7
Resultados

Nessa seção serão apresentados os índices de processamento do sistema desenvolvido segundo a Avaliação Internacional de Sistemas de Identificação de Entidades Mencionadas (HAREM) promovido pela Linguateca e depois uma comparação dos resultados do algoritmo de classificação da seção 6.1.1 com a abordagem tradicional de *bag-of-words* e utilizando *bag-of-lexems*.

7.1.
Pré-processamento

Para avaliar a qualidade do pré-processamento de textos, a medida mais utilizada na literatura de recuperação de informações é a medida F (van Rijsbergen, C. J., 1979) que combina a precisão com que o sistema extraiu as entidades, e abrangência do mesmo procedimento:

$$F_1(r,p) = \frac{2rp}{r+p}$$

onde p é a precisão e r a abrangência do resultado do algoritmo.

O HAREM funciona da seguinte forma: os sistemas participantes recebem uma coleção de textos no tamanho de aproximadamente 2000 páginas e devem efetuar a extração e classificação das entidades em 48 horas e devolver a base etiquetada. Depois de entregue essas bases são comparadas com o gabarito "Coleção Dourada". O gabarito consiste na mesma base textual etiquetada manualmente por 5 pessoas.

A medida-F é aplicada a subconjuntos da coleção de textos relacionados a diferentes categorias. Cada categoria avaliada é, então, uma seleção de textos sob um determinado critério. Os critérios são língua escrita (português brasileiro ou de Portugal), gênero (notícia, livro) e veículo (email, internet).

Na primeira coluna, F significa a avaliação do sistema segundo a medida-F, a segunda coluna "Saída" é o nome/instituição do sistema desenvolvido. A terceira coluna é a avaliação segundo apenas a precisão so sistema (utilizada no cálculo da medida-F), a terceira coluna também mostra o cáculo da abrangência utilizada na medida-F. A "Sobre-geração" é o percentual de entidades extraídas a

mais que o gabarito (Coleção Dourada) e a "Sub-geração" é a quantidade de entidades do gabarito não reconhecidas pelo sistema.

A Tabela 4 mostra, então, os resultados dos sistemas participantes do HAREM ordenados pela medida-F.

Tabela 4 – Resultado do Reconhecimento de Entidades Mencionadas para os textos da coleção escritos no Português Brasileiro.

Português Brasileiro

F	Saída	Precisão	Abrangência	Combinado	Sobre-geração	Sub-geração
0,78	iémen	75,89	80,38	0,31	0,09	0,09
0,78	mascate	72,84	83,00	0,31	0,12	0,04
0,78	asmara	72,88	82,82	0,31	0,12	0,04
0,77	nicósia	72,33	82,16	0,32	0,12	0,05
0,77	cairo	72,32	82,03	0,32	0,12	0,05
0,72	**PUC-Rio**	**61,02**	**86,60**	**0,40**	**0,27**	**0,02**
0,69	riad	71,08	67,60	0,42	0,14	0,21
0,64	damasco	75,51	55,50	0,49	0,04	0,34
0,60	doha	55,76	65,08	0,53	0,25	0,15
0,58	amã	51,96	65,08	0,55	0,30	0,15
0,58	abudhabi	52,31	64,12	0,56	0,30	0,16
0,56	bengazi	46,19	71,85	0,56	0,23	0,08
0,53	kuwait	49,43	57,65	0,58	0,13	0,17
0,52	oman	48,60	56,89	0,59	0,14	0,17
0,51	teerão	79,67	37,81	0,63	0,04	0,55
0,26	qatar	64,01	16,16	0,85	0,06	0,79
0,18	eritreia	50,03	11,23	0,89	0,10	0,82
0,11	túnis	46,52	6,46	0,94	0,13	0,88
0,10	dakar	43,43	5,55	0,95	0,13	0,89

A Tabela 4 mostra que o sistema desenvolvido ficou em sexto lugar para o processamento do português brasileiro. Porém, apresentou a maior abrangência, e, levando em conta que a precisão foi prejudicada por 0,27% de sobre-geração, podemos reavaliar melhor a medida F. A sobre-geração foi por conta de algumas entidades que o sistema extraía sem ter letra maiúscula, uma regra da avaliação. Vale ressaltar, também, que a medida abrangência vale por si só, pois expressa a eficiência do reconhecimento de entidades.

