspectivas pode ser observado na frase '**A região vive uma epidemia da doença**', extraída do *corpus* Bosque, CENTENFolha. Nesta frase há a sobreposição do sintagma nominal '**uma epidemia da doença**' sobre um sintagma preposicional, '**da doença**', e, subsequentemente, deste sobre um outro sintagma nominal, '**a doença**', que destaca um dos SN_L dessa oração. Veja a Figura 1.

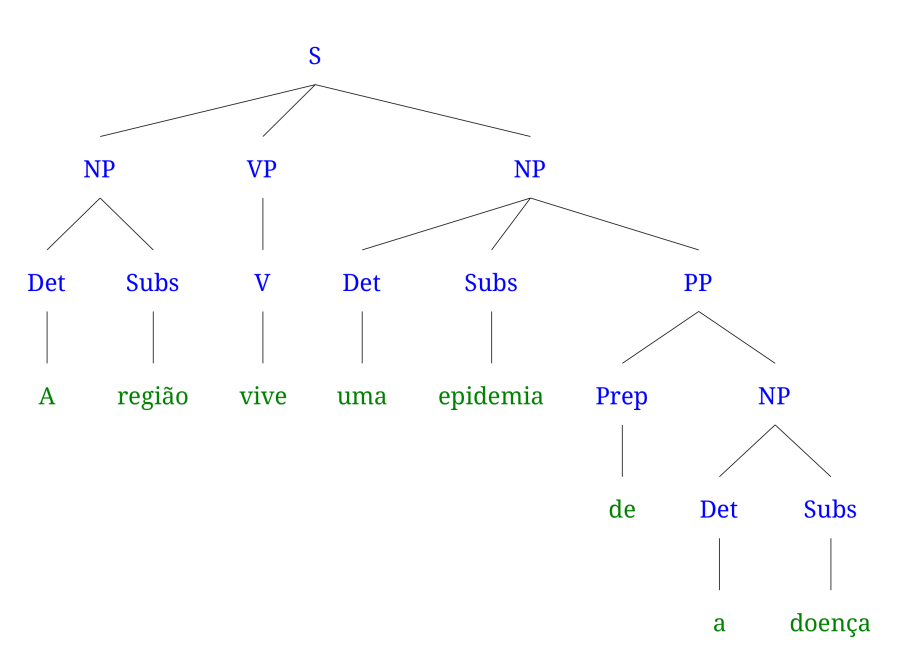

Figura 1. Estrutura sintática representando a gramática sobre o formalismo de constituintes.

Na Figura 2, para a mesma frase apontada anteriormente, também inserida no *corpus* Bosque, vemos a sua representação segundo a gramática UD. Nesta anotação, perpassamos a hierarquia recursiva de estruturas frasais apreendendo essas atribuições por meio das relações de dependência universal.

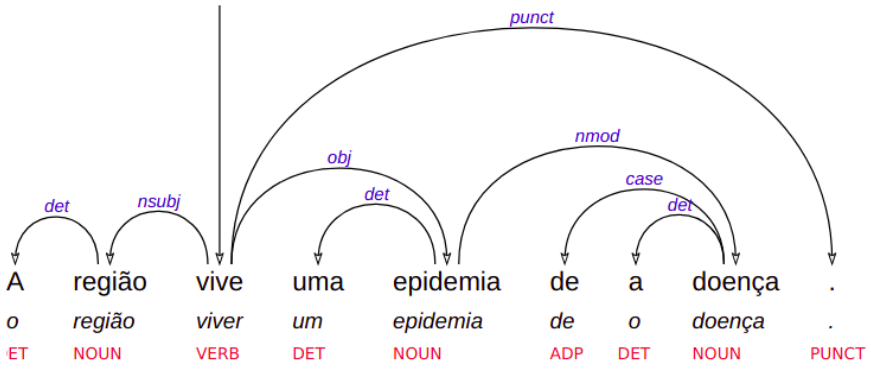

Figura 2. Estrutura sintática representando a gramática no formalismo UD.

