Proceedings of Symposium in Information and Human Language Technology. Natal, RN, Brazil, November 4–7, 2015.

n-Gramas de Caractere como Técnica de Normalização Morfológica para Língua Portuguesa: Um Estudo em Categorização de Textos

Guilherme T. Guimarães, Marcus V. Meirose, Sílvia M. W. Moraes

Faculdade de Informática
Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Caixa Postal 1429– 90619-900 – Porto Alegre – RS – Brazil

guilherme.guimaraes1991@gmail.com, mvmeirose@gmail.com, silvia.moraes@pucrs.br

Abstract. *This paper describes a study on text categorization using a character n-grams approach for the morphological normalization. In recent work, this approach has emerged as a way to simplify the normalization of terms. In our research, we compared this approach to the usual normalization methods of stemming and lemmatization. In our case study, we used a subset of the PLN-BR CATEG corpus and SMO classification algorithm from the Weka tool. The results show that the character n-gram approach is promising.*

Resumo. *Este artigo descreve um estudo em categorização de textos que utiliza n-gramas de caractere como método de normalização morfológica. Em trabalhos recentes, essa abordagem tem surgido como uma forma de simplificar a normalização dos termos. Em nossa investigação, comparamos essa abordagem a métodos usuais de normalização como stemming e lematização. Em nossos casos de estudo, usamos um subconjunto do corpus em PLN-BR CATEG e o algoritmo de classificação SMO da ferramenta Weka. Os resultados obtidos mostram que a abordagem de n-grama por caractere é promissora.*

1. Introdução

Com a evolução da era digital, a quantidade de informação que se encontra ao nosso alcance cresceu de uma forma significativa. Cresceu, também, a necessidade de organizar tais informações e transformá-las em dados úteis. Atualmente, tribunais, empresas, escritórios entre outros negócios necessitam de uma forma automatizada de organização dos textos. Para isso a técnica de classificação em categorias tem sido de grande ajuda. Como a organização desses artefatos exige uma grande demanda de tempo e trabalho manual, resultando em perda de efetivo para o "negócio", a solução tem sido o uso de formas automatizadas de organização. Tais formas incluem as técnicas de categorização de texto.

A categorização ou classificação de textos consiste em atribuir objetos (documentos textuais) de um universo a duas ou mais classes (ou categorias) [Manning e Schütze, 1999; Sebastiani, 2002]. Como a base de execução dessa tarefa está centrada nos termos[1] existentes nos textos, a normalização linguística acaba tendo um papel

1 Um termo pode ser a raiz de uma palavra, uma palavra, uma sequência de palavras ou mesmo uma sentença inteira.

importante nesse processo. É por meio de técnicas de normalização linguística que podemos reduzir o conjunto de termos, unificando as variantes de um termo a uma mesma forma de representação.

Nesse trabalho, estudamos o impacto, no processo de categorização de textos em português, da substituição de técnicas usuais de normalização morfológica como *stemming* (ou radicalização) e lematização por *n*-gramas[2] de caractere. Um *n*-grama de caractere é uma sequência consecutiva de *n* caracteres. Segundo a literatura na área, há várias razões que justificam o uso de *n*-gramas de caractere no processo de categorização de textos: essa técnica é puramente estatística, não é dependente de linguagem, não requer qualquer outro conhecimento sobre o texto para ser aplicada [Rahmound e Zakaria, 2007] e, ainda é mais tolerante tanto a erros ortográficos e sintáticos quanto a ruídos existentes em textos digitalizados [Cavnar e Trenkle, 1994].

Em nossos casos de estudo em categorização de textos, usamos o *corpus* PLN_BR CATEG[3]. Para este *corpus*, criamos, inicialmente, dois casos de estudo, ditos de referência, baseados em unigramas de palavra para cada forma de normalização usual: *stemming* e lematização. Em seguida, definimos vários casos de estudo baseados em *n*-grama de caractere para diferentes valores de *n*. Em todos os casos de estudo, utilizamos a técnica de limiar por *ranking* como forma de seleção de características. Na etapa de classificação, usamos o algoritmo da ferramenta Weka[4]: Sequential Minimal Optimization (SMO), que é uma versão do algoritmo Support Vector Machine (SVM). Os resultados obtidos em nosso estudo mostram que a abordagem baseada em *n*-gramas de caractere como forma normalização morfológica para língua portuguesa é promissora, no entanto mais estudos precisam ser realizados. Cabe mencionar que as principais vantagens dessa abordagem são a sua simplicidade, tolerância a erros diversos e sua indepedência de linguagem.

