POS-Tagging usando Pesquisa Local

João Laranjinho and Irene Rodrigues and Lígia Ferreira

Universidade de Évora

Abstract. Neste artigo apresenta-se um sistema de part-of-speech tagging, independente do domínio, para etiquetação gramatical de texto para o Português e Inglês.
O etiquetador usa informação morfo-sintática que vem de um dicionário local que completa a sua informação recorrendo a dicionários disponíveis na rede como o da Priberam e do LookWayUP.
Este etiquetador é baseado numa função heurística que é usada na optimização dos seus parâmetros e posterior etiquetação de texto.
Na optimização dos parâmetros da função heurística são usadas algumas das técnicas pesquisa local para reduzir o espaço de pesquisa.
Na avaliação do sistema usaram-se dois textos do corpora Reuters: testa (na fase treino) e testb (na fase de teste).

1 Introdução

Os sistemas de part-of-speech tagging classificam gramaticalmente átomos de um texto.

As formas das palavras são frequentemente ambíguas no part-of-speech tagging. Numa expressão, essas ambiguidades normalmente são resolvidas pelo contexto das palavras.

Os sistema de part-of-speech tagging dividem-se em dois grupos: baseados em regras e estocásticos.

Para o inglês, alguns dos sistemas actuais conseguem um valor que ronda os 96-97% de precisão.

Apresentamos um sistema automático independente do domínio para etiquetação gramatical de texto. Na etiquetação é usada informação morfo-sintáctica de dicionários que se encontram na WEB. Para reduzir o espaço de pesquisa são usadas algumas técnicas de pesquisa local.

O desempenho de um sistema de part-of-speech tagging pode ser medido com diversão métricas que representam o desempenho em valores numéricos.

As três métricas que normalmente são utilizadas para avaliar o desempenho são as seguintes: Abrangência (Recall), Precisão (Precision) e Medida-F (F-Measure).

- A Abrangência mede a relação entre o número de resultados correctos e o número de resultados existentes. A fórmula da Abrangência é a seguinte:
  $\text{Abrangência} = \frac{Resultados\ \ Correctos\ \cap\ Resultados\ \ Existentes}{Resultados\ \ Existentes}$
- A Precisão mede a relação entre o número de resultados correctos e o número de resultados obtidos. A fórmula da Precisão é a seguinte:
  $\text{Precisão} = \frac{Resultados\ \ Correctos\ \cap\ Resultados\ \ Obtidos}{Resultados\ \ Obtidos}$
- A Medida-F é uma métrica harmónica de Precisão (P) e Abrangência (A). A fórmula da *Medida-F* é a seguinte:
  $\text{Medida-F} = 2^* \frac{P*A}{P+A}$

2 Arquitectura do Etiquetador

O etiquetador contém duas etapas: optimização e etiquetação. Na figura 1 são apresentados os módulos de optimização e na figura 2 são apresentados os módulos de etiquetação.

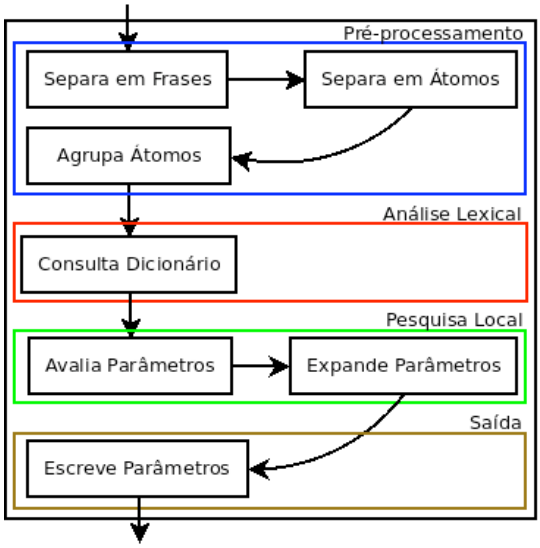

Fig. 1. Arquitectura do Optimizador

2.1 Optimização

A etapa de optimização contém os seguintes módulos: pré-processamento, análise lexical, pesquisa local e saída.

