JHI JOURNAL OF HEALTH INFORMATICS

Corpora Analysis: Journalistic and Scientific

Análise de Corpora: Jornalístico e Científico

Análisis Corpora: Periodístico y Científico

José Marcio Duarte[1], Kelsy Areco[1], Samuel Goihman[2], Edvane Birelo Lopes de Domenico[3], Felipe Mancini[4]

Keywords: Natural Language Processing; Medical Informatics; Information Science

ABSTRACT

Objective: This study aimed to compare two *Corpora*, one obtained from compiled newspapers – Journalistic *Corpus*, and the other from scientific papers – Scientific *Corpus*, with the hypothesis that the Scientific *Corpus* is more appropriated to Part-of-Speech information extraction in scientific similar texts. The aims were to analyze differences and similarities through: accuracy measurement; descriptive analysis; and independence of components in the *Corpora*. **Methods:** The analysis consisted on three steps: Descriptive Analysis; Accuracy Assessment; and Pointwise Mutual Information - PMI. **Results:** There was an important difference between words that do not match in both *Corpora*. The Scientific *Corpus* (92.95%) accuracy assessment was higher than Newspaper *Corpus* (88.32%). The PMI calculations for the bigrams of Newspaper and Scientific *Corpora* did not show statistically significant difference. **Conclusion:** The experiments carried out lead us to conclude that in order to extract PoS information with accuracy a better performance resulted with the association of scientific text with its specific *Corpus* and not a generic one, like Newspaper *Corpus*.

Descritores: Processamento de Linguagem Natural; Informática Médica; Ciência da Informação

RESUMO

Objetivo: Este trabalho realiza uma comparação entre um *Corpus* Jornalístico e um *Corpus* Científico. Queremos verificar se a especificidade do *Corpus* é mais adequada para extração de informação Part-of-Speech (PoS) em textos similares. Os objetivos são: analisar diferenças e similaridades por medida de acurácia; análise descritiva; e independência dos componentes nos *Corpora*. **Método:** A análise constituiu-se em três etapas: Análise descritiva; Avaliação da acurácia; e *Pointwise Mutual Information* - PMI. **Resultados:** Existe uma diferença entre as palavras que não coincidem nos *Corpora*. A avaliação da acurácia do Corpus Científico (92.95%) resultou em um valor maior comparado com o Corpus Jornalístico (88.32%). Os cálculos do PMI para os bigramas dos *Corpora* Jornalístico e Científico demonstraram não existir uma diferença estatisticamente significante. **Conclusão:** Os experimentos realizados nos levaram a concluir que para extrair, com acuracidade, informação PoS em textos é necessária associação com um *Corpus* de maior especificidade e não um genérico, como o jornalístico.

Descriptores: Procesamiento de Lenguaje Natural; Informática Médica; Ciencia de la Información

RESUMEN

Objetivo: Realizar una comparación entre un *Corpus* Periodístico y un *Corpus* Científico. Queremos verificar si la especificidad del *Corpus* es más adecuada para la extracción de información *Part-of-Speech* (PoS) en textos similares. Los objetivos: analizar diferencias y similitudes por medida de exactitud; análisis descriptivo; e independencia de los componentes. **Método:** El análisis se constituyó en tres etapas: Análisis descriptivo; Evaluación de la exactitud; Pointwise Mutual Information (PMI). **Resultados:** Hay una diferencia entre las palabras que no coinciden en los *Corpora*. La exactitud del modelo científico (92.95%) fue mayor que la exactitud del modelo periodístico (88.32%). Los cálculos del PMI para los bigramas de los *Corpora* Periodístico y Científico demostraron no tener una diferencia estadísticamente significativa. **Conclusión:** Los experimentos realizados nos llevaron a concluir que para extraer, con exactitud, información PoS en un texto demanda asociación con *Corpus* de mayor especificidad y no un *Corpus* genérico como el Periodístico.

