DOI: 10.5433/1981-8920.2017v22n1p293

REDES COMPLEXAS DE HOMÔNIMOS PARA ANÁLISE SEMÂNTICA TEXTUAL

REDES COMPLEJAS HOMÓNIMOS A ANÁLISISSEMÁNTICODEL TEXTO

Jadson da Silva Santos*
Felipe Coelho de Andrade**
Eduardo Manuel de Freitas Jorge***
João B. Rocha-Junior****
Hugo Saba*****

RESUMO:
Introdução: Estudos voltados ao processamento de linguagem natural já são bem difundidos e possuem aplicações diversas. Relacionado a essa área de pesquisa, o uso de técnicas para manipular um texto determinando a morfologia e sintaxe de suas palavras é bastante comum. Existem ferramentas que fazem esse tratamento, entretanto adicionar mecanismos de identificação semântica para essas palavras é fundamental para aumentar o entendimento automático da linguagem empregada. **Objetivo:** Com base nesse contexto, este artigo apresenta o processo de utilização de redes complexas como base de dados comparativa para determinar, através do contexto, o significado de palavras que expressam posicionamentos distintos. Além disso, são classificados com mesma morfologia e sintaxe, como ocorre com alguns homônimos. **Metodologia:** Através de uma metodologia experimental, o modelo aqui proposto baseia-se em pesquisa já consolidadas em Processamento de Linguagem Natural para montar uma rede complexa que recebe como vértices as palavras de um determinado texto e estabelece suas ligações a partir da ocorrência de adjacência entre esses termos. Assim observando as variações da rede, identifica-se como os

*Mestre em Computação Aplicada pela Universidade Estadual de Feira de Santana- (UEFS). E-mail: jadson.54nt05@gmail.com
**Bacharel em Análise de Sistemas pela Universidade do Estado da Bahia – (UNEB). E-mail: andrade.felipecoelho@gmail.com
***Doutor em Difusão do Conhecimento no Programa Multi Institucional pela Universidade Federal da Bahia- (UFBA). E-mail: emjorge1974@gmail.com.
****Doutor em Ciências da Computação pela Universidade Norueguesa de Ciência e Tecnologia e Docente do Programa de Pós-Graduação em Computação Aplicada (PGCA) da UEFS. E-mail: joao.rocha.jr@gmail.com.
*****Doutor em Difusão do Conhecimento no Programa Multi Institucional pela Universidade Federal da Bahia- (UFBA). E-mail: hugosaba@pq.cnpq.br.

homônimos do texto estão relacionados, e através da análise do contexto em que se encontram, verificar se é utilizado para expressar mais de um significado. **Resultados:** Um processo genérico com etapas de pré-processamento, montagem de Redes Complexas usando Processamento de Linguagem Natural para concepção de uma rede de homônimos para extrair informação textual semântica. **Conclusões:** A análise de homônimos selecionados e etiquetados é um processo não apenas morfossintático, acrescentado semântica em uma frase, parágrafo ou texto onde as palavras são empregadas. Assim, através de Processamento de Linguagem Natural acontecimentos mundiais e fatos filosóficos escritos textualmente podem ser melhor analisados, como por exemplo, o poder de argumentação e o perfil de escrita de um autor.

Palavras-chave: Processamento de Linguagem Natural. Redes Complexas. Processamento Textual. Semântica.

1 INTRODUÇÃO

Uma das formas de analisar um texto e definir seu grau de qualidade é por meio da observação de palavras usadas repetidamente. Define-se que um texto com uma variedade maior de palavras para expor o tema pretendido, costuma ser mais valorizado, indicando que o autor tem maior domínio de vocabulário e evita repetições desnecessárias.

Na língua portuguesa existem várias palavras que possuem mais de um significado, e são usualmente empregadas para diversos sentidos, inclusive num mesmo período. Os casos de repetição não afetam a qualidade do texto, desde que estejam bem elaboradas e sejam fatores essenciais ou insubstituíveis da mensagem. Palavras que são escritas de forma igual pertencem ao grupo das Homônimas classificadas como homônimas perfeitas, quando a escrita e a pronuncia são idênticas, e homônimas homógrafas, caso em que apenas as grafias são iguais, como definem Figueiredo e Figueiredo (2012). Baseado nessa premissa, para avaliar um texto é fundamental ter a compreensão de tais casos, a fim de atribuir um nível de qualidade mais preciso.

