Marchetti da SILVA, Edson; Rocha SOUZA, Renato

Fundamentos em processamento de linguagem natural: uma proposta para extração de bigramas

Encontros Bibli, vol. 19, núm. 40, mayo-agosto, 2014, pp. 1-31
Universidade Federal de Santa Catarina
Florianopolis, Brasil

Encontros Bibli,
ISSN (Versão eletrônica): 1518-2924
bibli@ced.ufsc.br
Universidade Federal de Santa Catarina
Brasil

ARTIGO

Recebido em:
05/05/2013

Aceito em:
13/05/2014

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, v. 19, n.40, p. 1-32, mai./ago., 2014. ISSN 1518-2924. DOI: 10.5007/1518-2924.2014v19n40p1

Fundamentos em processamento de linguagem natural: uma proposta para extração de bigramas

Fundamentals in natural language processing: a proposal for extraction bigrams

Edson Marchetti da SILVA[1]
Renato Rocha SOUZA[2]

RESUMO
É senso comum que o texto escrito é uma importante forma de registrar as informações e que atualmente grande parte desse conteúdo informacional está disponível em meio digital. Entretanto, de maneira geral, os computadores lidam com o texto como sendo uma cadeia de caracteres que não têm nenhum significado. A área de Processamento de Linguagem Natural (PLN) vem se empenhando em extrair significados do texto. Nesse sentido este trabalho apresenta uma revisão desse tema e propõe um método automatizado que utiliza uma heurística determinística denominada Heudet que visa extrair bigramas do texto. A meta é extrair o significado do texto através de um conjunto de expressões multipalavras identificadas. Os resultados obtidos foram melhores se comparados com aqueles que utilizam-se das técnicas de medidas de associação estatística obtidas pelo software Ngram Statistics Package (NSP).

PALAVRAS-CHAVE: Extração de expressões multipalavras. Medidas de associação estatísticas. Heudet.

ABSTRACT
It is common sense that the written text is an important way of to register information and currently much of this information content is available in digital form. However, in general, the computers consider a text is a string that have not significance. The area of Natural Language Processing (PLN) has been engaged in extracting meaning from text. Accordingly this paper presents a review of this issue and proposes an automated method that uses a deterministic heuristic called Heudet which aims extract bigram of the text. The goal is to extract the meaning of the text identifing a set of multiword expressions (MWE). The results were better compared to those using up the techniques of statistical association measures obtained from the software ngram Statistics Package (NSP).

KEYWORDS: Multiword expression extraction. Measures of association statistics; Heudet.

v. 19, n. 40, 2014.
p. 1-32
ISSN 1518-2924

[1] Cefet-MG - edson@div.cefetmg.br
[2] FGV/RJ - renato.souza@fgv.br

1 INTRODUÇÃO

Nesse sentido, Sarmento (2006) afirma que o texto não é um simples amontoado aleatório de palavras. A ordem da colocação das palavras no texto é que produz o significado. Portanto, o estudo da co-ocorrência das palavras traz consigo uma informação importante. Isso pode indicar que as palavras estão relacionadas, diretamente por composicionalidade ou afinidade, ou indiretamente por semelhança. Portanto, a base da linguística empírica consiste em encontrar, a partir da frequência de co-ocorrências observadas, as dependências significativas entre os termos. Esses termos adjacentes são denominados *n*-gramas ou Expressões Multipalavras (EM).

Evert (2005, citado por Sarmento) aponta como sendo quatro esses grupos de medidas utilizadas na identificação dos *n*-gramas:

- testes de significância estatística;
- coeficientes de associação;
- baseadas em conceitos da teoria da informação;
- baseadas em heurísticas diversas.

Zhang et al. (2009) corroboram com Sarmento ao afirmar que a capacidade de expressar sentido de uma palavra depende das demais palavras que a acompanham. Quando uma palavra aparece acompanhada por um conjunto de termos, maiores são as chances desse conjunto possuir um significado relevante. Isso significa que não apenas a palavra, mas também a informação contextual é útil para o processamento de informações. Tem ocorrido um crescente interesse, sobretudo na área de Processamento da Linguagem Natural (PLN), afinal essas formas fixas são tão numerosas em qualquer tipo de texto, que não podem ser ignoradas. Portanto, essas características das EM as tornam relevantes no tratamento dos recursos lexicais, os quais são importantes insumos informacionais para muitas aplicações relacionadas ao PLN, tais como: Recuperação da Informação (RI), tradução automática, sumarização de texto, etc.

É a partir dessa ideia simples e direta que pesquisas sobre EM são motivadas. Esperam-se capturar conceitos semânticos relevantes do texto expressos pelas EM. Nesse sentido, Villavicencio et al. (2010) destacam que muitas pesquisas têm buscado formas de automatização na aquisição lexical. Esses trabalhos buscam entender a formação dos recursos lexicais, uma área ainda carente de pesquisas.

Para melhor explicar o método proposto, este artigo está estruturado nas seguintes partes: na segunda seção apresenta-se uma visão sobre os trabalhos correlatos ao tema; na terceira seção, apresentam-se conceitos sobre os fundamentos linguísticos e as expressões multipalavras; na quarta seção, apresenta-se o funcionamento do método Heudet proposto; na quinta seção os resultados obtidos; e finalmente na sexta seção, apresentam-se as conclusões e as sugestões de trabalhos futuros.

2 TRABALHOS CORRELACIONADOS

No estudo apresentado por Ladeira (2010), o qual analisa os últimos 40 anos da produção científica nacional da área de PLN, realiza-se a avaliação de 621 publicações de forma horizontal e de 68 publicações através da análise de conteúdo traçando um abrangente panorama evolutivo dessa linha de pesquisa. Nesse cenário, foram constatadas as seguintes evidências: (1) a mudança do enfoque das aplicações: inicialmente, era dada maior ênfase às ferramentas linguísticas de processamento sintático e semântico, e recentemente uma nítida exploração das aplicações práticas; (2) ocorreu um expressivo crescimento de publicações após o ano 2000; (3) as áreas de Ciência da Computação (CC) e linguística atingem mais de 80% das publicações, sendo a participação da Ciência da Informação (CI) pouco representativa nesse contexto; (4) apenas doze pesquisadores foram os responsáveis por mais de 20% de toda a pesquisa nacional, sendo que desses, nenhum se declara como pertencente à área da CI; (5) observa-se que a RI foi a problemática que teve o maior destaque e com uma forte concentração dos trabalhos com publicação recente. Além disso, a maioria dos trabalhos analisados sobre RI estão voltados para técnicas de pré-processamento de documentos, o que, segundo a autora ainda seja um tema em aberto.

Ela também destaca que os resultados obtidos em alguns trabalhos sobre RI têm sido muito ruins, não apresentando melhorias significativas aos trabalhos anteriores.

As técnicas automatizadas de identificação de EM, que servem de base para encontrar os descritores nos documentos de referência, são temas de diversos trabalhos. Entretanto, não foram encontrados trabalhos que utilizam as EM extraídas, como descritores no processo de busca para identificação de documentos similares. Dentre os trabalhos que visam identificar as EM publicados nas últimas décadas destacam-se, a seguir, alguns dos mais relevantes.

Mais recentemente, conforme apontado por Wang e Liu (2010), muitos trabalhos têm como foco dominante a identificação e extração de EM. Devido à complexidade desses processos, diferentes abordagens têm sido empregadas. De maneira geral, essas abordagens de extração envolvem: (1) métodos estatísticos; (2) informações linguísticas; (3) métodos híbridos, os quais combinam essas abordagens.

Dentre esses trabalhos que visam identificar as EM destacam-se, dentre outros: Dias, Lopes e Guilloré (1999); Evert e Krenn (2005); Chen, Yeh, Chau (2006); Pedersen et al. (2011) que trabalham num contexto independente de linguagem e baseados em métodos estatísticos; Silva e Lopes (1999) que visam extrair *n*-gramas a partir da análise do texto em um contexto local denominado LocalMaxs; Cazolari et al. (2002); Pecina (2010); Portela, Mamede e Baptista (2011) levam em consideração as características morfo-sintáticas do texto, por isso demandam intensivo uso de recursos computacionais; Ramisch (2009) e Villavicencio et al. (2010) que utilizam método híbrido para identificação de EM para o processo de tradução automática.

Outros trabalhos que merecem destaque são os apresentados por Pearce (2002) e por Ramisch, Araújo e Villavicêncio (2012) os quais apresentam uma avaliação comparativa das principais técnicas e abordagens que vêm sendo adotadas por diversos pesquisadores sobre o tema de extração de EM em diversos *corpora* e idiomas.

Cada uma dessas abordagens citadas busca interpretar os conteúdos textuais escritos em linguagem natural, mas seguem caminhos diversos obtendo resultados de custo computacional[3] e de conteúdos distintos. Dessa forma, as vantagens e desvantagens de cada uma delas dependem do contexto para o qual estão sendo utilizadas.