A Tabela 5 mostra as avaliações dos sistemas para textos coletados de bases de emails. O sistema implementado nessa tese obteve a melhor pontuação. A Tabela 6 mostra o resultado da avaliação para a categoria de notícias de jornais. De forma coerente, o sistema obteve o mais alto índice de abrangência (91,73%), isto é, reconheceu o maior número de entidades corretamente. Isso ocorreu pelo fato de a base de treinamento do sistema ter sido justamente notícias provenientes de jornais brasileiros.

Tabela 5 – Resultado HAREM para a avaliação de bases de emails.

Base de Emails

F	Saída	Precisão	Abrangência	Combinado	Sobre-geração	Sub-geração
0,66	**PUC-Rio**	**57,72**	**76,56**	**0,45**	**0,18**	**0,03**
0,64	meca	55,99	73,65	0,48	0,18	0,04
0,64	asmara	55,89	73,65	0,48	0,18	0,04
0,63	riad	55,49	73,25	0,48	0,18	0,04
0,63	iémen	55,49	73,25	0,48	0,18	0,04
0,60	eritreia	54,89	66,20	0,51	0,17	0,12
0,58	amã	58,40	58,54	0,53	0,14	0,21
0,56	doha	52,92	60,22	0,56	0,17	0,11
0,56	ancara	52,41	60,22	0,56	0,18	0,11
0,56	tripoli	52,41	59,64	0,56	0,18	0,11
0,49	luxor	57,34	42,24	0,63	0,06	0,39
0,43	bagdad	33,90	59,69	0,68	0,16	0,09
0,43	cairo	36,53	51,63	0,69	0,20	0,18
0,43	mascate	36,41	51,60	0,69	0,19	0,18
0,39	túnis	61,07	28,43	0,74	0,11	0,60
0,26	manama	55,87	16,89	0,84	0,00	0,77
0,08	bahrein	28,22	4,42	0,96	0,08	0,88
0,05	damasco	30,43	2,68	0,97	0,05	0,93
0,05	kuwait	29,63	2,68	0,97	0,08	0,93

A abrangência de 91,73% é bastante significante, visto que o maior torneio de extração de entidades (MUC/ACE realizado nos EUA para a lingua inglesa) tem as maiores taxas de acerto por volta de 90%. (Weiss 2005)

Finalmente, na Tabela 7, é mostrada a avaliação para páginas na internet, também com um bom desempenho.

Além dos sistemas participantes do HAREM, existem sistemas para a língua inglesa que podem ser adaptados para a língua portuguesa, no entanto, apresentam resultados inferiores, o que pode ser melhorado com uma boa implementação. Bons exemplos são Docyoument, Penn Treebank, PreText, OpenNLP, Gate.

Quanto ao tempo de processamento, a implementação proposta nessa tese processa duas mil notícias em um minuto. O tradicional pré-processamento *bag-of-words* demora sete segundos nas mesmas condições. Já um pré-processamento *bag-of-lexems* feito por humanos (mesma base) leva horas ou até dias.

Normalmente, o erro humano, muito por causa do cansaço, atinge índices maiores que o de processadores automáticos. Para contornar esse problema, são utilizadas mais de uma pessoa para etiquetar a mesma e toda a base. Ao final utiliza-se um procedimento de alinhamento da case etiquetada, o que acaba demorando mais do que o processamento em questão.

Tabela 6 – Resultado HAREM para a avaliação de bases de textos de jornais.