[1] *Mestre em Ciências, Departamento de Informática em Saúde - DIS, Universidade Federal de São Paulo - UNIFESP, São Paulo (SP), Brasil.*
[2] *Professor Associado, Departamento de Informática em Saúde - DIS, Universidade Federal de São Paulo - UNIFESP, São Paulo (SP), Brasil.*
[3] *Professor Adjunto, Escola Paulista de Enfermagem, Universidade Federal de São Paulo - UNIFESP, São Paulo (SP), Brasil.*
[4] *Professor Adjunto, Universidade Aberta do Brasil - UAB, Universidade Federal de São Paulo - UNIFESP, São Paulo (SP), Brasil.*

Autor Coorespondente: **José Márcio Duarte**
e-mail: **jmarcio_duarte@yahoo.com.br**

Artigo recebido: **30/01/2017**
Aprovado: **28/05/2018**

INTRODUCTION AND BACKGROUND

Natural language processing (NLP) is a subfield of artificial intelligence that applies computational techniques for analyzing texts at one or more levels of linguistic analysis: phonological, morphological, lexical, syntactic, semantic and pragmatic. Thus, NLP aims to generate a communication structure close to human language[1-2].

NLP consists of natural language structuring of free texts for computational analysis. This approach is widely used for information identification, coding, and extraction in natural spoken or written language[3-5]. Part-of-speech (POS) tagging refers assignment of label or information to each word in a text corresponding to its grammatical class; for example, the word "school" is tagged as a noun[5].

The PoS tagging task is based on collections of texts called *Corpus*. In an annotated *Corpus*, every word in the text is assigned to its grammatical class. NLP typically uses an annotated *Corpus* that is easily available such as a *Corpus* of newspaper articles. Having a domain-specific *Corpus* for NLP is the key to solve several problems such as ambiguity, gender, and knowledge-specific domain. It is also a time-consuming and high cost task[6-8].

In Brazil, the frequently used Newspaper *Corpus* is formed by the CETEMPúblico[9] and CETENFolha[10]. They are a collection of print newspaper articles, respectively, from Portugal and Brazil. In a second step the text phrases were morphologically and syntactically parsed, and then this *Corpus* was denominated *Floresta Sintática*[11]. *Bosque* along with *Selva*, *Amazônia*, and *Floresta Virgem Corpora* make up *Floresta Sintática*. The *Corpus Bosque* is linguistically reviewed and it contains 9.368 sentences.

Hahn and Wermter[6] have reported that general purpose Corpus such as Newspaper Corpus can be used for specific purposes. However, recent studies have shown that domain-specific Corpus are more appropriated to extract information than Newspaper Corpus[8,12-13]. Scientific articles can turn out to be important data source in the construction of a domain-specific Corpus in health, given the access restrictions to clinical data registry[13].

Accuracy assesses the performance of classification PoS tagging models as demonstrated by Han[14]. Terms that are part of the construction of metrics include: True Positive (TP) - defined as positive elements that were correctly classified; True Negative (TN) - defined as negative elements that were correctly classified; False Positive (FP) - defined as negative elements that were erroneously classified; and False Negative (FN) - positive elements that were erroneously classified. These terms are used to generate the confounding matrix that will allow to assessing how the model recognizes elements of different grammatical classes. Accuracy is measured using the formula:

$$\textbf{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

This study aimed to compare Newspaper *Corpus* with written texts in Portuguese of Brazil and Portugal and a Scientific *Corpus* with written texts in Brazilian Portuguese. Thus, we will examine if the *Corpus* specificity is more appropriate to PoS information extraction in similar texts. The Scientific *Corpus* is applied to large area of Nursing, more specifically in Cancer and Chronic Diseases. The aims are: analyze differences and similarities through the accuracy measurement; descriptive analysis - use of specific words; and independence of components in the *Corpora*.

METHODS

To achieve our proposed objective, we produced a *Corpus* of scientific articles fetched in the MEDLINE[15] and LILACS[16] databases, accessible through Virtual Health Library (VHL)[17], selecting; Brazil as Country/ Region; Portuguese as Language; Nursing as Journal Subject; and 2010 to 2014 as Year of Publication.

As Subject, for producing this *Corpus*, we performed two searches on the VHL[17]. We included in our First search the following sets of words "(OR): *câncer*, *neoplasia*" (cancer, neoplasia); and in the Second search we used the term "*doença crônica*" (chronic disease). The searches were conducted on April 23, 2015.

The Scientific *Corpus* was elaborated from the two queries. In the First one, we retrieved 77 articles, of which ten duplicates and two were not available due to access restrictions. In the Second one we retrieved 40 articles, of which seven duplicates, one article was not available as full text and one article had been selected in the First search. Thus, 96 articles were used to produce Specific Scientific *Corpus*.