Este artigo está organizado em 5 Seções. A Seção 2 descreve de forma sucinta alguns métodos de normalização linguística. A Seção 3 descreve brevemente alguns trabalhos correlatos ao nosso. A Seção 4 detalha o nosso estudo em *n*-gramas de caractere aplicado à categorização de textos. E, por fim, a Seção 5 apresenta as nossas conclusões.

2. Normalização Linguística

O objetivo da normalização linguística é transformar as variantes de um termo em uma forma única de representação. A normalização linguística pode ser morfológica, léxico-semântica ou sintática [Galvez *et al,* 2005]. A normalização morfológica é aplicada a termos com formas semelhantes cujos conceitos, em geral, estão relacionados, tais como "conectado", "conexão" e "conectando". Nesse caso, as variantes poderiam ser representadas por "conexão". A normalização léxico-semântica é usada em termos com similaridade semântica como "estado emocional", "estado afetivo" e "sentimento". Esses termos poderiam ser reduzidos ao termo "sentimento". Já a normalização sintática é usada em termos com estruturas sintáticas diferentes que possuem significados semelhantes como em "desempenhou com eficiência", "desempenho eficiente" e "eficiência em desempenho". Todas essas formas poderiam ser transformadas em "desempenho eficiente".

2 *N*-gramas podem ser definidos em nível de palavras, caracteres ou bytes [Graovac, 2012].

3 Esta coleção foi obtida através do projeto Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil (PLN-BR), apoio CNPq #550388/2005-2.

4 http://www.cs.waikato.ac.nz/ml/weka/

A normalização morfológica é efetuada por meio de métodos de conflação. Conflação consiste em fundir variantes em uma única forma. Há vários métodos automáticos de conflação, dentre os mais usuais estão o *stemming* (radicalização) e a lematização [Gonzalez e Lima, 2003; Galvez *et al.,* 2005]. No *stemming,* a normalização é baseada na remoção de afixos, transformando as variantes em seus radicais. Por exemplo, "conectado" e "conectando" seriam representados por "conect". Já na lematização, as variantes são levadas a sua forma canônica (lema): os verbos vão para a forma infinitiva e os adjetivos e substantivos, para masculino singular (se existir). A conflação por lematização transformaria os termos exemplificados em "conectar".

Métodos baseados em *n*-gramas de caractere também podem ser usados como formas de conflação [Galvez *et al.,* 2005; Sharma, 2012]. Um desses métodos é o bigrama compartilhado. Nesse método, primeiramente, os termos são divididos em duas letras consecutivas, os bigramas. Por exemplo, o termo "filho" possui os bigramas "fi", "il","lh" e "ho". O processo de conflação é definido a partir da quantidade de bigramas compartilhados pelos termos. Para identificar esse compartilhamento são usadas, geralmente, medidas de similaridade, tal como o coeficiente Dice [Galvez *et al.,* 2005; Sharma, 2012]. Alguns autores classificam esse método como um algoritmo de *stemming* baseado em *n*-grama, com abordagem estatística [Jivani, 2011; Diyanati *et al.* 2014].

Hassan e Chaurasia em [Hassan e Chaurasia, 2012] descrevem outro método de conflação baseado em *n*-gramas de caractere. Eles definem *n*-gramas iniciais, médios e finais. Os *n*-gramas iniciais são gerados com os *n* primeiros caracteres do termo; os finais, com os *n* últimos e os médios, com os *n* mais centrais. Por exemplo, a palavra "casa" possui "ca" como bigrama inicial, "as" como bigrama médio e "sa" como bigrama final . Os *n*-gramas iniciais também foram usados por Mayfield e McNamee como método de conflação em [Mayfield e McNamee, 2003], no entanto eles os chamaram de pseudos-radicais (*pseudo-stem*). Para esses autores, essa abordagem é um método de *stemming* para o qual é gerado apenas único *n*-grama inicial de caractere (*Single N-gram Stemming*).

