Classificação de Argumentos Sintácticos Aproximação preliminar

João Sequeira, Teresa Gonçalves, and Paulo Quaresma

Universidade de Évora
m5071@alunos.uevora.pt, tcg@uevora.pt, pq@uevora.pt

Resumo Este artigo apresenta uma aproximação preliminar de uma vertente pouco explorada do processamento de linguagem natural para a língua Portuguesa, a classificação de argumentos sintácticos.
Primeiro é dada uma introdução à classificação de argumentos sintácticos, posteriormente são explicados os passos necessários à criação de um classificador utilizando a ferramenta MinorThird. O desempenho foi verificado nos argumentos sintácticos mais frequentes (predicado, sujeito e complemento directo) num subconjunto do Bosque 8.0. A mesma abordagem foi aplicada a um corpus da língua Inglesa utilizado no CONLL 2004 e os resultados foram comparados com os obtidos na tarefa conjunta do CONLL 2004.

1 Introdução

Actualmente existe uma grande quantidade de conteúdos digitais de cariz académico, pessoal, noticioso entre outros disponíveis para consulta na Internet. A tarefa de obter informação de conteúdos não tratados de fontes tão dispares tornou-se praticamente impossível [14,15].
Com o incremento de conteúdos digitais publicados existiu também um aumento na pesquisa de aplicações que consigam analisar e extrair informação automaticamente dos mesmos. Este factor tem proporcionando nos últimos anos uma crescente procura de aplicações de processamento de linguagem natural[1] [4].
A classificação de argumentos sintácticos[2] tem vindo a ser uma área de cada vez mais interesse, devido à sua crescente importância em sistemas de extracção de informação, pergunta-resposta, sumarização de documentos entre outras aplicações que necessitam de informação semântica [3]. Esta vertente do processamento de linguagem natural já possui vários recursos disponíveis para línguas como o Inglês, produto de vários projectos apresentados ou implementados para conferências internacionais [3]. Mas ainda existe muita matéria a ser explorada no âmbito de outras línguas, estando o Português entre elas.
Este trabalho explora a utilização da ferramenta MinorThird[3] [5] na tarefa de

[1] Do Inglês *Natural Language Processing (NLP)*, sendo possível para o português usar também a sigla PLN
[2] Nas conferências internacionais normalmente é usado o termo *semantic role labelling*, retratando as relações semânticas entre os diferentes constituintes duma frase
[3] http://sourceforge.net/apps/trac/minorthird/wiki

classificação de argumentos sintácticos para a língua Portuguesa. Para possuir um meio de comparação com os resultados obtidos internacionalmente é feita a mesma tarefa com o corpus em Inglês usado no CONLL[4] 2004 [3].

2 Classificação de Argumentos Sintácticos

A classificação de argumentos sintácticos é actualmente um dos subgrupos mais activos na área de processamento de linguagem natural. Nos últimos 10 anos o grupo SIGNLL[5] e todos os sistemas participantes abordaram este tema nas edições 2004 [3], 2005 [4], 2008 [22] e 2009 [9] das conferencias CONLL.
Consiste em identificar os verbos presentes numa frase e os seus argumentos sintácticos [4], tais como sujeito da acção, objecto da acção entre outros.

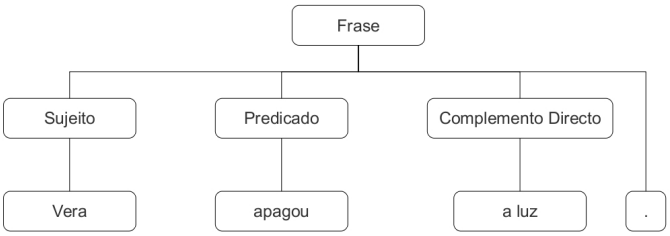

Figura 1. Frase do Bosque classificada com os respectivos argumentos sintácticos.