The construction process of the Scientific *Corpus* consisted on removing Author Names and bibliographic references, and then converting the articles from pdf into txt format preserving articles structure - title; abstract; methods; results; discussion; and conclusions.

For our study, we used the *Bosque* as the Newspaper *Corpus*. It is formed by 9.368 with sentences from CETEMPúblico[9] and CETENFolha[10] *Corpora*. We selected the *Corpus Bosque* with txt format, appropriate for processing.

The analysis consists of three steps of support for the specific PoS model development. As the first step, we selected in both *Corpora* a random sample of 10% and performed a descriptive analysis in relation to frequency and coincidence of words. We used Quartile distribution to demonstrate words frequency difference.

We then performed morphological annotation using the PALAVRAS[18] parser of the Visual Interactive Syntax Learning (VISL)[19] in each *Corpus*. Therefore, the tokens were generated, that is the word plus its grammatical class.

For each annotated *Corpus*, with accuracy measuring purpose by means of training and evaluation, we randomly sampled selected 100.000 tokens. As a basis for the accuracy test, we parted two standards containing 10% of tokens from each *Corpus*. The remaining 90%, we generate ten subsets of cumulative sizes (5.000; 10.000; 20.000;…; and 90.000) from each *Corpus* for training. The resulting unique models were tested by evaluating their accuracy.

Training was conducted using the PoS tool of Apache

OpenNLP[20] for generating models. Apache OpenNLP is a learning-based tool kit for the processing of words and includes tools such as maximum entropy and perceptron for tokenization, named entity extraction, and PoS tagging, among others. We use the maximum entropy algorithm as the basis for the training.

We obtain the learning curve from accuracy measurement of Scientific *Corpus* and Newspaper *Corpus* trained. The curve indicates the attribution performance of the Grammatical Information to words (Step II).

OpenNLP uses the PoS Tagger Evaluator tool to measure accuracy of PoS model. The accuracy of a PoS model for a test set of n words is determined by the Formula (1):

$$\textbf{Accuracy} = \frac{TP1 + TN1 + TP2 + TN2 + ... + TPn - 1 + TNn - 1 + TPn + TNn}{P1 + N1 + P2 + N2 + ... + Pn - 1 + Nn - 1 + Pn + Nn}$$

Formula (1)

In the third step, for each *Corpus*, we use PoS information bigrams and compare the components independence of texts according to the Pointwise Mutual Information (PMI) concepts. Mutual Information is defined as amount of information obtained about one random variable through another random variable. PMI is thus a measure of two particular points of distribution, i.e., a measure of the association between elements or events of a set, and quantifies the amount of information obtained from two elements or events. The measures of associations of events are based on information theory[5, 21].

PMI compares the probability of observing x and y together with the probabilities of observing x and y independently. If there is an actual association between x and y, then the joint probability of this association will have a positive value. If x and y are independent, there is no association between them, and PMI will provide no information and will be close to zero. The measure of association will have a negative value for events that occur together less frequently despite having a high frequency independently. PMI with positive or negative values close to zero indicates a strong independence of events[5, 21]. PMI of two elements x and y is given by the Formula (2):

$$\mathbf{I(x,y)} = \log 2 \left(\frac{P(X,\ Y)}{P(X)\ *\ P(Y)} \right)$$

Formula (2)

PMI has been applied to measure independence of *Corpus* components[8]. The events are components that make a bigram as seen in Formula (3).

$$\mathbf{I(PoS1,\ PoS2)} = \log 2 \left(\frac{P(PoS1PoS2\ \ )}{P(PoS1)\ \ *\ P(PoS2)} \right)$$

Formula (3)

For this experiment, we analyzed only PoS tagging from the *Corpora*. We use completely both Newspaper *Corpus* (191.178 PoS) and Scientific *Corpus* (356.342 PoS). The calculation of PMI values in each *Corpus* was based on bigrams and their frequencies and each element of a bigram and its frequency. The Frame 1 is an example of bigrams and its components.