Base de Textos de Jornais

F	Saída	Precisão	Abrangência	Combinado	Sobre-geração	Sub-geração
0,89	doha	90,08	88,92	0,15	0,02	0,05
0,89	bahrein	89,84	89,02	0,15	0,02	0,05
0,88	argel	89,88	86,90	0,16	0,01	0,06
0,88	rabat	89,87	86,81	0,16	0,01	0,06
0,88	marraquexe	87,70	88,11	0,17	0,03	0,05
0,83	**PUC-Rio**	**76,20**	**91,73**	**0,25**	**0,16**	**0,02**
0,83	luxor	90,85	76,44	0,26	0,02	0,19
0,70	manama	72,59	67,18	0,41	0,09	0,18
0,69	dakar	72,31	66,86	0,42	0,10	0,18
0,69	damasco	71,69	67,18	0,42	0,10	0,18
0,68	abudhabi	61,51	75,52	0,42	0,06	0,08
0,67	nicósia	84,04	55,87	0,46	0,00	0,36
0,55	sana	55,02	54,92	0,56	0,05	0,19
0,55	iémen	81,73	41,31	0,60	0,01	0,51
0,54	mascate	54,44	53,99	0,56	0,05	0,18
0,24	oman	74,77	14,44	0,86	0,00	0,82
0,21	qatar	50,51	13,12	0,87	0,07	0,78
0,11	amã	46,64	6,51	0,94	0,11	0,88
0,11	túnis	46,64	6,51	0,94	0,11	0,88

Tabela 7 – Resultado HAREM para a avaliação de bases de páginas da Internet.

Base de páginas da Internet

F	Saída	Precisão	Abrangência	Combinado	Sobre-geração	Sub-geração
0,74	abudhabi	69,63	79,99	0,35	0,10	0,03
0,74	rabat	69,68	79,84	0,35	0,10	0,03
0,74	teer	69,18	80,00	0,35	0,10	0,03
0,74	manama	69,23	79,85	0,35	0,10	0,03
0,73	argel	68,56	77,51	0,38	0,12	0,09
0,73	**PUC-Rio**	**65,39**	**81,99**	**0,37**	**0,16**	**0,02**
0,64	bagdad	69,31	60,11	0,48	0,11	0,26
0,64	casablanca	70,89	58,93	0,48	0,05	0,27
0,59	dakar	56,00	61,85	0,53	0,18	0,13
0,59	marraquexe	55,78	61,85	0,53	0,19	0,13
0,58	luxor	55,79	61,04	0,54	0,18	0,14
0,51	meca	40,81	67,87	0,61	0,18	0,07
0,46	eritreia	40,74	51,56	0,66	0,14	0,17
0,44	riad	71,76	31,79	0,70	0,04	0,58
0,44	bahrein	39,53	49,10	0,67	0,15	0,19
0,26	kuwait	55,03	16,75	0,84	0,02	0,75
0,17	mascate	42,77	10,92	0,90	0,07	0,78
0,12	bengazi	37,99	7,27	0,93	0,06	0,83
0,12	oman	37,60	7,12	0,93	0,06	0,83

7.2. Classificação

O algoritmo de classificação da seção 2.1.3 – Análise de Discriminante – foi aplicado em duas bases de notícias de um jornal da Internet. Para formar a base foi selecionada a seção de agronegócios do jornal para a classificação alvo e o universo de notícias erradas foi formado pelas outras partes do jornal, que não

tinham notícias de agronegócios. A aplicação na menor base teve o objetivo de testar o desempenho dos métodos com pouca quantidade de textos e também a interpretabilidade dos parâmetros do modelo. A aplicação na maior base teve o objetivo de testar a habilidade de gerenciar grandes volumes de textos, assim como a capacidade de generalização do modelo ajustado.

Primeiramente foi ajustado o modelo tradicional de bag-of-words, depois o método bag-of-lexems e finalmente o método cheio, bag-of-lexems mais a ontologia associada. O procedimento bag-of-lexems sem ontologia foi realizada apenas para se ter uma noção de interpolação dos resultados, pois o comparativo final será do método bag-of-words com o método bag-of-lexems rotulados.