O Processamento de Linguagem Natural (PLN) é a área que estuda problemas de geração e compreensão automática de linguagens humanas naturais, objetivando possibilitar uma comunicação mais atrativa entre homem e máquina (OTHERO, 1978). Considerando as ambiguidades das informações

e da interpretação de dados baseados no seu contexto. Existem várias aplicações para está área de pesquisa, uma delas é a Web Semântica com o objetivo de melhorar a busca de informações na Web, porém com problemas associados a semântica que são suavizados de forma semiautomática através de análises léxicas e aplicação de ontologias (CHISHMAN, 2009).

Para analisar um texto através de PLN é preciso considerar aspectos morfológicos, sintáticos e semânticos, de modo que permite o tratamento do texto e o seu entendimento (JÚNIOR, 2010). A análise morfológica trata de verificar a estrutura da palavra e identificar o radical e afixo, ajudando a classificá-las quanto à classe gramatical que pertencem, e em casos de mecanismos de buscas que usam o seu radical. O processamento sintático estuda as regras que determinam a formação de frases segundo uma gramática também conhecida como Modelo de Linguagem, verificando a sequência mais provável de organizá-las.

No que se refere a análise morfológica e sintática de textos uma das técnicas utilizadas no processamento da linguagem natural é a etiquetagem (tagging), que identifica a classe gramatical de cada palavra de um texto. O MXPost (AIRES *et al*. 2000), baseado no modelo de Ratnaparkhi (RATNAPARKHI, 1996) é uma ferramenta que faz isso, varrendo um texto/corpus e determinando a classe gramatical para as palavras contidas. Por exemplo, a precisão avaliada pela ferramenta MXPost em cadernos do jornal *Folha de São Paulo* podem variaram de 96.98% a 94.39% (FINATTO, 2011).

Existem palavras que podem ser usadas para expressar mais de um significado e recebem a mesma classificação morfossintática do etiquetador, o qual não determina o conceito semântico utilizado. Num caso de avaliação do texto, por exemplo, só a etiquetagem não possibilita o tratamento dessas ocorrências. Considerando este processo e suas diversas aplicações, o mesmo pode não ser efetivo, uma vez que demonstra fragilidade na coleta de propriedades semânticas que auxiliam no processamento de informações (METZ*et al.* 2007). Uma possibilidade de solução é o uso das redes complexas para auxiliar no processo de análise morfológica e sintática de textos,

facilitando as situações de ambiguidades e palavras repetidas com sentidos distintos.

O conceito de redes complexas está associado a um grafo, com um conjunto de vértices ou nós interligados por arestas, com topografia não trivial (BARABÁSI, 2003). Caracterizando dessa maneira uma rede complexa como um conjunto de elementos que estabelecem conexões entre si (NEWMAN, 2003). O estudo de redes iniciou por volta de 1735 por meio do matemático Leonard Euler que realizou seu primeiro trabalho de redes na forma de grafos para solucionar o problema das pontes de Königsberg, dando origem a teoria dos grafos.

Dentre as diversas áreas de estudos as quais redes complexas podem ser aplicadas está a de PLN. Dentro dessa gama de aplicação já existem vários trabalhos de utilização de redes complexas, por exemplo, para representar as relações entre palavras num texto. Como no trabalho de Antiqueira et al. (2005), que criaram uma rede complexa na qual os vértices simbolizam as palavras e as arestas relacionam as mesmas por adjacência. E através dessa modelagem analisam a qualidade do texto baseado nesse relacionamento. Outro exemplo está no trabalho de Caldeira (2005) que cria uma rede complexa para representar uma rede de palavras retiradas de um texto, além ainda do conceito de associação de palavras do trabalho de Nelson et al. (1999). Modelar textos através de redes complexas não é um método novo, pelo contrário já existe e é difundido a um tempo considerável, no entanto a análise de homônimos que é sugerida é pouco ou quase não trabalhada.

Através de uma metodologia experimental, o modelo aqui proposto baseia-se nesses trabalhos apresentados para montar uma rede complexa que recebe como vértices as palavras de um determinado texto e estabelece suas ligações a partir da ocorrência de adjacência entre esses termos. Assim observando as variações da rede, identifica-se como os homônimos do texto estão relacionados, e através da análise do contexto em que se encontram, verificar se é utilizado para expressar mais de um significado. Caso em que a repetição dessa palavra não indica falta de vocabulário para evitar uso contínuo de um mesmo termo.

Outra característica peculiar dessa abordagem é a possibilidade de extrair informação semântica por meio do processamento dessa rede, uma vez que dado um termo homônimo, que encontra-se ligado com palavras que expressam ideias distintas, existe a possibilidade de determinar, em cada caso a que sentido está sendo usado tal termo homônimo. Para facilitar a compreensão do modelo proposto apresenta-se um conjunto de textos com homônimos e as suas redes, permitindo assim apoiar na resolução de questões semânticas.