No pré-processamento separa-se o texto em frases e as frases em átomos. As frases são constituídas por átomos e os átomos por sequências de caracteres. Ainda no pré-processamento os átomos são agrupados em tripos para serem usados na função heurística.

Na análise lexical consulta-se em dicionários *on-line* a informação morfo-sintática das palavras que não se encontram no dicionário local, guardando-se essa informação no dicionário local.

Na pesquisa local geram-se conjuntos de parâmetros iniciais, que posteriormente serão avaliados. Quando um conjunto de parâmetros contém outros conjuntos vizinhos com valor de heurística superior, expandem-se os vizinhos e em seguida avaliam-se. A avaliação termina quando não são encontrados mais vizinhos com valor de heurística superior ou um critério de paragem ter sido alcançado.

Finalmente na saída transcreve-se para um ficheiro o conjunto de parâmetros que obteve o valor mais alto de heurística.

2.2 Etiquetação

A etapa de etiquetação contém os seguintes módulos: pré-processamento, análise lexical, avaliação e saída.

No pré-processamento separa-se o texto em frases e as frases em átomos. As frases são constituídas por átomos e os átomos por sequências de caracteres. Ainda no pré-processamento os átomos são agrupados em tripos para serem usados na função heurística.

Na análise lexical consulta-se em dicionários *on-line* a informação morfo-sintática das palavras que não se encontram no dicionário local, guardando-se essa informação no dicionário local.

Na avaliação são atribuídas classes gramaticais a cada átomo através de uma função que usa os parâmetros apurados na fase de optimização.

Finalmente na saída escreve-se num ficheiro para cada átomo a categoria correspondente.

3 Corpus

Nos testes com o etiquetador usamos os seguintes 2 ficheiro do corpus da Reuters: testa (na fase de treino) e testb (na fase de teste).

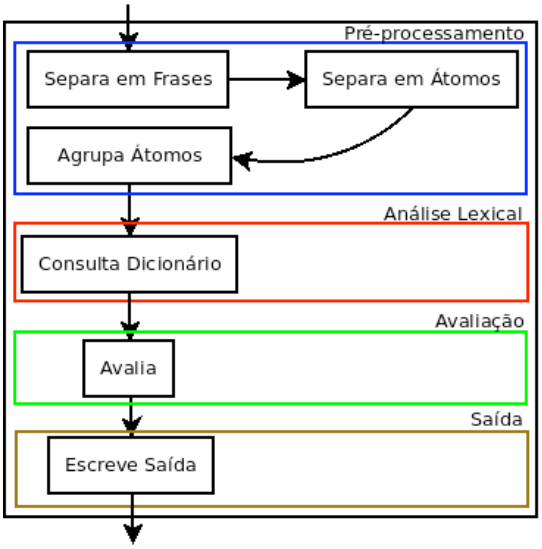

Fig. 2. Arquitectura do Etiquetador

Para conhecermos um pouco melhor o corpus fizemos testes com: átomos ambíguos, átomos sem contraditórios e todos os átomos. A distribuição dos átomos encontra-se na tabela 1

	testa	testb
Todos	51.360	46.435
Ambíguos	24.144	–
Sem Contraditório	45.184	–

Table 1. Distribuição dos átomos no Corpus

Os átomos ambíguos são átomos que se encontram no dicionário com entrada em mais que uma classe gramatical. Os átomos contraditórios são átomos com iguais características no dicionário que ocorrem no corpus com diferentes classificações.

4 Função de Avaliação

Na função de avaliação estuda-se o impacto das classes gramaticais na etiquetação gramatical de texto. O estudo inclui informação sobre: átomo anterior, átomo em análise e átomo seguinte.

No sistema de etiquetação de classes gramaticais são usadas as seguintes 20 classes gramaticais: ADJ, ADV, CONJ, DET, EX, FW, MOD, N, NP, NUM, PRO, P, TO, UH, V, VD, VG, VN, WH e SYM.