A abordagem estatística para extração de EM através da co-ocorrência de palavras em textos utiliza várias técnicas que buscam identificar as EM como sendo um conjunto de palavras adjacentes que co-ocorrem com uma frequência acima da esperada para uma sequência aleatória de palavras em um *corpus.* Dessa forma, a abordagem associativa nada mais é que a utilização de um conjunto de medidas de associação estatística que visam identificar as expressões candidatas a EM. Dentre as técnicas empregadas destacam-se: coeficiente de Chi Quadrado de Pearson; coeficiente de Dice; Informação Mútua Pontual (PMI, do inglês Pointwise Mutual Information); medida de Poisson Stirling dentre outras.

No trabalho de Dias, Lopes e Guilloré (1999), os autores questionam que muitos estudos se restringem a dar apenas um tratamento lexicográfico ao processo de extração de informação de textos. Portanto, sugerem o uso das EM como forma de obter melhor teor informacional do texto que possa ser utilizado pelas aplicações de RI e tradução automática. Desse modo, eles propõem em seu estudo a implementação de um sistema baseado exclusivamente em técnicas estatísticas para extrair EM, que ocorrem no texto de forma contígua e não-contígua[4]. Eles utilizaram um *corpus* paralelo com o debate político do parlamento europeu com cerca de trezentas mil palavras em cada um dos quatro diferentes idiomas francês, inglês, italiano e português. O sistema proposto reúne os conceitos de Expectativa Mútua (do inglês Mutual Expectation) proposto por Dias (1999 citado por DIAS, LOPES, GUILLORÉ) e o processo de aquisição de EM baseado no algoritmo LocalMax proposto por Da Silva (1999 citado por DIAS, LOPES, GUILLORÉ).

[3] Custo computacional, neste contexto, relaciona-se ao consumo de recursos computacionais de processamento demandados numa relação direta com o tempo de resposta.

[4] Segmentos de texto não contíguos são aqueles que as EM aparecem com quebra da sequência das palavras pela presença um ou mais palavras intercambiáveis dentro do segmento.

Esse sistema está estruturado nas seguintes etapas: A primeira transforma o conteúdo textual do *corpus* em tabelas de contingência contabilizando os *n*-gramas contíguos e não-contíguos. A segunda mede a coesão de todos os *n*-gramas através do cálculo da Expectativa Mútua para todos eles. A terceira elege as EM comparando todo o conjunto de *n*-gramas pelo valor de coesão utilizando o algoritmo LocalMax. Finalmente a qualidade das EM extraídas é testada e comparada com quatro outras medidas de associação calculadas para cada um dos idiomas existentes no *corpus.* Como resultado, os autores apontam que a técnica de Expectativa Mútua apresentando maior precisão na extração, além de superar o problema da palavra muito frequente que ocorre nas demais técnicas de medida de associação empregadas.

Dias, Lopes e Guilloré (1999) destacam que a maioria dos trabalhos de PLN têm se concentrado no reconhecimento e extração de informações explícitas no texto negligenciando os contextos implícitos compostos por unidades léxicas que devem ser consideradas como indivisíveis por terem um significado ou função que não é necessariamente o mesmo que analisar cada uma das palavras separadamente. Dessa forma, esses autores trabalham utilizando métodos estatísticos de expectativa mútua, conjugando o processo de aquisição lexical com o algoritmo de máximo local para identificação dos léxicos compostos.

A abordagem utilizada por Evert e Krenn (2005) é baseada no cálculo estatístico das medidas de associação das palavras contidas no texto. Nos testes empíricos esses autores utilizaram um subconjunto de oito milhões de palavras extraídas de um *corpus* constituído de um jornal escrito no idioma alemão. A abordagem proposta foi dividida em três passos. No primeiro, extraem-se as tuplas léxicas do *corpus* fonte contendo pronomes (*P*), substantivos (*S*) e verbos (*V*). Em seguida, esses dados são agrupados em pares (*P+S, V*) e colocados em uma tabela de contingência, representada por uma estrutura tridimensional, em que cada par está colocado num plano *P+S* por *V*. Finalmente, atribui-se, no terceiro eixo do plano, a informação da frequência representada por quatro células.

Dessa forma, realiza-se uma comparação entre todos os pares léxicos extraídos do *corpus* com as suas sentenças, contabilizando a cada sentença, uma das quatro possibilidade: existe *PS* e existe *V*; existe *PS* e não existe *V*; não existe *PS* e existe *V*; não existe *PS* e não existe *V*. Ou seja, é acrescida uma unidade para cada vez que uma das possibilidades ocorrerem.

No segundo passo, as medidas de associação são aplicadas às frequências coletadas no passo anterior. Desse processamento resulta uma lista de pares de EM candidatas com seus respectivos *scores* de associação, calculados e ordenados do mais fortemente associado para o menos fortemente associado. Os "*n*" primeiros candidatos da lista são selecionados para serem utilizados no próximo passo.

O terceiro passo constitui da avaliação da lista EM gerada por um especialista humano que retira manualmente os falsos positivos identificados pelo processo automatizado. Dessa forma, a abordagem proposta por esses autores se caracteriza por ser uma extração de EM semi-automática. Esses mesmos autores propõem o uso de uma técnica de extração de uma amostra aleatória e representativa do *corpus* em vez do conjunto completo dos documentos que visa minimizar o trabalho intelectual de um especialista.

Yagunova e Pivovarova (2010) trabalham para identificar a natureza das *collocations* no idioma russo. Esses autores utilizam medidas estatísticas que permitem identificar automaticamente as *collocations* no texto e ranqueá-las de acordo com o seu grau de estabilidade ou correspondência com o valor da medida escolhida. A lista das *collocations* identificadas é um reflexo das características linguísticas e extralinguísticas encontradas nos textos analisados, sendo que a ordenação pela relevância depende da técnica de medida de associação estatística empregada no processo de ranqueamento. Os autores utilizaram uma coleção de textos totalizando mais de 60 milhões de *tokens* extraídos de um site de notícias do período de abril a dezembro de 2009. O método automatizado envolveu inicialmente uma marcação morfológica da coleção, seguida por uma análise sintática a fim de remover parcialmente os homônimos.

O conteúdo resultante desse processamento foi separado em fragmentos de texto tomando como base os marcadores de pontuação. O próximo passo foi identificar as cem *collocations* melhor ranqueadas obtidas através das medidas estatísticas Mutual Information (MI) e T-Score que ocorram numa frequência acima de quarenta vezes. Finalmente, os resultados obtidos foram manualmente analisados, comprovando as hipóteses propostas: o método MI permite distinguir nome de objetos, termos e combinações complexas refletindo a área de conhecimento ou assunto do texto; enquanto que T-Score, trabalha melhor para distinguir as propriedades estilísticas do texto, ou seja, “combinações linguísticas gerais” (derivadas de palavras funcionais e palavras discursivas) e "construções agrupadas". Um dos problemas que esses autores identificaram no uso da técnica MI é que a medida depende do tamanho do *corpus* analisado e ela tende a sobre estimar ruídos, tais como palavras estrangeiras.

Outra forma de tratar o problema é através da abordagem simbólica, a qual busca encontrar o sentido sintático, morfológico e pragmático do texto baseando-se em um dicionário controlado de palavras e em um conjunto de regras visando à interpretação do mesmo. Nesse caso, o processamento é fortemente dependente do idioma e do domínio do *corpus*, enquanto a abordagem estatística procura dar um tratamento ao texto através do reconhecimento de padrões de comportamento baseados na frequência de co-ocorrência das palavras.

Cazolari et al. (2002) utilizam uma abordagem focada nas EM que são produtivas por um lado e, por outro, demonstram regularidades que possam ser generalizadas para as classes de palavras com propriedades semelhantes. Particularmente, eles buscam encontrar dispositivos gramaticais que permitam a identificação de novas EM motivados pelo desejo do reconhecimento o mais automatizado possível na aquisição das EM. Nesse sentido, a pesquisa desses autores estudou em profundidade dois tipos de EM: os verbos de suporte e os substantivos compostos (ou complexos nominais). Segundo eles, esses dois tipos de EM estão no centro do espectro de variação em composicionalidade que pode ser observado pela coesão interna juntamente com um elevado grau de variabilidade em lexicalização e variação dependente do idioma.

A pesquisa conduzida por Villavicencio et al. (2010) busca extrair as EM, combinando duas abordagens distintas: a abordagem associativa e a abordagem baseada em alinhamento lexical[5]. Na primeira, as medidas de associação são aplicadas para todos os bigramas e trigramas gerados a partir do *corpus* e o resultado dessas medidas é utilizado para avaliação. A segunda abordagem extrai de forma automatizada as EM, tomando como base os alinhamentos lexicais das versões de um mesmo conteúdo escrito nos idiomas português e inglês. O alinhamento final é gerado a partir da interseção dos alinhamentos em ambos os sentidos. Antes de usar a técnica de alinhamento, o *corpus* é etiquetado morfossintaticamente a fim de aplicar filtros de categorias gramaticais na lista inicial de EM extraídas. Para combinar os resultados obtidos pelas duas abordagens os autores utilizaram as redes bayesianas.