Frame 1 - Bigrams and its components example

Sentence:	As autoras apresentam o conceito
Annotated sentence:	as_DETFP autoras_NFP apresentam_VFIN o_DETMS conceito_NMS ._.
PoS Information:	DETFP NFP VFIN DETMS NMS
Bigrams:	{DETFP NFP}, {NFP VFIN}, {VFIN DETMS}, {DETMS NMS} {NMS.}

RESULTS

The results of this study, with focus on neoplasia and chronic diseases, are described in the experiments below, described in methods.

I – Descriptive Analysis

The Scientific *Corpus* (containing 356,342 words) had almost twice the number of words of the Newspaper *Corpus* (191,178 words). We selected random sampled of 10% in each *Corpus*, eliminating number and graphical signs (+, *, ? etc.). Thus, the sampled of Newspaper *Corpus* contains a "total" of 17.794 words and Scientific *Corpus* a "total" of 33.720 words

The words distribution in each *Corpus* is extraordinarily asymmetrical, in Table 1, we can observe the quartiles distribution. In tables 3 and 4, we can better examine this asymmetry by means "different" words distribution in each *Corpus*.

In Table 2 we can observe that the last quartile contain only four words in each *Corpus*. It is worth mentioning that given the large number of repetitions each quartile doesn't represent exactly 25%.

Distribution of number of "different words" - the Newspaper *Corpus* sample contains a "total" of 17.794 words with 5.238 "different". The Scientific *Corpus* sample contains a "total" of 33.720 words with 5.322 "different". There are 1.537 "different words" that match in the *Corpora*, taking into account the frequency of each word, which represents 12.745 words in Newspaper *Corpus* and 24.692 in Scientific *Corpus*. We found, 3.701 (representing 5.049 in total) in Newspaper *Corpus* and 3.785 (representing 9.028 in total) in Scientific *Corpus*, "different words" that don't match.

In Table 3 and Graph 1 we can observe, for each *Corpus*, the frequency distributions (absolute and relative) of "different words".

In Table 4 and Graph 2 we can observe, for each *Corpus*, the frequency distributions (absolute and relative) of "different words" that match in the two *Corpora*.

II – PoS evaluation of Newspaper and Scientific *Corpora*.

Scientific models were more accurate than Newspaper models. These models were tested in the Scientific texts samples that contained 10.000 tokens, as presented in methods. We can observe that when the number of tokens was increased, the model accuracy increased as well (Table

Table 1 - Newspaper *Corpus* and Scientific *Corpus* – Distribution in quartiles

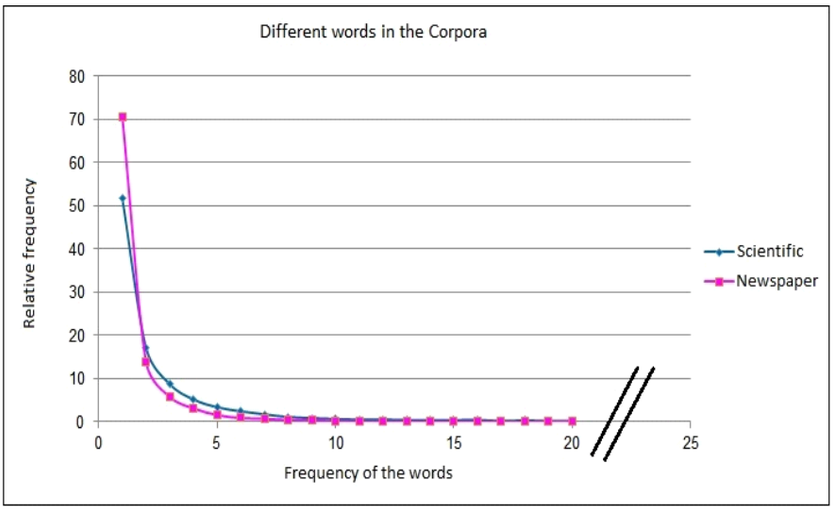
Quartiles	Newspaper *Corpus*			Scientific *Corpus*		
	Nº	%	Σ%	Nº	%	Σ%
First (Q1)	5.162	29.01	29.010	8.798	26.10	26.096
Median (Q2)	3.779	21.24	50.247	8.111	24.06	50.154
Third (Q3)	4.163	23.396	73.643	8.022	23.794	73.949

Table 2 - Newspaper and Scientific *Corpora* – Words from the last quartile

Words	Newspaper *Corpus*		Scientific *Corpus*	
	Nº	%	Nº	%
em	606	3.41	1.084	3.21
o	1.160	6.52	1.812	5.37
a	1.379	7.75	2.518	7.47
de	1.545	8.68	3.369	9.99
Total	4.690	26.36	8.783	26.04