Neste trabalho, usamos como método de conflação os *n*-gramas iniciais. Nossa escolha se baseou nos bons resultados encontrados por Hassan e Chaurasia em seus estudos em categorização de textos [Hassan e Chaurasia, 2012]. Outro motivo foi o fato de *n*-gramas iniciais serem mais simples, exigindo menos processamento computacional.

3. Trabalhos Relacionados

O primeiro trabalho que encontramos usando *n*-gramas de caractere para categorização de texto data de 1994. Nesse trabalho, Cavnar e Trenkle usam o método de bigrama compartilhado [Cavnar e Trenkle, 1994]. Inicialmente, os autores usam os termos dos textos do conjunto de treino para definir o perfil (baseado no modelo *bag-of-words*) de cada categoria de texto. Cada perfil é formado por *k n*-gramas de caractere mais frequentes. Em seguida, os autores definem um perfil para cada documento a ser classificado (do conjunto de teste) e, para definir a classe, medem a distância entre o perfil desses documentos em relação ao das categorias. Os autores usam, em seus experimentos, o *corpus* Usenet newsgroup em diferentes linguagens. No experimento cujo objetivo era classificar os artigos de acordo com a linguagem, a taxa de classificações corretas foi de 99,8%. Já naquele em que a meta era a classificação por assunto, a taxa ficou em torno de 80%. Rahmoun e Zakaria em [Rahmoun e Zakaria,

2007] utilizam uma abordagem semelhante a de Cavnar e Trenkle. Eles usam, no entanto, a medida χ^2 para associar os *n*-gramas de caractere aos perfis, e as medidas cosseno e de Kullback & Liebler para classificar os documentos. Na investigação deles, foram usados os *corpora* Reuters 21578 e 20Newsgroup, *n*-gramas de caractere com *n* variando de 2 a 7 e perfil com comprimento *k* entre 100 e 800. No caso do *corpus* Reuters 21578, a melhor média F1 foi de 70%, para *n=5* e *k=400* . Já no caso do *corpus* 20Newsgroup, F1 foi de 71% para *n=5* e *k=600.*

Em trabalhos mais recentes, Hassan e Chaurasia utilizam a categorização de textos para atribuir a autoria a documentos em língua inglesa [Hassan e Chaurasia, 2012]. Para isso, eles analisam o uso bigramas e trigramas de caractere iniciais, médios e finais na etapa de seleção de caracteristicas. Os autores realizaram vários testes e obtiveram bons resultados com bigramas e trigramas iniciais, alcançando mais de 95% de acurária. Já Kumari e outros em [Kumari *et al.*, 2014] aplicam a categorização de textos com outro fim. Eles buscam o aprimoramento da classificação de páginas *web* em relação aos seus gêneros (serviço, comércio, entretenimento,...). Em sua investigação, os autores utilizam o *corpus* 7-Genre, que contém 1.400 páginas *web* em inglês, e o algoritmo SVM como classificador. Nos estudos realizados por eles, foram testados *n*-gramas iniciais com *n* variando entre 3 e 8. O melhor resultado foi obtido com *n=5* para o qual a média F1 atingiu 95,8%.

Diferente dos trabalhos pesquisados, nosso estudo é voltado para língua portuguesa. É importante mencionar que não é de nosso conhecimento a existência de trabalhos que investiguem *n*-gramas de caractere como método de normalização linguística para o português. A seguir, descrevemos a nossa investigação usando essa abordagem em categorização de textos.