A abordagem de alinhamento lexical verifica se a EM encontrada em um documento escrito em um idioma, também ocorre na versão correspondente escrita em outro idioma. Para ser possível essa análise, os documentos necessitam estar alinhados através da correspondência das palavras entre as diferentes versões expressas em idiomas distintos. Entretanto, para ser possível o alinhamento é necessário que os documentos sejam analisados a partir de seus aspectos morfológicos tratados através de um pré-processamento de etiquetação[6]. Dessa forma, as classes gramaticais são utilizadas como informação adicional no processo de identificação das EM.

Na pesquisa desenvolvida por Zhang et al. (2009), é proposto um método denominado Enhanced Mutual Information and Collocation Optimization (EMICO) para extrair EM com foco nas entidades nomeadas. Estas se caracterizam por serem compostos contíguos com duas a seis palavras que descrevem conceitos com padrões sintáticos mais estáveis.

[5] Dois textos escritos em idiomas distintos são considerados como alinhados quando eles possuírem marcas que identifiquem os pontos de correspondência entre o texto original e a sua tradução.

[6] Programa de computador conhecido genericamente como etiquetador de categorias gramaticais. Gera uma saída, normalmente em XML, associando cada palavra à sua classe gramatical: substantivo, verbo, artigo, etc.

Esses autores empregam essa técnica em processamento de mineração de textos e a comparam com as técnicas de indexação tradicional do modelo de espaço vetorial conjeturando que o uso da EM para interpretação semântica do texto produz melhores resultados do que os modelos estatísticos e semânticos que lidam com palavras individuais.

No trabalho de Chen, Yeh e Chau (2006), apresenta-se um sistema alternativo para extrair EM, por considerar que os métodos estatísticos tradicionais lidam com uma grande quantidade de dados ruidosos e que consumem muito tempo de processamento. Desse modo, eles elaboraram um experimento baseado em um *corpus* constituído de 308 documentos escritos em chinês tradicional, que considera cada ideograma como sendo uma palavra e aplicaram uma metodologia dividida em quatro passos:

- Gerar segmentos – nesse passo, os autores utilizaram uma pequena e pré-definida lista de *stop words* como entrada inicial. O objetivo é utilizar as *stop words* como *tokens* para o processo de separar o documento em segmentos de texto. Durante esse estágio, a frequência do segmento de texto é calculada para o respectivo documento e também para o conjunto dos documentos.
- Calcular o peso dos segmentos de texto – no cálculo do peso são consideradas a quantidade de palavras contidas no segmento de texto, e as respectivas frequências do segmento no documento e no *corpus*.
- Fragmentar os segmentos de texto – nesse passo, é aplicada uma regra que considera que um segmento de texto só pode ser separado se, e somente se, ele estiver contido dentro do outro e, ao mesmo tempo, o seu peso for maior que do outro segmento. Esse procedimento faz com que segmentos de maior peso tenham menor possibilidade de serem segmentados, pois já representam uma EM.

Selecionar as EM – por fim aplica-se um filtro ao valor calculado do peso do segmento. Esse processo limita as EM extraídas como sendo somente aquelas que tiverem os maiores pesos tomando como base o valor do limite informado.

Um trabalho bastante afim encontrado na literatura, não pelo método empregado, mas pelo uso intuitivo do mesmo conceito é o modelo de verificação de aspectos combinados apresentado por Roussinov (2012). Diferentemente das técnicas de consulta convencionais, de expansão ou de tradução que estão limitadas a apenas buscar nos documentos os unigramas de forma independente, esse trabalho se caracteriza por propor uma função de similaridade para ranquear as respostas obtidas através da apropriação de dois aspectos: os aspectos presentes A_p e os aspectos faltantes A_f, sendo esses condicionados à presença dos aspectos presentes. Ou seja, considera a consulta do usuário como uma sequência de palavras (*n*-gramas) de tal forma que os termos de busca são avaliados também como termos dependentes. Desse modo, o método empregado considera os dois aspectos que são automaticamente identificados e tratados. O primeiro A_p, a presença do termo no documento, é verificada pelo casamento exato. O segundo A_f considera uma estimativa dos aspectos que não se manifestam explicitamente nos documentos da coleção. Desse modo, a presença implícita de um aspecto é obtida pela sua presença explícita, estatisticamente prevista. A predição é baseada nos indicadores contidos no texto. Esses indicadores são sequências (*n*-gramas) de palavras no documento composto de até três termos. Para cada indicador *i* oriundo do documento, o algoritmo estima $P(A_f \mid i, A_p)$, que é a probabilidade de ocorrência dos aspectos faltantes condicionada à ocorrência conjunta de A_p e *i*. Em geral, cada probabilidade condicional é estimada como mostrada na expressão (2.1).

$$P\left(A_f \mid i, A_p\right) = \left(\frac{NUM}{DEN}\right) \quad (2.1)$$

Sendo NUM (numerador), o tamanho da amostra dos documentos em que o aspecto faltante A_f ocorre com A_p e *i,* e DEN (denominador), o tamanho da amostra de documentos em que o indicador *i* ocorre juntamente com A_p. Para estimar o tamanho da amostra, foi utilizada a um subconjunto World Wide Web, (a Wikipedia) extraída através da API do motor de busca Bing da Microsoft.

O modelo empregado se baseia apenas nas características booleanas da linguagem de consulta, especificamente sobre os operadores AND (conjunção) e NEAR (proximidade). O objetivo é captar a presença conjunta de i e A_p através do operador NEAR, por exemplo, "Antartica NEAR station", já para captar a presença de A_f, A_p e i várias combinações diferentes dos operadores NEAR e AND são utilizadas. Desse modo, para cada consulta do usuário, o algoritmo analisa milhares de tais possíveis indicadores.

Um trabalho recente apresentado por Rayson et al. (2009) faz um retrospecto histórico da pesquisa sobre EM e destaca alguns dos principais grupos de pesquisa que atuam no mundo. A meta é facilitar o trabalho daqueles que estão desenvolvendo pesquisas nessa área. Esses autores relatam que, no início dos anos 1990, as EM passaram a receber maior atenção dos pesquisadores de PLN, nesse sentido eles citam a influência dos trabalhos de Smadja (1993), Dagan e Church (1994), Wu (1997); Daille (1995), dentre outros. Eles destacam que um importante marco ocorreu a partir de 2001 com o interesse despertado pelo Centre for the Study of Language and Information (CSLI), da universidade de Stanford, o qual visa investigar um meio para codificar a variedade de EM em gramáticas de precisão. Outro importante trabalho tem sido conduzido pela universidade de Lancaster, o qual resultou em uma grande coleção de termos semanticamente anotados. Com esses recentes desenvolvimentos de *corpus* linguísticos, pesquisadores de diferentes áreas têm se juntado, possibilitando desenvolver trabalhos que abordam as EM, a partir de diferentes perspectivas. Nesse sentido, desde 2003, essa comunidade de pesquisadores vem organizando workshops dentro de importantes congressos da área da CC, tais como ACL e LREC. Esse interesse reflete a importância desse tema dentro do campo de pesquisa da PLN. Sendo que, nos primeiros workshops, o tópico mais recorrente estava relacionado ao processo de identificação e extração automática de EM, surgindo propostas a partir de diferentes perspectivas, baseadas em análise linguística e em medidas estatísticas.

Com o decorrer das pesquisas e objetivando desenvolver algoritmos mais eficientes, as pesquisas se voltaram para a busca do entendimento mais profundo das propriedades estruturais e semânticas das EM, tais como padrões morfo-sintáticos, composionalidade semântica, comportamento semântico em diferentes contextos, propriedades de transformação de EM entre idiomas, etc. Após duas décadas de esforços, a comunidade de linguística computacional vem construindo valiosos recursos e ferramentas para aplicação no mundo real do PLN, tais como mapeamento de termos de consultas para sinônimos tanto para o uso em sistemas de recuperação da informação, quanto para sistemas de tradução automática e mineração de dados em textos que demandam a identificação de conceitos multipalavras.

Na busca de extrair sentido de um texto a partir de suas partes mais relevantes, outras estratégias têm sido adotadas. Nesse sentido, destaca-se o uso dos sintagmas nominais utilizados como descritores de busca, abordados pelos trabalhos de Kuramoto (1995) e Souza (2005) e da pesquisa de Maia & Souza (2010) que buscam utilizar os sintagmas para agrupar documentos correlatos. O método de identificação dos sintagmas nominais utiliza uma abordagem baseada na linguística, em que as palavras do texto são previamente etiquetadas a fim de identificá-las em classes gramaticais que servirão de base para a extração dos sintagmas. Entretanto, a identificação dos sintagmas exige um processamento analítico em profundidade das sentenças que demanda um exaustivo processamento computacional baseado em regras dependentes do idioma.