Observamos que nas sentenças selecionadas para nosso *corpus* há representações que expressam desde sentenças nominais e simples, como também aquelas extensas e

mais complexas nas quais estão presentes fenômenos, tais como: coordenação, elipses e expressões lexicais. Fenômenos linguísticos estes que nos dão exemplos e contraexemplos quanto a presença ou ausência dos SN_L.

Recorrendo a esses *corpora*, identificamos, pelas estruturas de constituintes frasais, quais ramos sintagmáticos correspondiam às principais características dos SN_L propostas por Oliveira e Freitas [Oliveira and Freitas 2006]. Uma vez conhecidos esses segmentos, os buscamos nas mesmas sentenças anotadas segundo o formalismo UD, lançando de modo manual a marcação no formato BIO (*Beginning-Inside–Outside tagging*). Nessa marcação BIO uma sentença é marcada em cada um de seus elementos expondo as diferentes 'assinaturas lexicais' que compõem um SN_L, ou não, nessa recente representação formal da sintaxe.

Baseado nos dois *coprrpora* mencionados, construímos nosso *córpus* com o total de 1.947 *tokens* marcados com as classes morfossintáticas (*Part of Speech*, PoS), as relações de dependência UD, além dos rótulos BIO. Reservamos 70% das sentenças para base de treino e 30% delas para base de teste.

Algoritmos

Para nossa pesquisa escolhemos um algoritmo baseado em AM tradicional (TBL) e outros reconhecidos como mais avançados e que formaram um comitê de algoritmos. O algoritmo *Transformation Based Learning*, (TBL) [Brill 1993] é uma abordagem supervisionada, baseada num *corpus* anotado e orientada à minimização de erros.

Pelo TBL um conjunto de regras de transformação e melhoramento são aplicados para corrigir os erros encontrados. Resumidamente, este é o seu funcionamento: São criadas regras baseadas num modelo padrão, um *template*. Esse *template* é definido de acordo extensão de elementos próximos ao termo em análise a ser considerado para formação das regras de transformação. Essas regras aplicadas ao *corpus* classificam segmentos das sentenças. Após a classificação, o algoritmo corrige os erros deixados na iteração anterior tendo o corpus de treino como referência. Essas regras são avaliadas por uma função de pontuação e a regra com a pontuação mais alta é selecionada pelo modelo TBL. Assim, as regras de transformação são adicionadas ao modelo até que um limite de pontuação seja alcançado ou nenhuma transformação de correção possa ser criada. Neste momento o algoritmo para.

Os demais algoritmos baseiam-se em modelos que remetem à de árvores de decisão e florestas de decisão, com e sem gradientes de regressão ponderados, tais como XGBClassifier – XGBCL, XGBRFClassifier – XGBRFC, DecisionTreeClassifier – DTC e RandomForestClassifier – RFCL, respectivamente, também em modelos baseados em *perceptron* e rede neurais (Perceptron – PCTR e MLPClassifier – MLPCL) em versões que permitem tratamento de dados categóricos após pré-processamento. Para esses algoritmos, realizamos três formas de apresentação de dados quanto aos atributos independentes: i) somente usando as marcações de dependência UD; ii) somente as marcações de PoS, e; iii) dados de relações UD associados as marcações PoS, orientando a classificação do atributo dependente usando rótulos BIO. Assim, procuramos representar o "conhecimento sintático" extraído do *corpus* de forma abstrata, atribuindo a cada termo de uma sentença as marcações BIO.