Table 3- Newspaper and Scientific *Corpora* - Frequencies (absolute and relative) of different words

Frequency	Different words			
	Newspaper *Corpus*		Scientific *Corpus*	
	Nº	%	Nº	%
1	3.702	70.68	2.754	51.75
2	730	13.94	916	17.21
3	307	5.86	468	8.79
4	166	3.17	278	5.22
5	84	1.60	182	3.42
6	48	0.92	132	2.48
7	39	0.74	95	1.79
8	24	0.46	62	1.16
9	20	0.38	47	0.88
Subtotal	**5.120**	**97.75**	**4.934**	**92.70**
≥10	118	2.25	388	7.29
Total	**5.238**	**100.00**	**5.322**	**100.00**



Graph 1 - Newspaper and Scientific *Corpora* - Frequencies (absolute and relative) of different words

5), however the increments percentage reduces as the number of tokens increases in both models, as seen when compared Models 1, 2 and 3 to Models 8, 9 and 10.

The Graph 3 shows accuracy curves for both Scientific *Corpus* and Newspaper *Corpus*. The two upward curves become parallel to the horizontal axis as the number of tokens in each model increases.

We observe that the accuracy of Newspaper PoS models tested for Scientific texts was higher than the accuracy of Scientific PoS models tested for Newspaper texts.

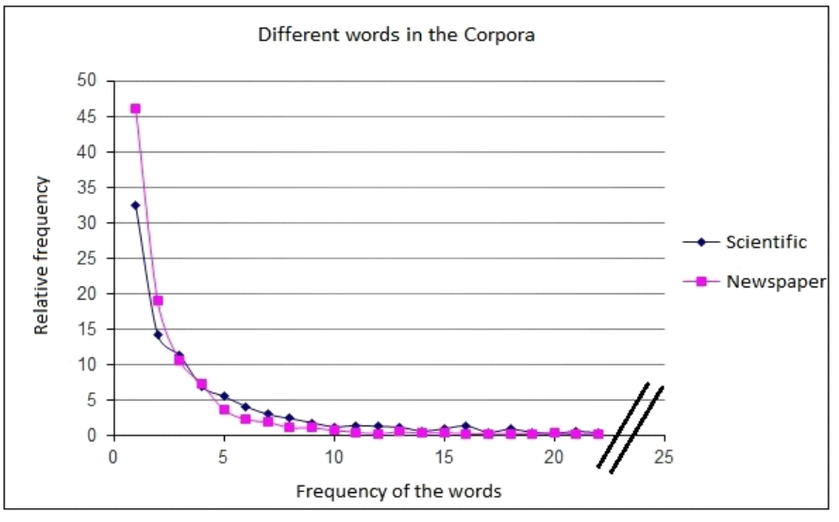
The Graph 4 shows the curves of accuracy of the Newspaper *Corpus* tested for Scientific texts and the accuracy of the Scientific *Corpus* tested for Newspaper texts. The analysis shows higher accuracy in the Newspaper *Corpus*. The two curves have a similar pattern (Graph 3), they become parallel to the horizontal axis as the number of tokens in the models increases.

III – *Pointwise Mutual Information (PMI)*

In the third step, to measure components independence

Table 4 - Newspaper and Scientific - Frequencies (absolute and relative) of different words that match in the two *Corpora*

Frequency	Different words			
	Newspaper *Corpus*		Scientific *Corpus*	
	Nº	%	Nº	%
1	709	46.16	499	32.49
2	291	18.95	218	14.19
3	162	10.55	174	11.33
4	112	7.29	106	6.90
5	56	3.65	85	5.53
6	35	2.28	63	4.10
7	29	1.89	46	2.99
8	17	1.11	37	2.41
9	17	1.11	28	1.82
Subtotal	**1.428**	**92.99**	**1.256**	**81.76**
≥10	108	7.03	280	18.23
Total	**1.536**	**100.00**	**1.536**	**100.00**



Graph 2 - Newspaper and Scientific - Frequencies (absolute and relative) of different words that match in the two *Corpora*

Table 5 - Accuracy of the Newspaper and Scientific models measured in standard text Scientific

Model	No. Tokens	Accuracy	
		Newspaper *Corpus*	Scientific *Corpus*
1	5.000	0.7701	0.8309
2	10.000	0.8115	0.8677
3	20.000	0.8448	0.8921
4	30.000	0.8536	0.8994
5	40.000	0.8663	0.9090
6	50.000	0.8731	0.9167
7	60.000	0.8773	0.9194
8	70.000	0.8789	0.9233
9	80.000	0.8814	0.9281
10	90.000	0.8832	0.9295

of texts according to the PMI concepts we select only the PoS information from both *Corpora*, as described in the methods.