4. Estudo em Categorização de Textos

Em nosso estudo, usamos um subconjunto do *corpus* em língua portuguesa PLN-BR CATEG. Este *corpus* possui em sua totalidade cerca de 30 mil textos do jornal Folha de São Paulo dos anos de 1994 a 2005. Usamos as seções do jornal como categorias dos textos. Selecionamos as categorias desse *corpus* considerando dois aspectos: quantidade de textos e uniformidade de conteúdo. Na tarefa de categorização de textos, o balanceamento e a qualidade das amostras do conjunto de treino interfere diretamente nos resultados [Batista *et al*., 2004]. Sendo assim, categorias com poucos textos como "Agrofolha" (166 textos apenas) ou com uma diversidade grande de contéudo como a categoria "Tudo" foram desconsideradas.

Como nosso foco era especificamente a investigação dos *n*-gramas de caractere, procuramos minimizar, na medida do possível, fatores que pudessem prejudicar a tarefa de classificação, tal como o desbalanceamento das amostras do conjunto de treino [Japkowicz e Stephen, 2002]. Por questões de desempenho, optamos por utilizar a técnica de balanceamento *under-sampling* (sub-amostragem), a qual permite a eliminação de amostras de classes majoritárias [Batista *et al*., 2004]. Sabemos que essa técnica pode levar a perda de informação, se o subconjunto de amostras do treino não for escolhido adequadamente, em decorrência da ausência de uma heurística que guie esse processo de seleção. No entanto, acreditamos que essa perda eventual não prejudica os resultados obtidos, visto que nossa investigação é de natureza comparativa. Se a perda ocorrer, ela se dará igualmente em todos os casos estudados.

Em nossa investigação, foram usadas apenas 6 categorias do *corpus* PLN-BR

CATEG: “Brasil” (5.606 textos), “Cotidiano” (6.458 textos), “Dinheiro” (4.153 textos), “Esporte” (4.632 textos), “Ilustrada” (2.935 textos) e “Mundo” (2.410 textos). Juntas elas totalizaram 26.194 textos. Após, alguns testes preliminiares, acabamos escolhendo 1.000 textos de cada categoria para formar o conjunto de treino. Os textos restantes foram usados como conjunto de teste.

Além disso, utilizamos, em todos os casos de estudo, o mesmo pré-processamento. Os textos foram tokenizados e removidos os *tokens* correspondentes a *stopwords*[5], pontuação, numeração e caracteres especiais. Aplicamos também o mesmo processo de seleção de características, que foi limiar por *ranking,* a exemplo de Hassan e Chaurasia em [Hassan e Chaurasia, 2007]. Em nosso estudo, testamos diferentes comprimentos *k* (quantidade de termos) para *bag-of-words*. Cabe mencionar que a *bag-of-words* final é resultante da união dos *k* termos mais relevantes (mais frequentes) de cada categoria.

A partir da *bag-of-words* determinada na etapa de seleção, os textos receberam uma representação vetorial cujos pesos foram definidos usando a medida TFIDF. Escolhemos essa técnica por ela ser muito usual na tarefa de categorização de textos. Por fim, usamos o mesmo algoritmo de classificação em todo o estudo: SMO da ferramenta Weka. Escolhemos esse algoritmo por sua aplicação ser recorrente em trabalhos correlatos ao nosso. Por fim, analisamos os resultados com base nas medidas comumente utilizadas para avaliar a tarefa em questão: *Precision*, *Recall* e F1.

Na seções seguintes descrevemos configuração dos nossos casos de estudo e os resultados obtidos.

4.1. Configuração dos Casos de Estudo

Para que pudéssemos analisar o impacto do uso de *n*-gramas de caractere como método de normalização morfológica no processo de categorização de textos, definidos 3 tipos de casos de estudo. Nesses casos, a principal diferença foi o método de normalização aplicado aos termos. Esses casos de estudos foram nomeados e organizados da seguinte forma:

- Caso de referência usando *Stemming*: utiliza unigrama em nível de palavra e usa como método de normalização o *Stemming*. Usamos o *stemmer*[6] de Caldas Junior e outros [Caldas Junior *et al.*, 2001] cuja implementação é baseada no algoritmo de Porter [Porter, 1980].
- Caso de referência usando Lematização: usa unigrama em nível de palavra também, mas utiliza como forma de normalização a lematização. Os textos que utilizamos do *corpus* PLN-BR CATEG foram lematizados pela ferramenta FORMA, desenvolvida por Marco Gonzalez e discutida em [Gonzalez *et al*., 2006].
- Caso de estudo usando *n*-gramas de caractere: aplica *n*-gramas iniciais em nível de caractere como método de normalização morfológica. Para este caso, foram testados os seguintes valores de *n:* {3,4,5,6,7}. Para definir esse intervalo, inicialmente, analisamos o comprimento médio das palavras existentes nos textos de nosso estudo. Descobrimos que, em média, as palavras possuíam