4. Resultados

Os resultados alcançados pelo aprendizado com o algoritmo TBL são apresentados na Tabela 1. Esta tabela está organizada de acordo com os dados utilizados nos treinamento e teste. Num primeiro momento, recorremos as classes morfossintáticas (PoS) dos termos e aos rótulos BIO. Num segundo experimento, usamos as relações UD que, da mesma forma, reportam as métricas obtidas para cada rótulo. Por essa Tabela 1 observamos que a utilização dos atributos morfossintáticos dos termos no algoritmo TBL alcançou uma acurácia ponderada maior, 87,0%, frente a utilização das relações UD que alcançaram 85,1%. Quanto às regras utilizadas pelo algoritmo, as principais regras aplicadas basearam-se na classificação dos termos imediatamente antes e depois da marcação BIO. Quando utilizadas as classes morfossintáticas, aquelas que alteraram estados iniciais dos rótulos 'BI' para o estado final 'O' alcançaram uma revocação média de 88,8%. Tomando como recurso as relações UD, as regras construídas atingiram uma revocação de 88,1% quanto a esse mesmo aspecto.

Tabela 1. Resultados percentuais do algoritmo TBL.

Rótulos	Métricas utilizadas			
	Precisão	*Revocação*	*Medida-F*	*Acurácia*
PoS				87,0
tag B	77,3	69,0	72,9	–
tag I	86,2	79,3	82,6	–
tag O	82,0	88,8	85,3	–
Relações UD				85,1
tag B	74,6	66,6	70,4	–
tag I	83,7	76,1	79,7	–
tag O	80,8	88,1	84,3	–

A Tabela 2 mostra as principais regras usadas pelo algoritmo TBL na marcação dos elementos que compõem os SN_L usando os rótulos *PoS*. Estas regras consideram o fator posicional anterior e posterior dos termos como determinante. Temos, como exemplo, a regra 044 para exclusão de elementos fora dos domínios do sintagma. Notamos que o algoritmo recorre a aplicação do padrão BIO para diferenciar suas principais regras. Por exemplo, na regra 044, quando são usados os dados morfossintáticos, esta regra considera que há uma grande probabilidade de, entre dois rótulos 'O', ocorrer outro 'O', alterando assim um estado inicial de marcação 'B' para 'O'. Note que nesta Tabela 2 algumas regras recebem o mesmo número, mas empregam argumentos distintos para marcação dos rótulos BIO.

O algoritmo TBL, ao processar relações UD, aponta regras de aprendizado que destacam as relações UD 'amod', 'det' e 'nsubj' para identificação de elementos que pertencem a um SN_L. Perceba também que, na Tabela 3, as relações tipo 'amod' (representante de modificadores, adjetivos) podem iniciar um SN_L, como mostrado na regra 030. Outro exemplo pode ser observado na regra 029 em que uma relação 'det' (representante de determinante, artigos) precede um termo que está inserido em um SN_L. Ainda, a regra 035 revela que a classificação sintática 'nsubj:pass' que representa três elementos anteriores ao *token* e insere o termo atual nesse sintagma.

Tabela 2. Regras compostas pelos rótulos morfossintáticos (PoS) e rótulos BIO.

Regra	*tag* Inicial	*tag* Final	Descrição
044	'B'	'O'	(*tag*[-1] 'O') e (*tag*[1] 'O')
046	'B'	'O'	(*token*[1] None)
044	'I'	'O'	(*tag*[-1], 'O') e (*tag*[1], 'I')
044	'I'	'O'	(*tag*[-1], 'O') e (*tag*[1], 'I')
053	'I'	'O'	(*token*[-1], 'NOUN') e (*token*[1], 'PRON')
053	'O'	'I'	(*token*[-1], 'NOUN') e (*token*[1], 'ADJ')
044	'I'	'O'	(*tag*[-1], 'O') e (*tag*[1], 'I')
044	'I'	'O'	(*tag*[-1]), 'O') e (tag[1], 'I')

Tabela 3. Regras compostas pelas relações UD e rótulos BIO.