A função heurística usada é a seguinte:

$F(A-1,A,A+1) = P1*ADJ(A)+P2*ADV(A)+P3*CONJ(A)+P4*DET(A)+P5*EX(A)+P6*FW(A)+P7*MOD(A)+P8*N(A)+P9*NP(A)+P10*NUM(A)+P11*PRO(A)+P12*P(A)+P13*TO(A)+P14*UH(A)+P15*V(A)+P16*VD(A)+P17*VG(A)+P18*VN(A)+P19*WH(A)+P20*SYW(A)+P21*ADJ(A-1)+P22*ADV(A-1)+P23*CONJ(A-1)+P24*DET(A-1)+P25*EX(A-1)+P26*FW(A-1)+P27*MOD(A-1)+P28*N(A-1)+P29*NP(A-1)+P30*NUM(A-1)+P31*PRO(A-1)+P32*P(A-1)+P33*TO(A-1)+P34*UH(A-1)+P35*V(A-1)+P36*VD(A-1)+P37*VG(A-1)+P38*VN(A-1)+P39*WH(A-1)+P40*SYW(A-1)+P41*ADJ(A+1)+P42*ADV(A+1)+P43*CONJ(A+1)+P44*DET(A+1)+$

$P45*EX(A+1)+P46*FW(A+1)+P47*MOD(A+1)+P48*N(A+1)+P49*NP(A+1)+P50*NUM(A+1)+P51*PRO(A+1)+P52*P(A+1)+P53*TO(A+1)+P54*UH(A+1)+P55*V(A+1)+P56*VD(A+1)+P57*VG(A+1)+P58*VN(A+1)+P59*WH(A+1)+P60*SYW(A+1)$

Na função heurística *A-1*, *A* e *A+1*, representam átomo anterior, átomo em análise e átomo seguinte.

5 Avaliação

Num dos testes fizemos 3 experiências nas quais apuramos os parâmetros usando a função heurística de forma isolada para cada classe gramatical com as seguintes informações do ficheiro testa: átomos ambíguos, átomos sem contraditórios e todos os átomos. Posteriormente com os parâmetros encontrados foi feita etiquetação do ficheiro testb, os resultados encontram-se na tabela 2.

	Ambíguos			Sem contraditórios			Todos		
CAT	PREC	COB	MED-F	PREC	COB	MED-F	PREC	COB	MED-F
ADJ	0,6899	0,7434	0,7157	0,6878	0,7934	0,7368	0,6909	0,8217	0,7507
ADV	0,7661	0,2161	0,3371	0,8461	0,6358	0,7260	0,8761	0,6424	0,7413
CONJ	0,9793	0,5569	0,7100	0,9922	0,9935	0,9928	0,9961	0,9908	0,9934
DET	0,9846	0,9825	0,9836	0,9774	0,9872	0,9823	0,9847	0,9882	0,9865
EX	0,8889	0,9412	0,9143	0,9655	0,8235	0,8889	0,8857	0,9118	0,8986
FW	1,0	0,0000	0,0000	1,0	0,0000	0,0000	1,0	0,0000	0,0000
MOD	0,9431	0,9888	0,9654	0,9462	0,9851	0,9653	0,9336	0,9963	0,9639
N	0,8019	0,8356	0,8184	0,7602	0,8750	0,8136	0,7713	0,8688	0,8172
NP	0,8023	0,5702	0,6666	0,9407	0,6032	0,7351	0,8891	0,6615	0,7586
NUM	0,9792	0,9890	0,9840	0,9816	0,9990	0,9902	0,9798	0,9863	0,9830
PRO	0,9965	0,9567	0,9762	0,9900	0,9867	0,9883	0,9955	0,9767	0,9860
P	0,9260	0,9835	0,9539	0,9284	0,9773	0,9522	0,9265	0,9766	0,9509
TO	1,0	0,4315	0,6029	1,0	0,9963	0,9982	1,0	0,9988	0,9994
UH	0,5556	0,7143	0,6250	0,8333	0,7143	0,7692	0,8000	0,5714	0,6667
V	0,9136	0,7305	0,8118	0,8828	0,7848	0,8309	0,8994	0,8001	0,8469
VD	0,8795	0,8546	0,8669	0,8769	0,9141	0,8951	0,8879	0,9182	0,9028
VG	0,8488	0,6033	0,7053	0,8229	0,9215	0,8694	0,8550	0,9256	0,8889
VN	0,8318	0,7136	0,7682	0,8411	0,7760	0,8072	0,8739	0,7206	0,7899
WH	0,9464	0,8689	0,9060	0,9579	0,9705	0,9642	0,9581	0,9738	0,9659
SYM	0,9854	0,0690	0,1289	0,9924	0,9990	0,9957	0,9929	0,9973	0,9951