No contexto deste trabalho, buscam-se identificar as partes semanticamente relevantes do texto representadas pelas EM de forma independente de idioma. O método proposto neste trabalho obtém as EM, a partir de um algoritmo determinístico que utiliza aspectos da estrutura física dos documentos denominado Heudet. Para verificar e eficiência do método extrair-se-ão as EM obtidas pela técnica Heudet comparado com os resultados obtidos por treze diferentes técnicas de medidas de associação estatística produzidas pelo pacote NSP, Pedersen (2011).

3 FUNDAMENTOS CONCEITUAIS

O objetivo desta seção é apresentar alguns conceitos que nortearam o processo de construção do método heurístico proposto para processar a identificação das EM.

3.1 - Fundamentos linguísticos

Cintra (1983) apresenta o termo linguagem como sendo uma faculdade natural, enquanto que o termo língua refere-se a um caso particular de linguagem.

> A linguagem é uma representação simbólica que expressa uma função psicossocial complexa. Corresponde a uma manifestação intelectual e multiforme dos seres, que recobre inúmeras formas de significar: linguagem verbal (oral e escrita), a pictórica, a musical, a cinética, a mímica, a documentária, etc. (CINTRA, 1983 p. 7)

Por ser a linguagem a forma utilizada para mediar as relações humanas na elaboração cognitiva do pensamento e na comunicação para a troca de informações, como era de se esperar, ela também é utilizada para registrar a fala e o pensamento na forma de texto. Este trabalho abrange apenas um recorte desse tema, ao lidar somente com o texto escrito, dentro de um domínio específico, expresso em linguagem natural e armazenado no formato digital.

Cipro Neto e Infante (2009, p. 9) definem **linguagem** como sendo a capacidade de se comunicar por meio de uma língua. Eles definem **língua** como sendo um sistema de signos convencionados e utilizados por membros de uma comunidade. Já o **signo** como sendo um elemento representativo com aspectos do significante e do significado unidos por um todo indissociável. Portanto, o conhecimento de uma língua demanda conhecer a identificação de seus signos e o uso adequado de suas regras combinatórias.

Na perspectiva de Cintra (1983 p. 7-13), que apresenta uma definição em um nível mais abstrato, signo é uma unidade que está no sistema e na consciência do falante. Os signos são compostos pelo léxico da língua e pela palavra. O léxico é não-quantificável, composto das unidades que alimentam o vocabulário. Ao se criar uma nova entrada no vocabulário, tem-se um vocábulo.

Os vocabulários são compostos de dois tipos de unidades: a) Morfema Lexical – contém o significado lexical, ou seja, expressam o "suporte de conceito" do mundo biossocial e b) Morfema Gramatical – contém significado gramatical, por isso mesmo é denominado "indicador de função". Alguns autores consideram a palavra como sendo uma unidade formal composta de morfemas definidos dentro de uma língua, outros como uma unidade de texto. Na prática, não há um consenso entre os linguistas sobre qual é a definição de palavra ou termo. A linguagem pode ser estudada essencialmente perante a perspectiva gramatical, semântica e pelo sistema que relaciona ambas como as subseções subsequentes mostrarão.

Cipro Neto e Infante (2009, p. 14-16) apresentam uma definição de **gramática** como sendo a designação para um conjunto de regras que garantem o uso modelar da língua. Ou seja, a gramática estabelece a norma culta e as regras que asseguram o uso correto da língua. O estudo da gramática é convencionalmente dividido em:

- **Fonologia** – Estuda os fonemas ou sons da língua e as sílabas que esses fonemas formam;
- **Morfologia** – Estuda a estrutura, a formação e os mecanismos de flexão das palavras, além de dividi-las em classes gramaticais;
- **Sintaxe** – Estuda as formas de relacionamentos entre as palavras ou entre orações, a qual inclui e regência, a colocação pronominal e a concordância.

Por ser a fonologia a parte que estuda os sons da língua, ela extrapola o escopo deste estudo, delimitada ao texto escrito. Portanto, tratar-se-á apenas de alguns dos aspectos da morfologia e sintaxe que são relacionados à compreensão do tema. Conforme define Cipro Neto e Infante (2009, p. 73-74), a morfologia estuda a estrutura, formação, flexão e classificação das palavras. Sendo que cada uma delas é formada por **morfemas**, que são os elementos que a constituem.

Esses elementos indecomponíveis são unidades de significação mínima que agregam significado à palavra. Segundo Faraco e Moura (1990, p. 132-138), os principais processos de formação das palavras são a derivação e a composição. A **derivação** é o processo de formação da palavra a partir de outra que já existe na língua. A **Composição** refere-se à junção de duas ou mais palavras ou radicais para formação de uma nova palavra. Os autores citam outros processos de formação de palavras como sendo: hibridismo, onomatopeia, siglonimização e abreviação vocabular.

Cintra (1983, p. 6) define morfologia como a disciplina que sintetiza parcialmente aspectos da semântica e da sintaxe, por se encarregar da identificação das partes da palavra e de suas condições de ocorrência.

Faraco e Moura (1990, p. 144-147) definem que cada palavra tem uma finalidade no ato de comunicação oral ou escrita. De acordo com essa finalidade as palavras se enquadram nas seguintes classes gramaticais, tais como: substantivo, adjetivo, verbo, pronome, etc.

As classes de palavras são normalmente divididas em duas. A de categoria léxica ou aberta tais como substantivos, verbos, adjetivos e advérbios os quais possuem um grande número de membros, e para os quais novas palavras são comumente adicionadas. E a categoria funcional ou fechada, que possui um número finito de palavras e tem claro uso gramatical, na qual enquadram-se os pronomes, os artigos, as preposições e as conjunções.

As palavras podem ser flexionadas mudando sua terminação para exprimir outros significados. **Flexão** é modificação sistemática da forma raiz por meio de prefixo ou sufixo para indicar distinções gramaticais tipo singular e plural. Flexão não muda a classe da palavra ou altera o seu significado, mas varia características tais como tempo, número e plural. Toda forma flexionada de uma palavra é frequentemente agrupada como manifestações de um morfema. A tipologia das flexões é relacionada a seguir:

- Flexão de número – é a mudança da terminação para indicar singular ou plural;
- Flexão em grau – é terminação utilizada para indicar tamanho nos substantivos e intensidade nos adjetivos e advérbios;

- Flexão de Tempo – existe apenas para os verbos, e indica o momento da ocorrência do fato presente, passado ou futuro;
- Flexão de modo – só existe para os verbos e serve para indicar as diferentes atitudes do emissor em relação ao fato que se quer expressar. Sendo três as possibilidades: indicativo, subjuntivo ou imperativo;
- Flexão de pessoa – permite flexionar o verbo de acordo com a pessoa gramatical: emissor, receptor, ou de que/quem se fala.

Cintra (1983, p. 6) define **sintaxe** como a disciplina que se ocupa das relações que se estabelecem a partir da organização sintagmática dos elementos e funcionamento do significado do signo, visto como elemento do sistema lexical de uma língua. Para Faraco e Moura (1990, p. 307-310) uma mensagem linguística é formada por palavras e o estudo da combinação e relação entre as palavras é denominado sintaxe. A análise sintática estuda um texto a partir de suas partes tais como: frase ou sentença, oração, sintagma, etc.

A análise sintática é uma técnica empregada no estudo da estrutura de uma sentença, seus períodos e orações. É um passo importante para o entendimento (semântica) de uma sentença em linguagem natural. Somente vocábulos não garantem o entendimento de uma sentença, é importante que a sua estrutura sintática seja analisada. Na análise sintática de uma oração em português deve levar em conta os seguintes sintagmas: termos essenciais (sujeito e predicado), termos integrantes (complementos verbal e nominal) e termos acessórios (adjunto adverbial, adjunto adnominal e aposto). A análise do período, por sua vez, deve considerar o tipo de período (simples ou composto), sua composição (por subordinação, por coordenação) e a classificação das orações (absoluta, principal, coordenada ou subordinada).

As primeiras formas utilizadas pelos bibliotecários para recuperar conteúdos foram proporcionadas pelas técnicas de classificação através da codificação apoiada em linguagens documentárias.

Essas linguagens, através de uma gramática com regras e instruções bem definidas, orientam o trabalho do indexador para extrair os descritores que melhor descrevem o conteúdo do documento. Nesse sentido, a linguagem documentária concretiza a capacidade simbólica do homem, mediante organização de seus termos e regras em um sistema próprio. A classificação dos documentos, realizada de forma manual por especialistas, é um processo intelectual exaustivo, pois demanda a leitura de cada texto pelo indexador a fim de selecionar as palavras-chave ou descritores, que representem o documento numa forma compatível com uma dada linguagem documentária.