5 Usamos a stoplist definida por Stanley Loh, que está disponível em http://miningtext.blogspot.com.br/2008/11/listas-de-stopwords-stoplist-portugues.html

6 http://www.nilc.icmc.usp.br/nilc/tools/stemmer.html

comprimento igual a 5. A partir dessa informação, decidimos investigar *n*-gramas iniciais de caractere cuja diferença de comprimento variava de -2 a +2 em relação à média. Consideramos que comprimento *n=2* seria muito pequeno para um pseudo-radical, assim como 8 seria muito longo. Cabe ressaltar que são definidos *n*-gramas de caractere apenas palavras nos quais o comprimento é maior que o valor de *n*. No caso do comprimento ser menor ou igual, a palavra é mantida e considerada na sua forma original (inteira).

Os resultados obtidos a partir desses casos de estudo são comentados nas seções a seguir.

4.2. Resultados do Caso de Referência usando *Stemming*

Neste caso de referência, realizamos o pré-processamento descrito anteriormente e aplicamos a normalização por *stemming*. Avaliamos diferentes comprimentos *k* para *bag-of-words,* onde *k*= {50,100, 150, 200, 250}. Os melhores resultados foram encontrados quando definimos *bag-of-words* de *150* termos para cada categoria. A Tabela 1 exibe os valores das medidas *Precision*, *Recall* e F1, por categoria, para a melhor configuração encontrada para este caso.

Tabela 1 – Melhor resultado para o caso de de referência usando *stemming*, com *k=150*

Categoria	*Precision*	*Recall*	*F1*
Mundo	0,51	0,82	0,63
Brasil	0,65	0,62	0,64
Cotidiano	0,74	0,66	0,71
Dinheiro	0,73	0,73	0,73
Esporte	**0,96**	**0,91**	**0,94**
Ilustrada	0,76	0,89	0,82
Média	0,73	0,77	0,75

A categoria Esporte foi a que obteve melhor classificação, provavelmente por utilizar um vocabulario mais constante. Os textos usados comentam 11 anos de esporte. Mesmo para um intervalo tão grande de tempo como esse, o termos usados nessa área pouco se alteraram. Diferente da categoria Mundo, para qual o classificador gerou o pior resultado em precisão. No espaço de 11 anos, muito do que se escreve sobre o mundo e é notícia mudou, o que deve ter provocado uma variação maior nos termos.

4.3. Resultados do Caso de Referência usando Lematização

Neste caso de referência, realizamos o mesmo pré-processamento, mas aplicamos a normalização por lematização. Também usamos o mesmo classificador e testamos igualmente diferentes valores para *k ={50,100, 150, 200, 250}*. O melhor resultado da também foi para *k=150*. A Tabela 2 exibe os valores das medidas *Precision*, *Recall* e F1, por categoria, para a melhor configuração encontrada para este caso.

Tabela 2 – Melhor resultado para o caso de referência usando lematização, com *k=150*

Categoria	*Precision*	*Recall*	*F1*
Mundo	0,39	**0,85**	0,53
Brasil	0,68	0,59	0,63
Cotidiano	0,73	0,64	0,68
Dinheiro	0,71	0,68	0,70
Esporte	**0,96**	**0,80**	**0,87**
Ilustrada	0,73	**0,83**	0,77
Média	0,73	0,70	0,71

Nesse estudo, os resultados gerais de classificação caíram um pouco, mas o *Recall* teve uma pequena melhora na maioria das categorias.