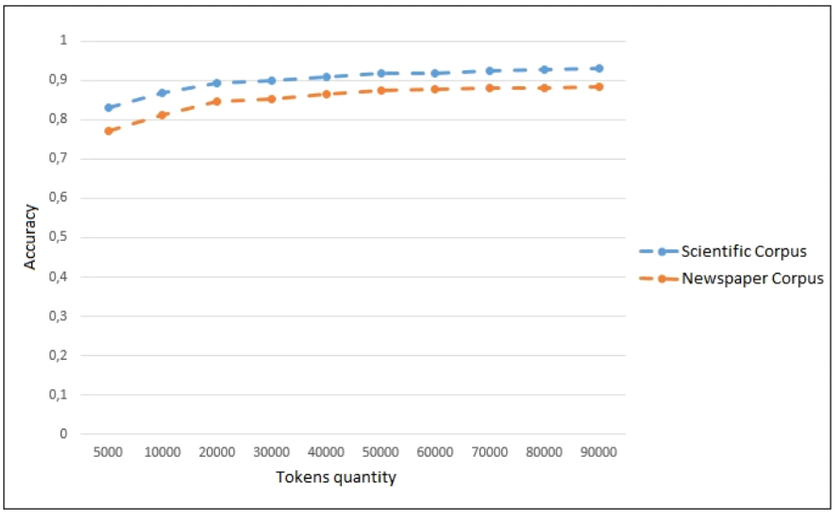
We can observe in the Table 7 that both positive and negative PMI averages for the Newspaper and Scientific *Corpora* did not show statistically significant difference.

DISCUSSION

The result denoted that specific and different words are used by means of the comparison of words in the two *Corpora*, as demonstrated by the descriptive analysis.

It is understood that the Newspaper articles having multiples genres contents contain a diversified lexicon. In addition, because the Newspaper is popular, it is written in language that is easily understood to general readers. In Scientific articles include texts of a particular genre with more specific lexicon.

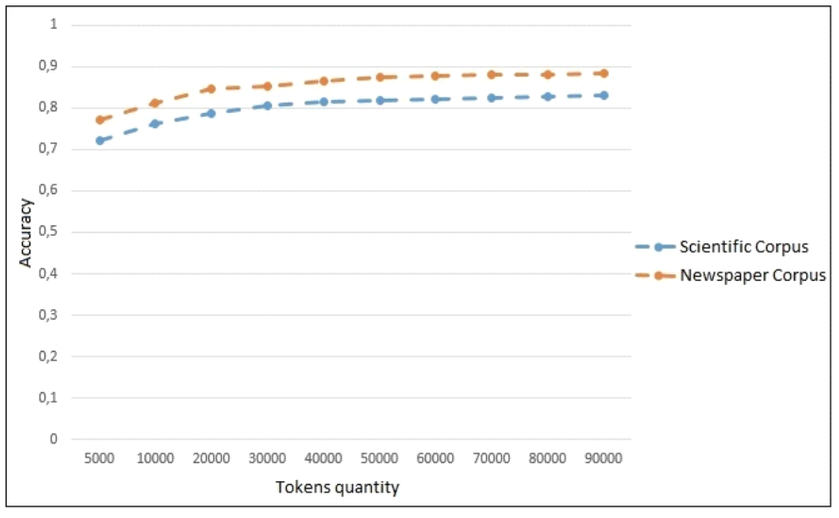
The PMI calculations for the bigrams of Newspaper *Corpus* and Scientific *Corpus* (Table 7) demonstrated similarity between used grammars. Differently from the result described by Campbell e Johnson[8], the occurrence



Graph 3 - Accuracy of the Newspaper and Scientific models measured in text Scientific