Segundo Cintra (1983, p. 6), existem dois procedimentos básicos para a depreensão dos descritores: a apreensão instantânea das unidades de informação e a apreensão por análise. Nesse sentido, devem ser observadas pelo indexador as partes relevantes do texto tais como: título, resumo, introdução e conclusão, os termos que possuem maior frequência de ocorrência e que traduzem melhor a percepção do indexador no sentido do texto. O autor destaca ainda a dificuldade em expressar através da linguagem documentária as nuances decorrentes da linguagem natural tais como: a polissemia, a homonímia, a sinonímia e os modos e expressões de relações complexas.

Observados esses aspectos para escolha do descritor, deve-se ter especial atenção, nessa operação de delimitação dos significados, com a sua semântica contextual. Nesse sentido, o autor define **semântica** como a disciplina que se ocupa do sentido ou da significação dos elementos; e **sintaxe** como a disciplina que se ocupa das relações que se estabelecem a partir da organização sintagmática dos elementos; e **morfologia** como a disciplina que sintetiza parcialmente aspectos da semântica e da sintaxe, encarrega-se da identificação das partes da palavra e de suas condições de ocorrência.

Os descritores de um texto podem ser analisados em termos de: (1) os sintagmas de símbolos notacionais (números, letras, pontuação, marcas) isto é, unidades resultantes da combinação de formas menores em unidades de nível superior. (2) lexemas como combinação de fonemas, ou seja, como combinação de unidades capazes de promover a distinção entre os signos da língua. Nas linguagens de indexação, os signos, em geral, lidam com os termos de uma

língua particular, fixando significados de forma a anular o sentido simbólico do signo linguístico. Dessa forma, os signos documentários exigem do indexador uma percepção do contexto para que a tradução tenha um forte poder de partilha na comunidade à qual o documento se destina.

3.2 Expressões multipalavras

Inicialmente, faz-se necessário definir três conceitos fundamentais para este trabalho: Expressão Multipalavras (EM), Termo Técnico-Científico (TCC) e Termo Multipalavras (TM).

A definição de EM é ampla, pois engloba diversos fenômenos distintos como compostos nominais, expressões idiomáticas e termos compostos. As EM são necessariamente compostas por mais de uma palavra.

Os TCC e os TM são fenômenos linguísticos ligados ao texto técnico-científico, definidos como locuções que possuem estatuto terminológico. Sendo que os TCC podem ser unidades lexicais únicas, aceitam pouca variabilidade (morfológica, raramente sintática) e representam um único conceito. Enquanto os TM não correspondem ao conceito de fraseologia do domínio, são altamente flexíveis e normalmente possuem uma estrutura complexa que associa mais de um conceito. A seguir apresenta-se uma definição para cada um desses termos citados no trabalho de Ramisch (2009, p. 65):

- EM é um conjunto de duas ou mais palavras com semântica não-composicional, ou seja, o sentido do sintagma não pode ser compreendido totalmente através do sentido de suas componentes (SAG et al., 2002).
- TCC é uma unidade lexical ou multilexical com significado não ambíguo quando empregada em textos especializados, ou seja, a terminologia de um domínio é a representação linguística dos seus conceitos (KRIEGER, FINATTO, 2004).
- TM é um termo composto por mais de uma palavra. (SAN JUAN et al., 2005; FRANTZI et al., 2000).

Na realidade, não existe uma definição formal consensual na literatura sobre EM. Em linhas gerais, considera-se que as EM são formações compostas de duas ou mais palavras que, quando associadas, possuem uma expressividade semântica mais forte do que quando cada um de seus termos são postos separadamente. Para Sag et al. (2002 p. 2) EM são: "interpretações idiossincráticas que cruzam os limites (ou espaços) entre as palavras".

Segundo Ranchhod (2003, p. 2), as expressões fixas são objetos linguísticos que apresentam divergências terminológicas e a ausência de critérios de análise que os levaram ser consideradas como objetos linguísticos excepcionais, não integráveis na gramática das línguas. Entretanto, tem ocorrido um crescente interesse, sobretudo na área de PLN, afinal essas formas fixas são tão numerosas em qualquer tipo de texto, que não podem ser ignoradas. Portanto, essas características das EM as tornam relevantes no tratamento dos recursos lexicais, os quais são importantes insumos informacionais para muitas aplicações relacionadas ao PLN, tais como: tradução automática, sumarização de texto, etc.

Calzolari et al. (2002, p. 1934) corroboram com a classificação apresentada por Sag et al. (2002) e ainda incluem um "etc" no final. Ou seja, como os próprios autores definem, EM é utilizada para descrever diferentes, mas relacionados fenômenos, que podem ser descritos como uma sequência de palavras que agem como uma unidade em algum nível de análise linguístico e que apresentam alguns ou todos dos seguintes comportamentos: reduzida transparência sintática e semântica; redução ou ausência de composicionalidade; mais ou menos estável; passível de violação de alguma regra geral sintática; elevado grau de lexicalização (dependendo de fatores pragmáticos); alto grau de convencionalidade. Ainda segundo esses mesmos autores, as EM estão situadas na interface entre a gramática e o léxico. Eles apresentam também algumas das causas das dificuldades ocorridas no âmbito teórico e computacional para o tratamento das EM, como sendo: a dificuldade de estabelecer limites claros para o domínio das EM; a falta de léxicos computacionais de tamanho razoável para auxiliar no PLN; perante a perspectiva multilingue, muitas vezes não é possível encontrar uma

correspondência direta lexical equivalente; dificuldade generalização dos léxicos (geral e terminológico) para um contexto específico.

Segundo Moon (1998, citada por VILLAVICENCIO et al.), as EM são unidades léxicas formadas por um amplo contínuo entre os grupos composicionais e os não-composicionais ou idiomáticos. Nesse contexto, entende-se por expressão composicional aquelas que, a partir das características de seus componentes, determinam as características do todo. E não-composicional ou expressões idiomáticas aquelas cujo significado do conjunto de palavras nada tem a ver com o significado de cada uma das partes. A ocorrência das EM nas línguas, de maneira geral, são muito frequentes conforme é apontado por Biber et al. (1999, citado por Wang e Liu 2011). Segundo esses autores, na língua inglesa, as EM representam de 30% a 45% do idioma falado e cerca de 21% da escrita acadêmica. Entretanto, esses números podem estar ainda subestimados, se se considerar que o surgimento de novas EM ocorrem com frequência, como por exemplo: computação em nuvens, energia limpa, etc. Wang e Liu (2011) reafirmam ainda que as EM são uma questão ainda a ser mais bem resolvida pelas aplicações que lidam com PLN.

Após revisar a literatura na busca de encontrar uma definição consensual para EM, percebe-se que o termo tem um uso genérico o qual engloba vários conceitos ou subtipos conforme descritos anteriormente. Desse modo, empregam-se diferentes métodos ora estatísticos, ora linguísticos, ora uma combinação de ambos para identificar as EM de forma mais estrita. Portanto, faz-se necessário apresentar a definição de EM a qual será utilizada neste trabalho – EM são expressões fixas que co-ocorrem em um documento com uma frequência acima de um limite predefinido, considerando-se as características da estrutura do documento.

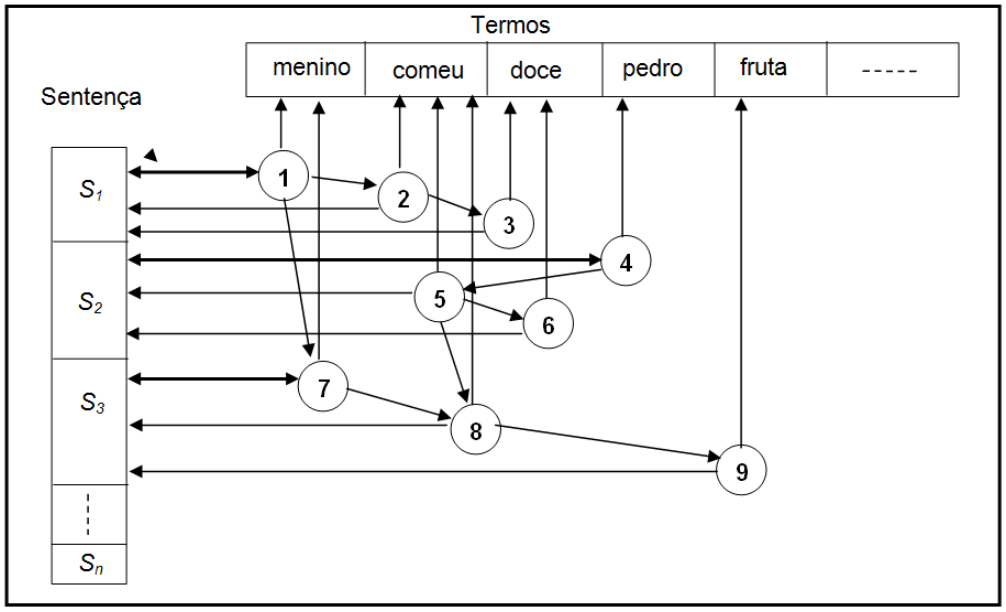
4 METODOLOGIA

Nesta seção, o objetivo é apresentar como funciona o método Heudet, o qual foi implementado em um programa de computador desenvolvido na linguagem C++ pelos autores. Por fim, o resultado produzido pelo método

Heudet foi comparado com o resultado produzido pelo pacote NSP proposto por Perdersen (2011).