Table 2. Etiquetação isolada usando no treino átomos ambíguos, átomos sem contraditórios e todos os átomos

Num outro teste fizemos também outras 3 experiências na quais etiquetamos o ficheiro testb usando os parâmetros encontrados de forma isolada para cada uma das classes gramaticais com as seguintes informações do ficheiro testa: átomos ambíguos, átomos sem contraditórios e todos os átomos. Neste teste escolhemos para cada átomo a classe gramatical que obteve o valor mais alto de heurística. Os resultados alcançados com os átomos ambíguos, átomos sem contraditórios e todos os átomos, foram respectivamente 0.8348, 0.8522 e 0.8598.

Finalmente num outro teste fizemos 3 experiências nas quais marcamos: adjectivos, substantivos e nomes próprios. No ficheiro de treino testa estas classes são aquelas que têm maiores frequências de átomos e onde a etiquetação teve menor desempenho. Neste teste etiquetamos sucessivamente as classes que obtiveram maiores percentagens de erro durante 3 iterações. Os ganhos na etiquetação das classes adjectivo, substantivo e nome próprio, foram respectivamente 0.0746, 0.0673 e 0.0708.

		MARCAÇÃO		
		ADJ	N	NP
% DE ERRO	ADJ	–	0,0285	0,0160
	ADV	0,0424	0,0030	0,0037
	CONJ	0,0	0,0	0,0017
	DET	0,0030	0,0007	0,0075
	EX	0,0	0,0	0,0
	FW	0,0	0,0	0,0020
	MOD	0,0	0,0019	0,0018
	N	0,0811	–	0,0335
	NP	0,1436	0,1327	–
	NUM	0,0089	0,0064	0,0292
	PRO	0,0	0,0	0,0023
	P	0,0050	0,0004	0,0040
	TO	0,0	0,0	0,0017
	UH	0,0	0,0	0,0006
	V	0,0026	0,0440	0,0021
	VD	0,0073	0,0037	0,0002
	VG	0,0083	0,0064	0,0005
	VN	0,0069	0,0010	0,0020
	WH	0,0	0,0	0,0003
	SYM	0,0	0,0	0,0018

Table 3. Percentagem de erros na marcação de adjectivos, substantivos e nomes próprios

6 Conclusão e Trabalho Futuro

Na experiência em que usamos os parâmetros que foram apurados com todos os átomos do ficheiro testa conseguimos melhores resultados. No entanto, a diferença em relação à experiência na qual usamos os átomos sem contraditórios não foi significativa (não chegou a 0.01 de erro). Já na experiência na qual usamos apenas átomos ambíguos existiu uma perda de cerca de 0.02.

Etiquetar sucessivamente as classes que obtêm maior percentagem de erro permite alcançar ganhos no desempenho do sistema.

Como trabalho futuro pensamos fazer estudos nos quais:

- retiramos os contraditório com menores frequências;
- retiramos átomos não ambíguos;
- marcamos sucessivamente as classes que obtêm maior percentagem de erro até não existir perda no desempenho;
- etiquetamos nomes próprios antes de etiquetar todas as outras classes gramaticais;
- adicionamos informação de dois ou mais átomos anteriores e de dois ou mais átomos seguintes ao átomo em análise;