É visível na Figura 1 a classificação de uma frase presente no corpus usado para a língua Portuguesa, Bosque[6] [1]. Ao analisar a frase verifica-se que possui um sujeito ("Vera"), um verbo que compõe o predicado ("apagou") e um complemento directo que sofre a acção realizada pelo sujeito ("a luz").
Gildea e Jurafsky, pioneiros na classificação de argumentos sintácticos, enumeram dois métodos proeminentes para realizar a análise de textos, um baseado em gramáticas[7] e outro orientado a dados. O processo de criar gramáticas é muito moroso visto serem criadas à mão e necessitarem de incluir uma descrição para cada caso existente na língua[8]. Para os sistemas orientados a dados são necessários corpora classificados, e aplicações capazes de criar modelos a partir dos mesmos. Esses modelos são posteriormente usados para classificar textos sem marcações. Exemplos dessas aplicações são os participantes nas tarefas conjuntas[8] das edições enumeradas anteriormente das conferências CONLL [3,4].

[4] *Conference on Computational Natural Language Learning*, `http://ifarm.nl/signll/conll/`
[5] Special Interest Group on Natural Language Learning
[6] http://www.linguateca.pt/Floresta/corpus.html
[7] Do Inglês *grammar based systems*
[8] Do Inglês *Shared Task*

3 Classificador de Argumentos Sintácticos

3.1 Pré-Processamento dos Corpora

O corpus usado para criar o modelo baseia-se no Bosque 8.0[9] incorporado no projecto Floresta Sintá(c)tica[1]. A Floresta Sintá(c)tica consiste em texto corrido dividido em frases analisadas sintacticamente em estruturas de árvore pelo analisador sintáctico PALAVRAS [2] [1]. O Bosque 8.0 é composto por 9368 frases dos primeiros 1000 extractos do CETEMPúblico e do CETEMFolha, priorizando a qualidade em detrimento da quantidade [12]. Para o CETEMPúblico foram usados excertos de noticias retiradas do jornal Público [20] e para o CETEMFolha foram usados excertos de noticias retiradas do jornal Folha de S. Paulo [6].

```
'source' => 'CP429-7 Vera apagou a luz.',
'number' => 1,
'cod' => 'CETEMPúblico n=429 sec=clt sem=96a',
't' => [
  'fcl||STA',
  [
    'np||SUBJ',
    'prop(\'Vera\' F S)||H::Vera'
  ],
  [
    'vp||P',
    'v-fin(\'apagar\' PS 3S IND)||MV::apagou'
  ],
  [
    'np||ACC',
    'art(\'o\' <artd> F S)||>N::a',
    'n(\'luz\' <np-def> F S)||H::luz'
  ],
  'jjpunct(-.-)'
]
```

Figura 2. Representação da frase '`Vera apagou a luz.`' presente no Bosque.

Neste trabalho foi usado o CETEMPúblico, tendo sido removidas as frases consideradas títulos de noticias. O pré-processamento do corpus, no final do mesmo com 4416 frases, foi dividido nos seguintes passos:

1. extraiu-se as palavras e respectivas categorias sintácticas das frases. As frases estavam dispostas na forma visível na Figura 2;

[9] Obtido em: `http://www.linguateca.pt/Floresta/corpus.html`

2. converteu-se as categorias sintácticas em etiquetas XML[10], formando um conjunto de frases com a forma visível na Figura 3.

```
<SUBJ>Vera<\SUBJ> <P>apagou<\P> <ACC>a luz<\ACC>.
```

Figura 3. Representação da frase '`Vera apagou a luz.`' com etiquetas XML.

Um processamento similar foi realizado com o corpus utilizado na tarefa conjunta do CONLL de 2004[11] [3]. Foi seleccionado o corpus do CONLL 2004 visto ter sido a primeira abordagem à tarefa de classificação de argumentos sintácticos realizada nestas conferências. Como este artigo documenta uma primeira aproximação à mesma tarefa mas para o Português achou-se por bem tentar estar no mesmo patamar para obter uma melhor medição do desempenho do classificador. Este corpus foi composto por seis secções do jornal de Wall Street [7] (15 a 18 para treino (8936 frases), 20 para desenvolvimento (1671 frases) e 21 para teste (2012 frases)) presentes no Penn Treebank [17,13] ao qual foi acrescentado a informação de estruturas sintácticas de predicado-argumento presentes no PropBank [10,16].