Para realizar esse processamento, considera-se que o texto já foi pré-processado por um parser[7] a fim de quebrá-lo em sentenças, sendo que cada sentença e quebrada novamente em uma lista de termos normalizados[8]. Todas essas sentenças e termos identificados são também numerados sequencialmente e processados, a fim de ordená-los em uma estrutura de dados em memória que permita a extração das EM. Essa técnica conhecida como Positional Index foi descrita por Manning, Raghavan & Shütze (2009, p. 41-43). Ela consiste em adicionar na estrutura de lista invertida indexada em memória a(s) posição(ões), controlada(s) a partir de uma sequência numérica, com a localização do número da sentença e do número da posição do termo na sentença. Ou seja, uma precisa localização de onde o termo se encontra no documento. Essa estrutura de dados em memória utilizada para viabilizar o processo de extração é mostrada na Figura 1.

Figura 1. **Representação da estrutura de dados criada para extrair EM.**



Fonte: Elaborada pelos autores.

[7] Programa de computador que recebe como entrada um texto puro, na forma de uma sequência de caracteres sem formatação, e processa a divisão desse um uma lista de termos ou palavras.
[8] Tratamento dado a um termo ou palavra no sentido de padronizar a lista de palavras extraídas de um texto após processar o parser. Esse processo inclui: descartar palavra muitos comuns, as chamadas *stop words*; transformar todas as letras que compõem as palavras em minúsculas, etc.

Para entender como essa estrutura dados é utilizada no processo automatizado de extração de EM apresenta-se a seguir um exemplo o qual utiliza um pequeno documento de referência composto por três sentenças sem formatação. Desse modo, o conteúdo do documento de referência d_{ref} da busca é composto pelas sentenças S_1, S_2, S_3. Ou seja, d_{ref} = { S_1, S_2, S_3 }, conforme mostrado em (3.1).

$$d_{ref} = \begin{matrix} S_1 \rightarrow \text{O menino comeu o doce.} \\ S_2 \rightarrow \text{Pedro comeu o doce.} \\ S_3 \rightarrow \text{O menino comeu a fruta.} \end{matrix} \qquad (3.1)$$

Considera-se também, que, após executada o parser no documento, tem-se como resultado o conjunto de termos normalizados V = { T_1, T_2, T_3, T_4, T_5 }, conforme mostrado na Tabela 1.

Tabela 1. **Termos normalizados**

Identificação	Termo	Obtido da sentença
T_1	Menino	S_1 e S_3
T_2	Comeu	S_1, S_2 e S_3
T_3	Doce	S_1 e S_2
T_4	Pedro	S_2
T_5	Fruta	S_3

Fonte: Elaborada pelos autores.

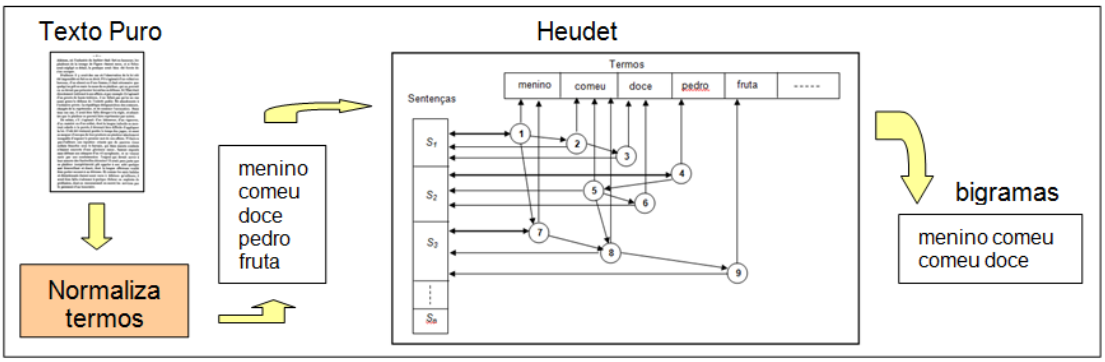
Os artigos “o” e “a” existentes nas sentenças foram descartados por serem considerados *stop words*.

Finalmente, considera-se que o conjunto de nós N = {1, 2, 3, 4, 5, 6, 7, 8, 9} representa cada uma das nove palavras do texto a ser indexado considerando o exemplo proposto.

Como o processamento é realizado na ordem em que as sentenças são lidas, ao ler a sentença S_1 serão processados os termos T_1, T_2 e T_3 referenciados pelos três nós 1, 2 e 3 respectivamente. Ao ler a sentença S_2 serão processados os termos T_4, T_2 e T_3 referenciados pelos nós 4, 5, 6 e, assim, sucessivamente.

Após processadas todas as sentenças, a estrutura proposta permite identificar quais são as frases existentes na cadeia de caracteres e também em quais sentenças um determinado termo ocorre. Dessa forma, para extrair as EM, o algoritmo Heudet percorre as sentenças verificando para cada palavra, que ainda não foi processada, quais são as suas adjacentes. Em seguida, verifica-se a frequência em que a repetição dos termos adjacentes ocorre. Os termos, com a frequência de repetição maior ou igual a uma quantidade informada em parâmetro de entrada para o processamento do algoritmo, serão adicionados na lista dos bigramas identificados. Os demais são descartados. O valor utilizado desse parâmetro nos experimentos foi igual a quatro. Desse modo, somente os bigramas com frequência de co-ocorrência para cada documento superior ou igual a esse valor informado é que foram considerados para representar semanticamente o documento. A figura 2 apresenta um esboço desse processo.

Figura 2. **Esboço do processo de identificação de *n*-gramas do *software* elaborado.**

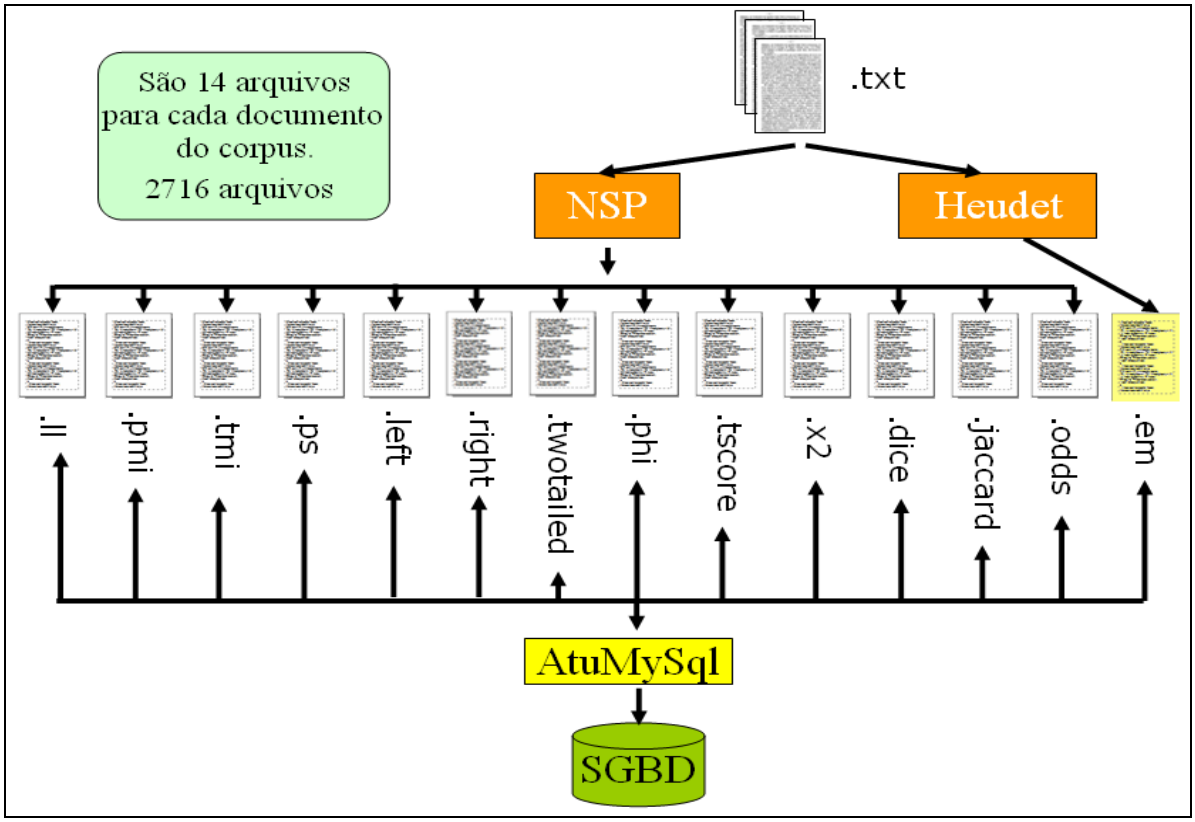


Fonte: Elaborada pelos autores.