Regra	*tag* Inicial	*tag* Final	Descrição
030	'O'	'B'	(*token*[1], 'amod')
029	'O'	'I'	(*token*[-1], 'det')
035	'O'	'I'	(*token*[-3, -2, -1]), 'nsubj:pass')
020	'B'	'I'	(*tag*[-1], 'I')
032	'I'	'O'	(*token*[2], 'conj')
037	'I'	'B'	(*token*[-1], 'case') e (*token*[1], 'det')
021	'I'	'O'	(*tag*[1], 'B')
037	'I'	'O'	(*token*[-1]), 'amod') e (*token*[1], 'det')

Os demais algoritmos empregados permitem o tratamento de dados de maneira conjunta ou não, em outras palavras, os rótulos morfossintáticos e os rótulos das relações UD de cada termo podem ser submetidos separadamente ou em conjunto para o treinamento e teste. A Tabela 4 mostra os resultados obtidos por meio da apresentação, tanto de dados morfossintáticos como também das relações UD, modo este último que obtivemos os melhores resultados. Tais algoritmos foram dispostos sequencialmente a um mesmo fluxo de dados e a cada iteração forneciam as métricas que são apresentadas na Tabela 4. O algoritmo XGBCL destacou-se, frente os demais, em todas as métricas. A medida de precisão, 81,2%, foi +1,2 p.p. mais eficiente que seus concorrentes DTC e RFCL. A métrica de revocação atingiu 81,4%, +1,2 p.p. diante algoritmo MLPCL. O cálculo de medida-F de 81,2% foi +1,1 p.p. que o MLPCL. Por fim, sua acurácia de 81,4% reservou +1.2 p.p. perante do MLPCL que, nesse cenário, foi o que mais se aproximou de seus resultados.

Tabela 4. Resultados do comitê de algoritmos.

Algoritmo	Métricas utilizadas			
	Precisão	*Revocação*	*Medida-F*	*Acurácia*
DTC	80,1	80,0	79,9	80,0
RFCL	80,1	*80,0*	*79,9*	*80,0*
PCTR	81,0	78,5	78,6	78,5
MLPCL	80,5	80,2	80,1	80,2
XGBCL	81,2	81,4	81,2	81,4
XGBRFC	79,6	79,7	79,3	79,7

5. Conclusão

Diante da quantidade de dados utilizados para treinamento dos algoritmos, o TBL demonstra ser mais assertivo do que demais algoritmos, tanto no cenário em que utilizou dados morfossintáticos, quanto no cenário em que foram utilizadas as relações de dependência universal (UD). Considerando a utilização de atributos morfossintáticos no TBL, percebe-se que suas regras são caracterizadas pela exclusão de termos inicialmente inseridos num SN_L. Quando tomadas as relações UD, a composição das regras são mais definidas pela inserção de termos nesse sintagma. Ainda mais, por esse recurso, pode-se identificar, para esse *corpus*, quais relações UD foram mais significativas para na composição de um SN_L alcançando maior acurácia se comparada a métrica obtida no comitê formado. O desempenho dos algoritmos do comitê expõe a dependência desses últimos a um maior volume e variabilidade de dados para expressarem seu potencial, conforme argumenta Diana Santos [Santos 2021]. O aprendizado computacional que recorre a classificação de rótulos BIO permite a identificação de fragmentos que compõe um SN_L e, com isso, suas configurações mais extensas podem ter seus limites mal definidos ou descontinuados. Ainda considerando o TBL, a exclusão de elementos inicialmente considerados pertencentes a esse sintagma foi a estratégia mais utilizada por este algoritmo. Para o TBL as marcações morfossintáticas proporcionaram classificações melhores em +1,9 p.p. do que com as relações UD. Por fim, destacamos como contribuições futuras: i) a ampliação do *corpus* com maior quantidade de sentenças anotadas para reafirmar ou não do desempenho do TBL frente aos tais algoritmos no estado da arte; ii) a revisão do *corpus* por linguistas; iii) aproximações que incrementem a precisão e a revocação alcançadas até este momento e iv) a identificação desse tipo específico de sintagma em outros idiomas para reafirmar a proposta do projeto *Universal Dependencies*, bem como a correlação dos SN_L nas diferentes línguas naturais.

Referências

Branco, A. and Silva, J. (2004). Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Branco, A. and Silva, J. R. (2006). A suite of shallow processing tools for Portuguese: LX-suite. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, pages 179–182, Trento, Italy. Association for Computational Linguistics.