3.2 MinorThird

O MinorThird é um conjunto, em código aberto, de classes implementadas na linguagem de programação Java para realizar tratamento de textos como por exemplo guardar, anotar e categorizar textos, aprendizagem e extracção de entidades mencionadas. Foi criado pelo professor William W. Cohen da Universidade de Carnegie Mellon e actualmente é mantido Frank Lin [5].
O MinorThird usa colecções de documentos para criar uma base de dados denominada *TextBase* sobre a qual são realizadas afirmações lógicas para posteriormente serem guardadas num objecto do tipo *TextLabels*. Como a anotação presente no objecto *TextLabels* é independente do conteúdo dos documentos podem existir vários tipos de anotações para o mesmo conjunto de documentos [5]. As anotações no *TextLabels* enumeram as categorias ou propriedades, podendo ser sintácticas ou semânticas, de uma palavra, documento ou *span*[12]. Estas anotações podem ser criadas manualmente, ou automaticamente através de uma aplicação. Os *TextLabels* e as *TextBases* a eles associadas podem ser guardados num repositório previamente configurado [5].
Os métodos de aprendizagem de extracção e classificação de *spans*, subconjuntos de palavras de um documento ou documentos inteiros, presentes no MinorThird são numerosos. Entre os métodos de aprendizagem sequencial estado-da-arte estão os campos condicionais aleatórios[13] [24,11] e métodos de treino de mo-

[10] Sigla do Inglês *Extensible Markup Language*
[11] Obtido em: `http://www.lsi.upc.edu/~srlconll/st04/st04.html`
[12] Em português enumera um conjunto de palavras
[13] Do Inglês *Conditional Random Fields*

delos escondidos de Markov[14] [21] [5]. Sendo o MinorThird uma ferramenta de aprendizagem supervisionada orientada a dados os passos do seu funcionamento estão representados na Figura 4:

- num primeiro passo é criado um modelo (*TextLabels*) utilizando ficheiros classificados (*TextBases*);
- num segundo passo esse modelo é usado para classificar textos sem marcações dando como resultado os textos classificados.

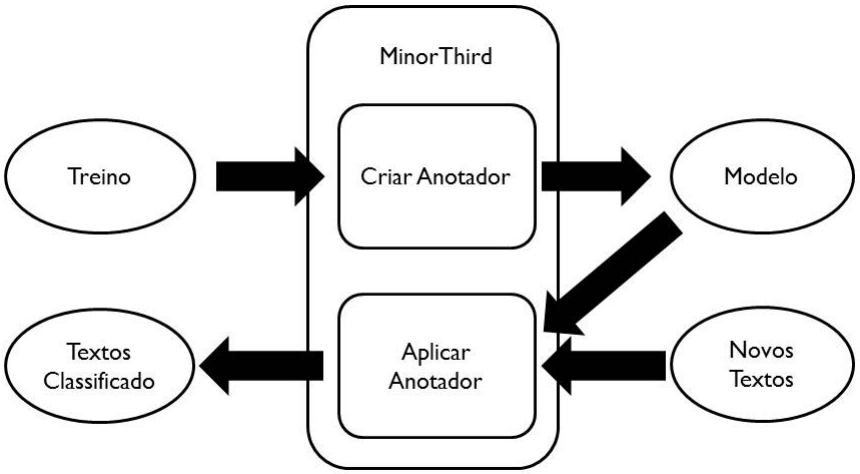

Figura 4. Funcionamento do MinorThird, criando um modelo com base em ficheiros de treino e posteriormente usando esse modelo para classificar novos textos.

3.3 Cenário Experimental

Os corpora obtidos no pré-processamento, e renomeados para uma melhor compreensão na análise dos resultados:

- BosqueXML: corpus criado com base no Bosque 8.0. Das classes presentes no Bosque 8.0 foram utilizadas no BosqueXML apenas as que forneciam viabilidade estatística para a tarefa da classificação de argumentos sintácticos.
- CONLL'2004: corpus obtido após processamento do usado na tarefa conjunta do CONLL 2004.

A Tabela 1 apresenta o número de *spans* para cada argumento sintáctico estudado em cada corpora.

[14] Do Inglês *Hidden Markov Models*

		BosqueXML	CONLL'2004	
Etiqueta	Descrição	#	# Treino	#Teste
`P`	Predicado	7268	19098	3627
`Arg0`	Sujeito	4673	12709	1671
`Arg1`	Complemento Directo	3802	18046	3429

Tabela 1. Etiquetas, descrição e contagens para os corpora BosqueXML e CONLL'2004.