4.4. Resultado do Caso de Estudo usando n-Gramas Iniciais de Caractere

Neste caso também aplicamos o mesmo pré-processamento, mas usamos conflação por *n*-gramas iniciais de caractere. Repetimos o estudo usando o comprimento *k=150,* que foi o que gerou melhores resultados nos casos de referência apresentados. Para esta configuração, foram testados os valores de *n={3,4,5,6,7}*. A Tabela 3 exibe os valores das medidas *Precision*, *Recall* e F1, por categoria, para a melhor configuração encontrada, que foi para *n=5*.

Tabela 3 – Melhor resultado para o caso de *n*-gramas iniciais de caractere, com *k=150* e *n=5*

Categoria	*Precision*	*Recall*	*F1*
Mundo	0,49	0,82	0,62
Brasil	0,67	0,62	0,64
Cotidiano	0,78	0,68	0,72
Dinheiro	0,72	0,72	0,72
Esporte	**0,96**	**0,91**	**0,94**
Ilustrada	0,78	0,88	0,82
Média	0,76	0,74	0,75

Os resultados obtidos com esta abordagem se aproximaram muito a do caso de referência usando *stemming*. Esse resultado é interessante, pois indica que uma abordagem mais simples e, portanto, com menor exigência computacional pode ser uma alternativa quando o tempo de resposta é tão importante quanto bons resultados de classificação.

Na seção seguinte, comparamos os casos de estudo apresentados.

4.5. Análise Geral dos Resultados

No gráfico apresentado pela Figura 1, comparamos as medidas *Precision*, *Recall* e F1 dos melhores casos de referência para *stemming* e lematização com os casos baseados em *n*-gramas iniciais de caractere (NGC).

Analisando o gráfico percebemos que a partir de *n=5* para *n*-gramas de caractere

(NGC_n=5), os resultados de categorização não melhoraram. Isso aconteceu provavelmente porque o comprimento médio das palavras do *corpus* era 5, ou seja, não deveriam existir muitas palavras com comprimento maior ou igual a 6. Em relação aos casos de referência, no estudo que fizemos, os *n*-gramas iniciais de caractere foram competitivos e resultaram em valores próximos ou ligeiramente melhores que as normalizações tradicionais. No entanto, precisamos realizar um estudo mais abrangente incluindo outros *corpora* e outros algoritmos de *stemming* e lematização para considerarmos os resultados mais conclusivos. De qualquer forma, acreditamos que a abordagem é uma alternativa atrativa, pois é simples, demanda pouco processamento e não requer praticamente tratamento linguístico.