A próxima seção descreve como é realizado o processamento do texto para montar a rede de palavras e averiguar a ocorrência dos homônimos. Na seção 3 é detalhada a parte do processamento que realiza a remoção de homônimos em textos com base em rede complexa. Por fim a seção 4 discorre sobre as observações acerca dos experimentos realizados e as conclusões feitas por meio dos resultados encontrados.

2 ANÁLISE TEXTUAL

Para analisar os textos é necessário fazer um pré-processamento a fim de preparar seus dados para serem utilizados. Para ampliar a semântica do texto o enquadramento do mesmo em um domínio é uma estratégia fundamental. Uma forma seria trabalhar com uma Ontologia de Fundamentação em um primeiro nível de abstração, para posteriormente conduzir ao enquadramento em uma Ontologia de Domínio (CAMPOS, CAMPOS, MEDEIROS, 2011). Após esta etapa, inicia-se o processamento textual onde são retiradas as *stopwords*, que são os termos com pouca representatividade, como artigos e preposições. Após o texto passa pelo processo de lematização, onde os termos são reduzidos a sua forma básica, Biber (1998), facilitando o agrupamento padrão de diversas variantes de um mesmo signo, Gallison (1983), afim de reduzir ambiguidades na fase de lematização o texto é etiquetado pelo MXPost. Dados estes passos os termos dos textos encontram-se estruturados para a construção da rede.

Para a rede ser montada são identificados os homônimos homógrafos e perfeitos contidos no texto, e localizadas as palavras adjacentes a esses

termos. Por meio desses procedimentos já é possível montar a rede de homônimos sugerida e avaliar suas características. Neste modelo as *n* palavras serão representadas por vértices e a relação de adjacência entre esses termos determinará as conexões. As arestas receberão como pesos o número de vezes que um determinado termo relaciona-se a outro, ou seja, a quantidade de associações entre as palavras. Por meio dessa representação é possível ter informação a respeito dos termos com maior força de associação e analisar a variância de conexões entre termos.

No aspecto de análise semântica a rede aqui proposta permite avaliar como os homônimos estão relacionados com o significado expresso no texto. No exemplo da figura 1 encontra-se um trecho do texto do “Poema às Notas” de Luísa Ducla Soares (SOARES, 2004) que demonstra um exemplo de uso de homônimos para expressar mais de um significado (Figura 1).

Figura 1:"Poema às Notas", de Luísa Ducla Soares**.**

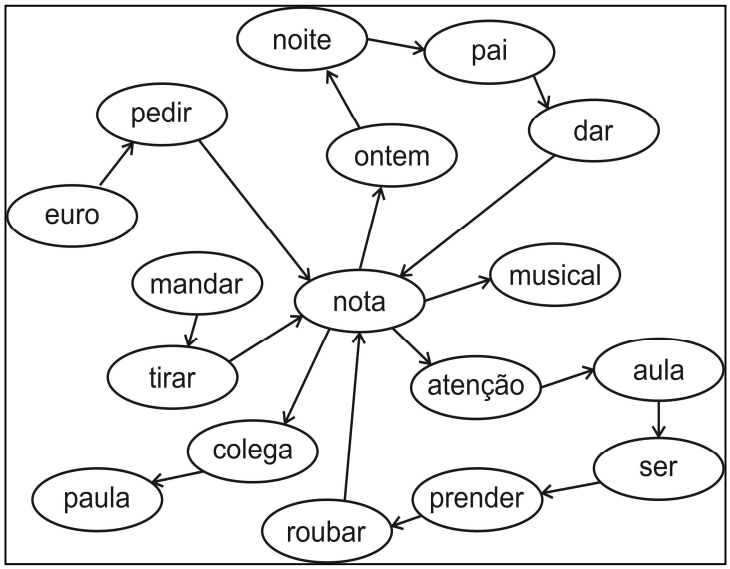
Sem um euro pedi notas
ontem à noite, aos meus pais.
Deram-me notas e notas,
mas só notas musicais.

Mandaram-me tirar notas
e ter atenção na aula.
Fui preso por roubar notas
da minha colega Paula.

Fonte: Soares, 2015.