Foram testados os algoritmos existentes no MinorThird sendo que os melhores resultados foram obtidos com os algoritmos SVMCMM e CRF. O algoritmo SVMCMM é constituido por modelos condicionais de Markov[15] [19,18] treinados com máquinas de vectores de suporte[16] [23]. O algoritmo CRF utiliza campos condicionais aleatórios [24,11].
Todos os testes feitos usaram as características pré-definidas de cada algoritmo com uma janela de contexto de tamanho três.

Foi aplicado o procedimento validação cruzada 10 pastas ao BosqueXML e um procedimento treino/teste ao CONN'2004. O desempenho dos modelos foi analisado com recurso à precisão (π), cobertura (ρ) e medida F_1 obtidas na classificação dos *spans*.

3.4 Resultados

A Tabela 2 mostra os resultados obtidos com os algoritmos SVMCMM e o CRF utilizando o corpus BosqueXML. Observa-se que o algoritmo CRF apresenta melhores valores de precisão enquanto que o SVMCMM apresenta melhores valores de cobertura.

Para ambos os algoritmos verifica-se que os valores da precisão estão pelo menos 0.1 acima dos da cobertura para todas as etiquetas (para o algoritmo CRF o `Arg0` e `Arg1` apresentam valores de precisão superiores em 0.2 quando comparados com a cobertura). A etiqueta *Predicado* apresenta os melhores resultados com valores de F_1 acima dos 54% enquanto que o *Complemento Directo* apresenta valores de F_1 abaixo dos 21%.

A Tabela 3 mostra os resultados obtidos com os algoritmos SVMCMM e CRF utilizando o corpus CONLL'2004.

O corpus CONLL'2004 apresenta valores de precisão e cobertura similares para ambos os algoritmos (excepto para a etiqueta `Arg0` onde o CRF tem um valor 0.1 superior na precisão). Mais uma vez o *Predicado* apresenta melhores resultados com valores de F_1 acima dos 82%, enquanto que o *Complemento Directo* tem valores de F_1 abaixo dos 24%.

Comparando a Tabela 2 e a Tabela 3 pode-se concluir que os resultados obtidos com o corpus da língua Portuguesa estão abaixo dos obtidos com o

[15] Do Inglês *Conditional Markov Models*
[16] Do Inglês *Support Vector Machines*

	SVMCMM			CRF		
Etiqueta	π	ρ	F_1	π	ρ	F_1
`P`	.603	.503	.548	.660	.475	.545
`Arg0`	.416	.283	.337	.447	.237	.308
`Arg1`	.285	.161	.206	.361	.117	.175

Tabela 2. Precisão, cobertura e F_1 do corpus BosqueXML utilizando os algoritmos SVMCMM e CRF.

	SVMCMM			CRF		
Etiqueta	π	ρ	F_1	π	ρ	F_1
`P`	0.850	0.823	0.836	0.842	0.805	0.823
`Arg0`	0.599	0.464	0.523	0.699	0.463	0.557
`Arg1`	0.372	0.170	0.234	0.414	0.151	0.221

Tabela 3. Precisão, cobertura e F_1 do corpus CONLL'2004 utilizando os algoritmo SVMCMM e CRF.

corpus da língua Inglesa. Uma possível explicação para esta diferença poderá ser o tamanho dos corpora: o CONLL'2004 é sensivelmente 3 vezes maior que o BosqueXML. Outra explicação possível é a estrutura sintáctica da língua Inglesa ser mais simples que a da língua Portuguesa.

A Tabela 4 compara os melhores valores de F_1 para as etiquetas `Arg0` and `Arg1` obtidos com o MinorThird (`Arg0` com o algoritmo CRF e `Arg1` com o algoritmo SVMCMM) com o melhor e pior obtidos na tarefa conjunta do CONLL 2004 como reportado em [3] (os valores dos *Predicados* não são mostrados visto estes não terem sido considerados na avaliação da tarefa conjunta).

		CONLL 2004	
Etiqueta	MinorThird	Melhor	Pior
`Arg0`	0.557	0.814	0.562
`Arg1`	0.234	0.716	0.490

Tabela 4. Máximo, mínimo dos valores de F_1 dos argumentos mais comuns da tarefa conjunta do CONLL 2004 comparados com os valores obtidos usando o corpus CONLL'2004 com o MinorThird.