Um ponto que deve ser destacado é que, apesar de esse processamento extrair apenas bigramas, não significa que expressões com *n*-gramas não sejam consideradas. Na prática, o processamento pode lidar com qualquer quantidade de termos consecutivos que tenham uma frequência observada igual ou acima da quantidade definida em parâmetro. Isso pode ser feito, pois qualquer conjunto de *n*-gramas pode ser transformado em pares de bigramas. No exemplo a seguir, o pentagrama "Universidade Federal de Minas Gerais" é transformado em 3 bigramas: "Universidade Federal", "Federal Minas", "Minas Gerais". O termo "de" é descartado por ser uma *stop word.*

Após finalizar a implementação do método Heudet, tornou-se necessário comparar o resultado dos bigramas obtidos com o produzido por outras técnicas a fim de avaliar o seu desempenho. Nesse sentido, o pacote NSP foi escolhido por disponibilizar um processo de identificação de bigramas através do processamento de treze medidas de associação estatísticas distintas. O *corpus* utilizado é composto por 194 artigos publicados no Enancib de 2010, totalizando 682.537 termos normalizados, sendo 46.888 distintos. Desse modo, o resultado produzido pelo processamento dessas 14 técnicas produziu 2716 arquivos textos, os quais foram inseridos em tabelas do MySql através do processamento de um programa "AtuMySql" elaborado pelos autores. O Objetivo é de facilitar a comparação dos resultados obtidos utilizando consultas SQL. A figura 3 apresenta um esboço desse processamento realizado, sendo que a extensão (.em) representa os arquivos gerados pelo processamento da técnica Heudet.

Figura 3. **Esboço do processamento de comparação da EM extraídas pelas 14 técnicas distintas.**



Fonte: Elaborada pelos autores.

5 RESULTADOS

Para todo o processamento descrito pela sessão anterior foi utilizado quatro como sendo a frequência de co-ocorrência do bigrama no documento, para todas as técnicas empregadas. Após carregar os resultados no banco de dados foram realizadas consultas SQL a fim de comparar os conteúdos obtidos por cada uma dessas técnicas. Constatou-se que as EM identificadas por todas as técnicas do pacote NSP foram as mesmas, totalizando 7.832. A única diferença no resultado produzido por essas técnicas foi a ordenação dos bigramas em função da relevância calculada. Já pela técnica Heudet foram obtidas 7.734 EM. Ao analisar o porque dessas diferenças verificou-se que, como as técnicas de parser utilizadas pelo NSP e Heudet são distintas era necessário normalizar o resultado. Em primeiro lugar, enquanto Heudet desconsidera as palavras formadas por apenas uma letra, o pacote NSP as mantém. Portanto, foram descartadas 86 EM geradas exclusivamente pelo pacote NSP que continham no bigrama pelo menos uma das palavras formadas por apenas uma letra. Afinal palavras de apenas uma letra são consideradas de pouco teor semântico. Desse modo, as EM geradas exclusivamente pelo pacote NSP ficaram reduzidas em 7.746 na busca de padronizar os resultados. Ao analisar quantas são as EM idênticas para esses dois conjuntos verificou-se que correspondiam à 7.636. Ao retirar essa parte comum dos conjuntos obtidos pelo pacote NSP e Heudet foram encontrados 110 casos de EM obtidas de forma exclusiva pelo pacote NSP e 98 exclusivas pela técnica Heudet. Ao analisar esses casos tornou possível comparar o resultado entre os métodos.

Dos 110 casos exclusivos de NSP:

- 83 casos (75,5%) erro gerado por não avaliar a estrutura do documento;
- 18 casos (16,4%) erro no delimitador de palavra (&, % etc)
- 4 casos (3,6%) diferença de critério na tokenização (@, < >)
- 5 casos (4,5%) falso positivo do (.) algoritmo Heudet.

Já para os 98 casos exclusivos de Heudet a diferença está relacionada ao de critério de quebra de palavras adotado pelos dois programas para os

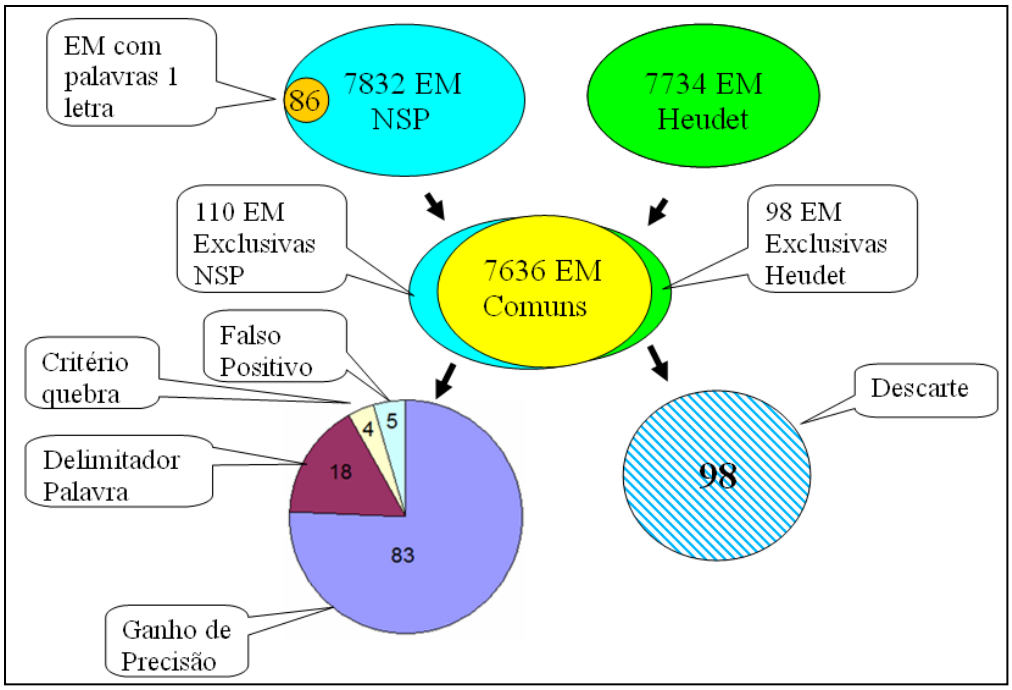
caracteres (@, < >). Enquanto Heudet considera os caracteres @, < e > como parte da palavra, o pacote NSP considera como separador da palavra, portanto, divide esse termo em duas palavras. Cabe observar que Heudet adotou essas estratégias por considerar semanticamente o @ como parte do nome de um endereço de e-mail. Portanto, semanticamente não faz sentido quebrá-lo em duas palavras, e no caso de < e > são caracteres usados como delimitadores de tags em documento html, que pelo mesmo motivo são incorporados à palavra. Portanto, essas diferenças ocorrem apenas pelas definições dos critérios de separação de palavras adotados por cada um dos algoritmos, podendo desse modo, ser desconsideradas.

O que realmente importa para avaliar o ganho semântico da técnica proposta são os 83 casos identificados pelo NSP, que ao considerar a estrutura semântica do texto composto apenas como um conjunto de termos dispostos como em um saco de palavras do inglês "*bag of words*" identifica EM que não possuem ligação semântica no texto. Vejamos um exemplo:

> Melhorar o bem estar da humanidade é uma tarefa das ciências.
> Da informação surge o insumo para a tomada de decisão.

O algoritmo do NSP pode considerar nas sentenças acima "Ciência Informação" como um bigrama, o que para o algoritmo implementado em Heudet não é considerado. Afinal Heudet avalia a estrutura física do texto dividido em sentenças. Desse modo o ponto final após a palavra "ciências" quebra o link semântico dessa palavra com a que ocorre na próxima sentença de tal forma que termos pertencentes às sentenças adjacentes nunca formam uma EM, devido à separação caracterizada pela divisão das sentenças no texto, as quais são expressas pelo ponto final, ponto de exclamação ou de interrogação.
Ao descartar esses ruídos, medidos empiricamente, demonstra-se que o método Heudet possibilitou um ganho de precisão de 1,2% nas EM identificadas. Ele identificou todas as EM encontradas pelo pacote NSP, mas descartou 83 casos considerados como ruídos. A figura 4 demonstra como as EM foram agrupadas para realizar a análise dos resultados.