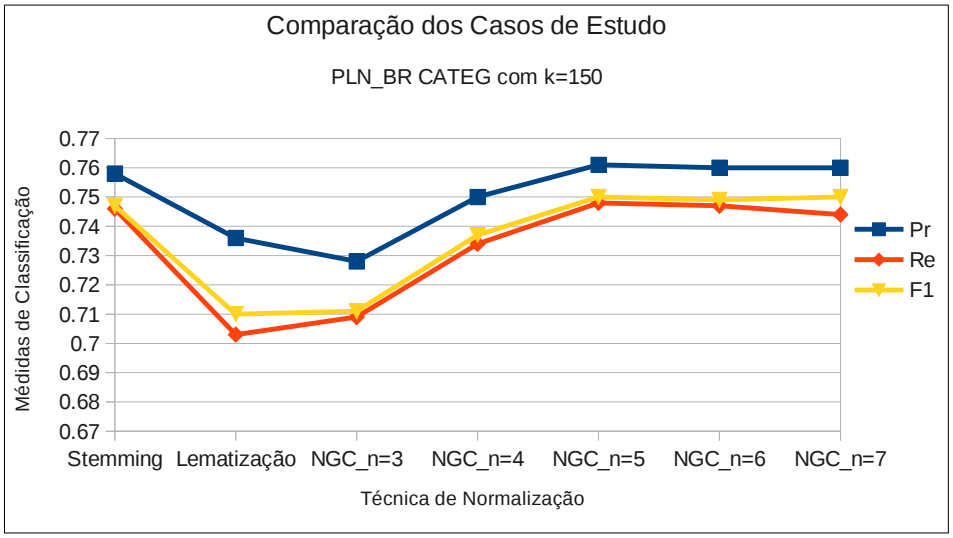

Figura 1 – Comparação dos casos de estudo

5. Conclusão

A classificação de textos é uma área que tem muita aplicabilidade, especialmente na *web,* onde há muitos documentos disponíveis em formato digital. Dado ao grande volume de textos nesse âmbito, é importante aprimorar tanto a precisão quanto a escalabilidade de sistemas classificadores. Para isso é imprescindível que exista um pré-processamento mais efetivo dos textos, com baixo custo computacional, para tratar e selecionar somente os termos mais relevantes, a fim de reduzir a alta dimensionalidade comum nessa tarefa.

Acreditamos que a abordagem de *n*-gramas iniciais pode ser usada para língua portuguesa como método de normalização linguística, pois, além de ser simples, seu custo computacional é baixo. As vantagens oferecidas pela abordagem a tornam competitiva em ambientes onde o custo computacional é muito relevante. Embora mais estudos precisem ser realizados, acreditamos, com base no nosso estudo, que o tamanho médio das palavras do *corpus* possa ser um bom valor inicial para definir o comprimento (*n*) do *n*-gramas de caractere. Faz parte de nossos trabalhos futuros, expandir nosso estudo testando outros *corpora* e outros algoritmos de normalização morfológica para a língua portuguesa.

Referências

Batista, G., Prati, R.C. e Monard, M.C. (2004). "A study of the behavior of several methods for balancing machine learning training data". SIGKDD Explor. Newsl.6, 1 (June 2004), 20-29.

Caldas Junior, J.; Imamura, C.Y, M. e Rezende, S.O. (2001)."Avaliação de um Algoritmo de Stemming para Língua Portuguesa. In the Proceedings of the 2nd Congress of Logic Applied to Technology, Vol. 2, 267-274.

Cavnar, W. B e Trenkle, J. M. (1994). "N-Gram-Based Text Categorization". In *Ann* Arbor MI, Vol. 48113, No. 2, 161-175 .

Diyanati, M.H, Sadreddini, M. H., Rasekh, A. H, Fakhrahmad, S. M. e Taghi-Zadeh, H. (2014). "Words Stemming Based on Structural and Semantic Similarity". In Computer Engineering and Applications, Vol. 3, No. 2, 89-99.

Galvez, C., Moya-Anegón, F. e Solana, V. H. (2005). "Term conflation methods in information retrieval: non-linguistic and linguistic approaches". In Journal of Documentation , Vol. 61, No. 4, 520-547.

Gonzalez, M. e Lima, V. L. S. (2003). "Recuperação de Informação e Processamento da Linguagem Natural." In XXIII Congresso da Sociedade Brasileira de Computação, Campinas, Anais do III Jornada de Mini-Cursos de Inteligência Artificial, Volume III, 347-395.

Gonzalez, M., Lima, V. L. S. e Lima, J. V. (2006) "Tools for Nominalization: an Alternative for Lexical Normalization", In the Proceedings of the 7th Workshop on Computational Processing of the Portuguese Language – Written and Spoken, PROPOR 2006, Springer-Verlag, p.100-109.

Graovac, J. (2012). "Serbian text categorization using byte level n-grams". In Proceedings CLoBL, 93–97.

Hassan, F. I. H e Chaurasia, M. A. (2012). "N-Gram Based Text Author Verification". In International Conference on Innovation and Information Management (ICIIM 2012), Vol. 36, 67-71.

Japkowicz, N. e Stephen, S. (2002). "The class imbalance problem: A systematic study", Intell. Data Anal.6, 5 , 429-449.

Jivani, A. G. (2011). "A Comparative Study of Stemming Algorithms". In International Journal Comp. Tech. Appl., Vol 2 ., No. 6, 1930-1938

Kumari, K.P., Reddy, A.V. e Fatima, S . (2014). "Web Page Genre Classification: Impact of n-Gram Lengths". In International Journal of Computer Applications, Vol. 88, No.13, 13-17.

Manning, C.D. e Schütze, H. (1999). "Foundations of Statistical Natural Language Processing". MIT Press.