7.2.1. Amostra Pequena

Para formar a base foi selecionada da seção de agronegócios do jornal um total de 100 notícias. Das outras partes do jornal, que não tinham notícias de agronegócios, foram coletadas mais 4900 notícias dando um total de 5000 documentos.

Seguindo a sequência de análise, para a fase de treinamento, o modelo tradicional de pré-processamento *bag-of-words* foi processado e apresentou os resultados daTabela 8. Foram executados o mesmo procedimento para diferentes quantidades de informação (QIP), dado que processar todos os termos da base é inviável computacionalmente. As diferentes quantidades de entrada de informação demandam um tempo cada vez maior de processamento de máquina. Por esse motivo foram feitos cinco formatos diferentes cortando a lista de termos em 10, 20, 30, 40 e 50% do seu tamanho original.

Tabela 8 – Resultados do modelo bag-of-words para a fase de treinamento com 5000 notícias. A tabela apresenta os dados do modelo e o quadro apresenta o formato de leitura dos parâmetros do modelo.

bag-of-words

QIP*	Acerto	Erro	Termos
10%	45	6	41
20%	59	6	157
30%	64	6	279
40%	68	6	362
50%	72	6	487

* QIP: Quantidade de Informação Processada

Trecho da saída:

```
...
termo:agricultores AND text:carvalhaes
termo:agroeconômica
termo:böel
termo:cafeicultura
termo:camardelli
termo:cogo
termo:coopercentral
termo:culturas AND text:fertilizantes AND text:importaram
termo:cutrale
...
```

O mesmo procedimento, com excessão do número de termos ajustados, foi realizado para o método *bag-of-lexems*. Essa etapa foi realizada com objetivos apenas de interpolação do índice de acerto. Os resultados desse modelo são apresentados na Tabela 9.

Tabela 9 – Resultados do modelo bag-of-lexems para a fase de treinamento com 5000 notícias. A tabela apresenta os dados do modelo e o quadro apresenta o formato de leitura dos parâmetros do modelo.

Bag-of-Lexems

QIP*	Acerto	Erro	Termos
10%	64	7	--
20%	78	7	--
30%	83	8	--
40%	85	6	--
50%	88	6	--

Trecho da Saída

```
...
termo:agropecuária sachetti ltda
termo:agência paulista de tecnologia do agronegócio
termo:antonio camardelli
termo:cafeicultura
termo:carlos cogo
...
```

Finalmente, na Tabela 10, o resultados do modelo cheio, com a estratégia bag-of-lexems mais a classificação ontológica.

Tabela 10 – Resultados do modelo bag-of-lexems com ontologia para a fase de treinamento com 5000 notícias. A tabela apresenta os dados do modelo e o quadro apresenta o formato de leitura dos parâmetros do modelo.

Bag-Of-Lexems e Ontologia

QIP*	Acerto	Erro	Termos
10%	66	7	149
20%	81	7	324
30%	87	6	542
40%	90	6	779
50%	92	6	1028

Trecho da Saída

```
...
local:iowa
empresa:agropecuária sachetti ltda
empresa:agência paulista de tecnologia do agronegócio
empresa:cogo consultoria agroeconômica
empresa:conselho nacional do café
empresa:implementos associados
empresa:ministério da agricultura, pecuária e abastecimento
 AND substantivo:saca
empresa:scot consultoria
pessoa:antonio camardelli
pessoa:carlos cogo
...
```

Os resultados de todos os métodos da fase de treinamento foram consolidados na Tabela 11.

Tabela 11 – Comparação entre as acurácias dos tres métodos: BOW (bag-of-words), BOL(bag-of-lexems) e BOLO(bag-of-lexems com ontologia), para as diferentes quantidades de informação processada.

Comparativo de Acurácia

QIP*	BOW	BOL	BOLO
10%	39%	57%	59%
20%	53%	71%	74%
30%	58%	75%	81%
40%	62%	79%	84%
50%	66%	82%	86%

Tabela 12 – Comparativo do número de termos significantes usados no modelo de classificação.