Figura 4. **Processo de análise das EM obtidas.**



Fonte: Elaborada pelos autores.

Em trabalho publicado anteriormente por Silva e Souza (2012), cuja frequência de correlação utilizada foi igual a três, também ficaram demonstrados resultados similares com ganho de precisão de 1,5%. Portanto, isso possibilita verificar que com o número de co-ocorrências menor, implica em um número ainda maior de bigramas para esse mesmo *corpus*, o que resultou em um acréscimo do percentual de descarte.

6 CONCLUSÕES E TRABALHOS FUTUROS

Espera-se que esse estudo sirva de ponto de partida para que outros pesquisadores da área possam apresentar novas ideias a partir dos resultados dos experimentos que se utilizaram da técnica Heudet, a qual melhorou a precisão na identificação das EM. Com relação ao desempenho do *software* na identificação das EM, verificou-se que o tempo consumido pelo processo de identificação dos bigramas, processados em documento com cerca de 30 páginas em formato textos demorou cerca de 2 a 3 segundos.

Vistos esses números, conclui-se que o algoritmo Heudet apresenta vantagens em relação ao uso das técnicas estatísticas. Isso se dá pelo fato de ele levar em consideração a estrutura do documento. Afinal, ele considera o documento como um conjunto de sentenças, em vez de um conjunto de palavras, como é trabalhado pelos algoritmos estatísticos do pacote NSP. Em relação ao desempenho, verificou-se que há viabilidade de uso, porque os tempos consumidos durante os testes de processamento do *software* elaborado foram bem razoáveis.

Desse modo, foi demonstrado empiricamente que o processamento determinístico apresentou ganhos de precisão em relação a todas às treze técnicas estatísticas disponibilizadas pelo pacote NSP. Portanto, Heudet pode ser empregado com vantagens como parte de um Processamento de Linguagem Natural que demande a identificação de Expressões Multipalavras em um documento.

Como contribuição o *software* elaborado com a técnica Heudet pode ser disponibilizado através de uma API (Application Program Inteface) para grupos de pesquisa interessados em extrair *n*-gramas de um texto sem formatação, de tal modo que esse conjunto de pares de termos expresse automaticamente uma representação semântica do texto completo. Obviamente, deve-se relativizar essa representação do texto através de *n*-gramas, com as limitações representacionais que essa técnica impõe. Um conjunto de bigramas não consegue capitar todo significado global de um texto. Mas, de certo modo captura o seu sentido principal, ao considerarmos as palavras adjacentes que co-ocorrem com uma determinada frequência. Mesmo com essas limitações o resultado produzido pode ser utilizado para diversos fins, tais como: busca por semelhança, clusterização de documentos por temas similares e etc.

REFERÊNCIAS

CALZOLARI, Nicoletta et al. Towards best practice for multiword expressions in computational lexicons. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOUCES AND EVALUATION, 3rd, 2002, Las Palmas. **Proceedings...** Las Palmas, 2002, p. 1934–1940.

CHEN, Jisong, YEH; Chung-Hsing; CHAU, Rowena. A multi-word term extraction system. PRICAI 2006, LNAI 4099, p. 1160–1165, 2006. Springer-Verlag Berlin Heidelberg, 2006.

CINTRA, Anna Maria Marques. Elementos de linguística para estudos de indexação. **Ciências de Informação,** Brasília, v. 12, n. 1, p. 5-22, 1983.

CIPRO NETO, Pasquale; INFANTE, Ulisses. **Gramática da língua portuguesa**. São Paulo: Scipione, 2009. 584p.

DIAS, Gael; LOPES, José Gabriel Pereira; GUILLORÉ, Sylvie. Mutual expectation: a measure for multiword lexical unit extraction. In: VEXTAL. 1999, Veneza. **Proceedings ...** Veneza, 1999, p. 133-138.

FARACO, Carlos Emílio; MOURA, Francisco Marto, **Gramática**. São Paulo: Ática, 1990. 487p.

EVERT, Stefan; KRENN, Brigitte. Using small random samples for the manual evaluation of statistical association measures. **Computer Speech and Language**, London, v. 19, p. 450–466, Oct. 2005.

KURAMOTO, Hélio. Uma abordagem alternativa para o tratamento e a recuperação da informação textual: os sintagmas nominais. **Ciência da Informação**, Brasília, v. 25, n. 2, p. 182-196, mai/ago 1995.

LADEIRA, Ana. Paula. **Processamento de linguagem natural:** caracterização da produção científica dos pesquisadores brasileiros. 2010. 262 f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2010.

MAIA, Luiz Cláudio Gomes; SOUZA, Renato Rocha. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 15, p. 154-172, 2010.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. **An introduction to information retrieval**. Cambridge, 2009.

PEARCE, Darren. A comparative evaluation of collocation extraction techniques. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 3rd 2002, Las Palmas. **Proceedings...** Las Palmas, 2002, p. 1530–1536.

PECINA, Pavel. Lexical association measures and collocation extraction. In: INTERNATIONAL CONFERENCE LANGUAGE RESOUCES AND EVALUATION, 2010. **Proceedings**… 2010, 44(1-2), p. 137-158.

PEDERSEN, Ted et al. **The Ngram Statistics Package***.* 2011. Disponível em: http://www.d.umn.edu/~tpederse/nsp.html. Acesso em: ago. 2011.

PORTELA, Ricardo; MAMEDE Nuno; BAPTISTA, Jorge. Mutiword Identificação. In: SIMPÓSIO DE INFORMÁTICA, 3º, 2011. **Anais**... out. 2011, p. 1-12.

RAMISCH, Carlos. **Multiword terminology extraction for domain specific documents,** 2009*.* Dissertação (Mestrado em Mathématiques Appliqueées) - École Nationale Supérieure d'Informatiques, Grenoble, 2009.

RANCHHOD, Elisabete Marques. O lugar das expressões 'fixas' na gramática do Português. in Castro, I. and I. Duarte (eds.), Razão e Emoção, vol. II, Lisbon: INCM, p. 239-254, 2003.

RAYSON, Paul; PIAO, Scott; SHAROFF, Serge; EVERT, Stefan. MOIRÓN, Begoña Villada. Multiword expressions: hard going or plain sailing? **Springer Science Business Media B. V**. p. 1-5, 2009.

ROUSSINOV, Dmitri. Towards Combined Aspect Verification Model. (no prelo).

SAG, I. A. et al. Multiword expressions: a pain in the neck for nlp. In: Em Proceedings of the Third INTERNATIONAL COFERENCE ON COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING, 3rd, 2002, London. **Proceedings...** London: Springer-Verlag, 2002, v. 2276, p. 1–15.

SARMENTO, Luís. Simpósio **Doutoral Linguateca**. 2006. Disponível em: http://www.linguateca.pt/documentos/SimposioDoutoral2005.html. Acesso em: out. 2011.

SILVA, Edson Marchetti; SOUZA, Renato Rocha. Information Retrieval System using multiwords expressions (MWE) as descriptor. **JISTEM – Jounal of Information System and Technology Management**, v. 9, n. 2, p. 213-234, May/Aug. 2012.

SILVA, Joaquim Ferreira; LOPES, Gabriel Pereira. A local maxima method and fair dispersion normalization for extracting multi-word units from corpora. In: MEETING ON MATHEMATICS OF LANGUAGE, 6th , 1999. **Proceedings**... p. 369-381, 1999.

SOUZA, Renato Rocha. **Uma proposta de metodologia para a escolha automática de descritores utilizando sintagmas nominais**. Tese (Doutorado em Ciência da Informação) - Escola de Ciências da Informação, Universidade Federal de Minas Gerias, Belo Horizonte, 2005.

VILLAVICENCIO, Aline et al. Identificação de expressões multipalavra em domínios específicos. **Linguamática**, v. 2, n. 1, p. 15-33, abr. 2010.

WANG, Lijuan; LIU, Rong. A Rapid Method to Extract Multiword Expressions with Statistic Measures and Linguistic Rules. WISM 2011, Part II, LNCS 6988, p. 234–241, 2011.

YAGUNOVA, E. V.; PIVOVAROVA, L.M. The Nature of Collocations in the Russian Language. The Experience of Automatic Extraction and Classification of the Material of News Texts. **Automatic Documentation and Mathematical Linguistics**, Allerton Press, v. 44, n. 3, p. 164–175, 2010.

ZHANG, Wen; et al. Improving effectiveness of mutual information for substantival multiword expression extraction. **Expert Systems with Applications**, Elsevier, v. 36, 2009.