Mayfield, J. e McNamee,P. (2003) "Single N-gram Stemming," In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, 415-416.

Porter, M. (1980). "An algorithm for suffix stripping". *Program*, **14**(3), 130-137. http://www.tartarus.org/~martin/PorterStemmer/def.txt.

Rahmoun, A. e Zakaria, E. (2007). "Experimenting N-grams in text categorization", In International Arab Journal of Information Technology, Vol. 4, No. 4, 377-385.

Sharma, D. (2012). "Stemming Algorithms: A Comparative Study and their Analysis". In International Journal of Applied Information Systems (IJAIS), Vol. 4, No. 3, 7-12.

Sebastiani, F. (2002). "Machine Learning in Automated Text Categorization", In A*CM Computing Surveys*, Vol. 34, No. 1, 1-47, ACM.

Part II

IV Jornada de Descrição do Português

Preface

Evento satélite do X Brazilian Symposium in Information and Human Language Technology (STIL 2015), em Natal.

A Jornada de Descrição do Português (JDP), mais uma vez, visa aproximar as comunidades de linguistas e de pesquisadores da área da Computação. A intenção é integrar, ainda mais efetivamente, essas duas áreas que, especialmente no âmbito brasileiro, precisam reforçar a atuação de forma interdisciplinar para promover avanços no processamento automático da língua portuguesa. A Linguística Descritiva, em especial, tem enorme potencial para aportar conhecimentos ao Processamento Automático de Língua Natural (PLN), de maneira a colocar a língua portuguesa numa posição de destaque no cenário mundial, fazendo frente à grande produção de recursos computacionais para outras línguas (como o inglês, francês ou espanhol), que vislumbraram essa interdisciplinaridade já na década de 1960.

Os trabalhos aqui apresentados vinculam-se aos grandes temas da descrição linguística do português, a saber: Estudos de Fonética e Fonologia, Estudos do Léxico (Lexicologia, Lexicografia e Terminologia), Estudos de Sintaxe, Estudos de Semântica, Estudos de Texto e Discurso, nas mais diversas correntes teóricas. Os trabalhos selecionados são apresentados em formato de comunicação oral ou de pôster, segundo a orientação do nosso Comitê Científico. Esperamos que os trabalhos aqui reunidos inspirem novas participações no nosso evento.

Nesta edição da JdP, temos os seguintes trabalhos, apresentados por colegas de diversas regiões do Brasil e de Portugal:

Coordenação

Lucelene Lopes (PUCRS, Porto Alegre, RS, Brasil)
Maria José Bocorny Finatto (UFRGS, Porto Alegre, RS, Brasil)

Organização

Maria José Bocorny Finatto (UFRGS, Porto Alegre, RS, Brasil)
Lucelene Lopes (PUCRS, Porto Alegre, RS, Brasil)
Andrea Jessica Borges Monzon (UFRGS, Porto Alegre, RS, Brasil)
Alena Ciulla (UFRGS, Porto Alegre, RS, Brasil)
Aline Evers (UFRGS, Porto Alegre, RS, Brasil)
Bianca Pasqualini (UFRGS, Porto Alegre, RS, Brasil)

Comitê Científico

Alena Ciulla (Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil)
Aline Evers (Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil)

Aline Villavicencio (Universidade Federal do Rio Grande do Sul, Porto Alegre RS, Brasil)
Ariani Di Felippo (Universidade Federal de São Carlos, São Carlos, SP, Brasil)
Éric Laporte (Université Paris Est, Marne-La-Vallée, França)
Gladis Maria de Barcellos Almeida (Universidade Federal de São Carlos, São Carlos, SP, Brasil)
Guilherme Fromm (Universidade Federal de Uberlândia, Uberlândia-MG, Brasil)
Lucelene Lopes (Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, RS, Brasil)
Maria José Bocorny Finatto (Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil)
Oto Araújo Vale (Universidade Federal de São Carlos, São Carlos, SP, Brasil)
Renata Vieira (Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, RS, Brasil)
Stella Esther Ortweiler Tagnin (Universidade de São Paulo, São Paulo-SP, Brasil)

Chapter 4

Apresentação Oral