Table 6 - Accuracy of the Newspaper and Scientific models measured in text Newspaper

Model	No. Tokens	Accuracy	
		Newspaper *Corpus*	Scientific *Corpus*
1	5.000	0.7701	0.7214
2	10.000	0.8115	0.7614
3	20.000	0.8448	0.7869
4	30.000	0.8536	0.8071
5	40.000	0.8663	0.8134
6	50.000	0.8731	0.8174
7	60.000	0.8773	0.8202
8	70.000	0.8789	0.8244
9	80.000	0.8814	0.8271
10	90.000	0.8832	0.8296



Graph 4 - Accuracy of the Newspaper and Scientific models measured in relation to the standard text Newspaper

of PoS bigrams in the medical texts has been shown to be less independent compared to those in a Newspaper texts and the medical text has a less complicated grammar according to the calculations of PMI.

Although we did not find statistically significant difference between the grammars of *Corpora*, it is worth mentioning that Campbell e Johnson [8] compared Newspaper *Corpus* formed by discharge summaries written English language. In this study we used a Scientific *Corpus* with written texts in Brazilian Portuguese applied to large area of Nursing Care and a Newspaper *Corpus* with written texts in Portuguese of Brazil and Portugal.

The challenge of PMI is that bigrams composed of low-frequency words receive a higher score than bigrams composed of high-frequency words, and therefore can lead to a biased interpretation of dependence of bigrams. To overcome these effects, we analyzed PMI in which words occurred at least three times in the *Corpus*[5].

The scientific model showed an accuracy 4.63% higher than that in the newspaper model measured in scientific

Table 7 - Number, Mean and Confidence Interval to PoS bigrams

Corpus		PMI	
		Positive	**Negative**
Newspaper	**Nº**	736	916
	Mean	1.546	-2.324
	CI (95%)	[1.4566,1.6356]	[-2.4338,-2,2135]
Scientific	**Nº**	746	1029
	Mean	1.556	-2,354
	CI (95%)	[1.4620,1.6491]	[-2.4629,-2.2443]

texts. In contrast to previous studies reporting manual assignment of tags, we automatically assigned PoS tags in our study. This finding is corroborated by Campbell and Johnson[8], and Oleynik et al.[12] studies showing that models trained on newspaper *Corpus* are less effective than those trained on medical *Corpus* for the analysis of medical texts. Furthermore, Hahn and Wermter[6] argued that models trained on a Newspaper *Corpus* can be used in medical texts because the difference in accuracy is minimal. However, since specificity is key for the extraction of medical information, the lexicon of a Newspaper *Corpus* may prove ineffective.

The accuracy of Scientific *Corpus* in the newspaper texts was 5.36% lower than the accuracy of Newspaper *Corpus* in the scientific texts. This is because the Newspaper *Corpus* is composed of texts of multiple genre including health-related texts whereas the Scientific *Corpus* is composed of texts of a specific genre.

Accuracy results are important for validating Nursing Care oriented scientific articles for the elaboration of a specific *Corpus*. A difference of 4% in accuracy represents one error per sentence, which limit the number of sentence that can be parsed correctly[13].

CONCLUSION

This study compared Newspaper *Corpus* with a Scientific *Corpus* applied to large area of Nursing, more specifically in Cancer and Chronic Diseases.

The result denoted that specific and different words are used by means of the comparison of words in the two *Corpora*. Thus, the model trained on a Newspaper *Corpus* will not be effective in recognizing the grammatical class of specific words of Nursing Care oriented texts requiring the addition of other more domain-specific *Corpus*.

The result of accuracy evaluation is consistent with that of other similar studies, reaffirming the importance of the development of a specific *Corpus* to information extraction. The PMI calculations for the PoS bigrams of Newspaper *Corpus* and Scientific *Corpus* demonstrated a similarity between grammars used.

The experiments carried out lead us to conclude that in order to extract, with accuracy, PoS information demands association of the text with a specific *Corpus* and not a generic like the Newspaper *Corpus*.

ACKNOWLEDGEMENTS

The authors would like to thank the Department of Health Informatics at UNIFESP and CNPq for their support.

FUNDING

This work is part of a project, Conceptual Mapping: construction and evaluation of automatic corrector health clinical cases, process number 457715/2014-6, supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (Universal 2014 – MCTI/CNPq).