Da Tabela 4 pode-se observar que o uso de informação linguística adicional como por exemplo categorias gramaticais, sintagmas, orações e entidades mencionadas é proveitosa para a tarefa de classificação de argumentos sintácticos enquanto que métodos sequenciais de classificação não são suficientes. O uso desta informação melhora o desempenho dos sistemas como é visível quando se compara os valores de `Arg0` e `Arg1` obtidos com o MinorThird e dos sistemas do CONLL 2004. A diferença é superior no *Complemento Directo* do que no *Sujeito*.

4 Conclusões e Trabalho Futuro

Este artigo tentou realizar uma pesquisa para a língua Portuguesa numa área ainda pouco explorada. Verificou-se que os resultados obtidos com um corpus da língua Portuguesa estão abaixo dos obtidos com um corpus da língua Inglesa. Como já foi mencionado anteriormente a diferença de desempenho pode dever-se com o diferente tamanho dos corpora. Outra possível explicação é o uso de estruturas sintácticas mais complexas e as muitas contracções de palavras existentes na língua Portuguesa quando comparada com a língua Inglesa.

Verificou-se que informação linguística tais como palavras, categorias gramaticais, sintagmas, orações e entidades mencionadas são úteis para a tarefa de classificação de argumentos sintácticos e o uso de apenas métodos de aprendizagem sequenciais não produzem bons resultados.

Como trabalho futuro temos a intenção de aumentar o tamanho do corpus da língua Portuguesa e desenvolver um classificador que faça uso de toda a informação linguística mencionada anteriormente. Apenas nesses termos uma comparação entre ambas as línguas será justa.

Referências

1. S. Afonso, E. Bick, R. Haber, and D. Santos. Florestra sintá(c)tica: A treebank for portuguese. 2002.
2. E. Bick. *The Parsing System "PALAVRAS": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, 2000.
3. X. Carreras and L. Màrquez. Introduction to the conll-2004 shared task: Semantic role labeling. In *Proceedings of CoNLL-2004*, 2004.
4. X. Carreras and L. Màrquez. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, 2005.
5. W. Cohen. Minorthird: methods for identifying names and ontological relations in text using heuristics for inducing regularities from data, 2004.
6. Empresa Folha da Manhã S.A. Folha.com. `http://www.folha.uol.com.br`, 1921.
7. Inc. Dow Jones & Company. The wall street journal. `http://europe.wsj.com/home-page`.
8. D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28:245–288, 2002.
9. J. Hajic, M. Ciaramita, R. Johansson, D. Kawahara, M. Martí, L. Márquez, A. Meyers, J. Nivre, S. Padó, J. Štěpánek, P. Straňák, M. Surdeanu, N. Xue, and Y. Zhang. The conll-2009 shared task: syntactic and semantic dependencies in multiple languages. In *CoNLL '09 Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, 2009.
10. P. Kingsbury and M. Palmer. From treebank to propbank. `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.7566`, 2002.
11. J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of 18th International Conference on Machine Learning*, pages 282–289, 2001.
12. Linguateca. Florestra sintá(c)tica, 2009.

13. M. Marcus, B. Santorini, and M. Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330, 1993.
14. N. Miranda, R. Raminhos, P. Seabra, J. Sequeira, T. Gonçalves, and P. Quaresma. Reconhecimento de entidades nomeadas com svm. In *Actas das Jornadas de Informática da Universidade de Évora 2010*, Novembro 2010.
15. N. Miranda, R. Raminhos, P. Seabra, J. Sequeira, T. Gonçalves, and P. Quaresma. Named entity recognition using machine. 2011.
16. M. Palmer, D. Gildea, and P. Kingsbury. The preposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31, 2005.
17. The Penn Treebank Project. The penn treebank project. `http://www.cis.upenn.edu/~treebank/`, 1999.
18. L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.
19. L. Rabiner and B. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, Janeiro 1986.
20. PÚBLICO Comunicação Social SA. Público. `http://www.publico.pt`, 1990.
21. M. Stamp. A revealing introduction to hidden markov models, 2004.
22. M. Surdeanu, R. Johansson, A. Meyers, L. Màrquez, and J. Nivre. The conll-2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Natural Language Learning*, pages 159–177, 2008.
23. V. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, Setembro 1998.
24. H. Wallach. Conditional random fields: An introduction. `http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.124.6711`, 2004.