Brill, E. D. (1993). *A corpus-based approach to language learning*. PhD thesis, University of Pennsylvania.

Choi, M.-S., Lim, C. S., and Choi, K.-S. (2005). Automatic Partial Parsing Rule Acquisition Using Decision Tree Induction. In Dale, R., Wong, K.-F., Su, J., and Kwong, O. Y., editors, *Natural Language Processing – IJCNLP 2005*, pages 143–154, Berlin, Heidelberg. Springer Berlin Heidelberg.

da Silva, J. R. M. F. (2007). *Shallow processing of Portuguese: From sentence chunking to nominal lemmatization*. PhD thesis, Universidade de Lisboa, Faculdade de Ciências.

Garrido Alenda, A., Gilabert Zarco, P., Pérez-Ortiz, J. A., Pertusa, A., Ramírez Sánchez, G., Sánchez-Martínez, F., Scalco, M. A., and Forcada, M. L. (2004). Shallow parsing for Portuguese-Spanish machine translation. In *Workshop Notes of TASHA'2003*, pages 21–24, Lisboa, Portugal. Edições Colibri.

Hammerton, J., Osborne, M., Armstrong, S., and Daelemans, W. (2002). Introduction to Special Issue on Machine Learning Approaches to Shallow Parsing. *Journal of Machine Learning Research*, 2:551–558.

Lacroix, O. (2018). Investigating NP-Chunking with Universal Dependencies for English. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Molina, A. and Pla, F. (2002). Shallow Parsing using Specialized HMMs. *Journal of Machine Learning Research*, 2(4):595–613.

Oliveira, C. and Freitas, M. C. d. (2006). Um modelo de sintagma nominal lexical na recuperação de informações. *XI Simpósio Nacional e I Simpósio Internacional de Letras e Linguística (XI SILEL)*, pages 778–786.

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and de Paiva, V. (2017). Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206, Pisa,Italy. Linköping University Electronic Press.

Radford, A. (1981). *Syntactic Theory and the Structure of English: A Minimalist Approach*. Cambridge Textbooks in Linguistics.

Ramshaw, L. A. and Marcus, M. P. (1999). Text Chunking Using Transformation-Based Learning. In Armstrong, S., Church, K., Isabelle, P., Manzi, S., Tzoukermann, E., and Yarowsky, D., editors, *Natural Language Processing Using Very Large Corpora*, pages 157–176. Springer Netherlands, Dordrecht.

Sang, E. T. K. (2002). Memory-Based Shallow Parsing. *Journal of Machine Learning Research*, 2:559–595.

Santos, D., Simões, A., Frankenberg-Garcia, A., Pinto, A., Barreiro, A., Maia, B., Mota, C., Oliveira, D., Bick, E., Ranchhod, E., et al. (2004). Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa. In *Iberomerican Conference on Artificial Intelligence*, pages 147–154. Guillermo de Ita Luna, Olac Fuentes Chávez and, Mauricio Osorio Galindo.

Santos, D. S. M. (2021). Grandes quantidades de informação: um olhar crítico. In *II Congresso Internacional em Humanidades Digitais*, Online. UFRJ.

Sharma, A., Gupta, S., Motlani, R., Bansal, P., Shrivastava, M., Mamidi, R., and Sharma, D. M. (2016). Shallow Parsing Pipeline – Hindi-English Code-Mixed Social Media Text. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1340–1345, San Diego, California. Association for Computational Linguistics.

Topsakal, O., Açikgöz, O., Gürkan, A. T., Kanburoglu, A. B., Ertopçu, B., Özenç, B., Çam, I., Avar, B., Ercan, G., and Yildiz, O. T. (2017). Shallow parsing in Turkish. In *2017 International Conference on Computer Science and Engineering (UBMK)*, pages 480–485.

Acknowledgement

This work was carried out at the Center for Artificial Intelligence (C4AI–USP), with support by the São Paulo Research Foundation (FAPESP Grant #2019/07665-4) and by the IBM Corporation.