REFERÊNCIAS

1. Liddy ED. Enhanced text retrieval using Natural Language Processing. Bull Am Soc Inform Sci. 1998 Abr/May; 24(4):14-6.
2. Coppin B. Compreensão de linguagem. In: Inteligência artificial. Rio de Janeiro: LTC; 2012. p. 495-524.
3. Ferreira LS. Medical information extraction in European Portuguese [dissertação]. Aveiro: Universidade de Aveiro: Departamento de Electrónica, Telecomunicações e Informática; 2011.
4. Abulkhair M, ALHarbi N, Fahad A, Omair S, ALHosaini H, AlAffari F. Intelligent integration of discharge summary: a formative model. Proceedings of the 2013 Fourth International Conference on Intelligent Systems, Modelling and Simulation (ISMS); 2013 Jan 29-31; Bangkok, TH. US: IEEE Xplore Digital Library; 2013.
5. Manning CD, Schuitze H. Collocations. In: Foundations of statistical natural language processing. Massachusetts: The MIT Press; 1999. p. 151-225.
6. Hahn U, Wermter J. High-Performance tagging on medical texts. Proceedings of the 20th International Conference on Computational Linguistics; 2004 Aug 23-27; Geneva, SE. Pennsylvania: Association for Computational Linguistic; 2004.
7. Wermter J, Hahn U. An annotated German-language medical text *Corpus* as language resource. Proceedings of the 4th International Conference on Language Resources and Evaluation; 2004 May 26-28; Lisbon, PT. Pennsylvania: CiteSeerX; 2004.
8. Campbell DA, Johnson SB. Comparing syntactic complexity in medical and non-medical *Corpora*. Proceedings of the AMIA Symposium; 2001 Nov 3-7; Washington, DC.
9. Rocha PA, Santos D. CETEMPúblico: Um *Corpus* de grandes dimensões de linguagem jornalística portuguesa. Anais do V Encontro para o processamento computacional da língua portuguesa escrita e falada; 2000 Nov 19-22; São Paulo, SP. São Paulo: ICMC/USP; 2000.
10. Linguateca [Internet]. CETENFolha [cited 2015 Nov 05]. Available from: http://www.linguateca.pt/
11. Afonso S, Bick E, Haber R, Santos D. Floresta sintá(c)tica: um treebank para o português. Anais do XVII Encontro Nacional da Associação Portuguesa de Linguística (APL 2001); 2001 Out 02-04; Lisboa. Lisboa: APL; 2001.
12. Oleynik M, Nohama P, Cancian PS, Schulz S. Performance analysis of a POS tagger applied to discharge summaries in

portuguese. Proceedings of the 13th World Congress on Medical and Health Informatics; 2010 Sep 12-15; Cape Town, ZA.
13. Smith L, Rindflesh T, Wilbur WJ. MedPost: a part of speech tagger for bioMedical text. Bioinform J. 2004 Sept;20(14):2320-1.
14. Han J, Kamber M, Pei J. Data mining: concepts and techniques. 3a ed. Massachusetts: Elsevier; 2012.
15. MEDLINE®/PubMed® Resources Guide [Internet]. Maryland:U. S. National Library of Medicine. [cited 2015 Oct 31]. Available from: https://www.nlm.nih.gov/bsd/pmresources.html
16. LILACS [Internet]. São Paulo: BIREME – OPAS – OMS. [cited 2015 Oct 31]. Available from: http://lilacs.bvsalud.org/
17. Portal Regional da BVS. Informações e conhecimento para saúde [Internet]. São Paulo: BIREME – OPAS – OMS. [cited 2015 Oct 31]. Available from: http://pesquisa.bvsalud.org/portal/advanced/
18. Bick E. The parsing system palavras: automatic grammatical analysis of portuguese in a constraint grammar framework [dissertação]. University of Århus (DK): Department of Linguistics; 2000.
19. Visual Interactive Syntax Learning [Internet]. [cited 2015 Nov 05]. Available from: http://beta.visl.sdu.dk/
20. openNLP [Internet]. The Apache Software Foundation [cited 2015 Nov 07]. Available from: https://opennlp.apache.org/
21. Church KW, Hanks P. Word association norms, mutual information, and lexicography. Comput Linguistics. 1990 Mar;16(1):22-9.