Comparativo de Termos

BOW	BOLO
41	149
157	324
279	542
362	779
487	1028

Pela melhora nos resultados com o método BOLO, reafirmamos a tese de que o pré-processamento dos textos funciona como um atalho eficiente para o cálculo da matriz de covariância entre os termos.

Analisando os trechos das saídas Tabela 8, Tabela 9 e Tabela 10 pudemos notar a clareza semântica do modelo ajustado. No primeiro trecho temos como relevante o termo “cogo”, que a princípio não dá para saber o que é. No segundo, aparece o termo “carlos cogo” que permite a inferencia de ser o nome de uma pessoa. No último trecho, são apresentados dois termos com “cogo”, um que é

uma pessoa realmente "pessoa:carlos cogo" e outro que é a empresa dele "empresa: cogo consultoria agroeconômica".

Apesar de o número de termos significantes para o método BOLO ter sido maior, o que indicaria menor parsimônia, podemos ver, que para a mesma acurácia de 58% aproximadamente, o ajuste BOW usou 279 termos contra 149 do ajuste BOLO. Nesse caso o modelo BOLO apresenta melhor acurácia, melhor interpretabilidade dos resultados e melhor performance.

7.2.2. Amostra Grande

O algoritmo de classificação foi aplicado em um volume maior de textos para testar a performance em grandes volumes de textos, ambição de qualquer algoritmo de mineração de textos. O objetivo dessa aplicação é, além de testar a habilidade com grandes volumes, testar a capacidade de generalização dos métodos.

A base textual continha um total de 37049 notícias das quais 1300 notícias eram da seção de agronegócios e 35749 de outras seções do jornal. Das 1300 foram separadas 1000 notícias para a fase de treinamento (*in-sample*) ou ajuste dos parâmetros do modelo e 300 para a fase de teste de generalização (*out-of-sample*). Desse modo, sobraram 35749 notícias para compor a base de textos errados.

Os resultados para o método *bag-of-words* se encontram na Tabela 13, os resultados para o modelo *bag-of-lexems* na Tabela 14 e o resultado para o modelo *bag-of-lexems* rotulados pela ontologia na Tabela 15.

Tabela 13 – Resultados de performance e generalização utilizando bag-of-words.

bag-of-words

QIP*	Acerto	Erro	Termos	Tempo (seg)	Generalização
10%	848	2118	876	30	224
20%	860	1317	1103	113	211
30%	872	1252	1508	268	215
40%	872	1283	1991	323	201

Tabela 14 – Resultados de performance e generalização utilizando bag-of-lexems.

bag-of-lexems

QIP*	Acerto	Erro	Termos	Tempo (seg)	Generalização
10%	771	1578	915	**3**	119
20%	840	1535	1462	12	155
30%	858	619	1724	30	166
40%	875	616	2201	57	153

Tabela 15 – Resultados de performance e generalização utilizando bag-of-lexems rotulados usando a ontologia definida.

bag-of-lexem e Ontologia

QIP*	Acerto	Erro	Termos	Tempo (seg)	Generalização
10%	873	1375	**808**	3	168
20%	903	1303	1302	20	195
30%	919	**490**	1545	45	**234**
40%	**929**	517	2044	71	230

Para comparar, escolhemos o melhor modelo de BOW que, segundo os resultados, é o que utiliza 30% de QIP, e o melhor de BOLO que é o de, também, 30%. No critério Acerto verificamos 872 x 929 indicando superioridade de BOLO, no critério Erro, verficamos 1283 x 517 indicando superioridade significativa de BOLO, no critério Quantidade de Termos verificamos 1991 x 2044 indicando uma leve superioridade de BOW, no critério Tempo de Processamento verificamos 323 x 71 segundos indicando superioridade significativa de BOLO e, por fim, no critério generalização 201 x 230 indicando superioridade de BOLO.