Dado o trecho do poema a rede de associação de palavras descreve o relacionamento entre os termos e demonstra a ocorrência de associações do homônimo “nota” para indicar nota musical, nota da ficha de avaliação, nota em Euro e apontamento. Na figura 2 encontra-se a rede gerada a partir desse texto e percebe-se que o termo “nota” tem o maior grau de entrada e saída, justamente por ser o termo homônimo utilizado no trecho do poema. Cada vértice relaciona-se uma vez com os demais, porém em uma representação de um texto maior as ligações entre mesmos nós tendem a ocorrer com maior frequência.

Figura 2: Rede do trecho do poema da Figura 1.



Fonte: Autores, 2015.

3 REMOÇÃO DE HOMÔNIMOS EM TEXTOS COM BASE EM REDE COMPLEXA

Por meio dessa abordagem é possível utilizar os artifícios da teoria dos grafos para analisar a rede, além da ocorrência destas palavras com mais de um sentido semântico a fim de identificar, por meio de suas associações, seus significados.

Na figura 3 encontra-se um texto que demonstra a utilização de homônimos com significados distintos e o processo de identificação e classificação destes termos baseados nas redes de homônimos criadas.

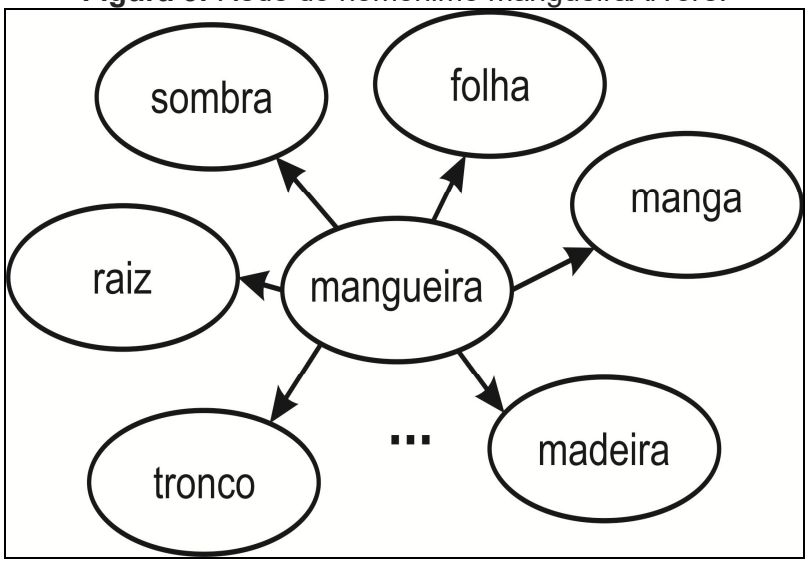
Figura 3

Ao sentar-se abaixo de uma mangueira de tronco largo sentiu o cheiro da manga madura no pé e foi surpreendido com uma manga pobre caindo logo em sua roupa nova sujando a manga de sua camisa. Saiu triste de perto da mangueira que o fazia sombra e foi em busca de água para se lavar, quando avistou ao lado da casa uma mangueira encaixada na torneira, lavou a manga de sua vestimenta e retornou a sombra da mangueira.

Fonte: Autores, 2015.

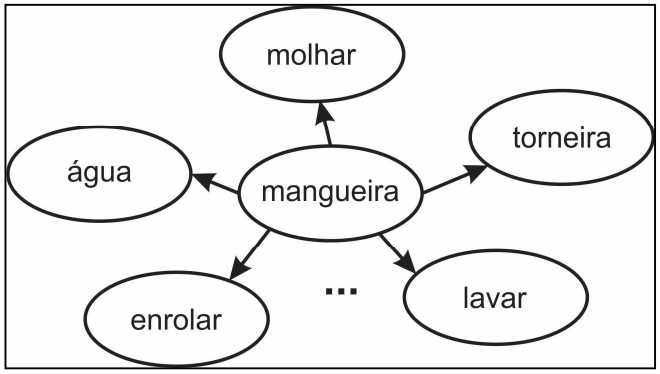
O processamento desse texto possibilita a identificação e a classificação dos termos homônimos segundo as redes criadas para este fim. É verificado no texto acima a ocorrência dos termos “mangueira”e “manga” que são utilizados em mais de um sentido semântico. Cada uma dessas palavras está associada a um conjunto de termos que ajudam a definir seu significado. Foram criadas redes de homônimos para cada tipo de utilização dessas palavras (Figuras 4 e 5), como é demonstrado seguir, na qual há a associação com os termos que usualmente as acompanham.

Figura 3: Rede do homônimo mangueiraArvore.



Fonte: Autores, 2015.

Figura 4: Rede do homônimo mangueiraObjeto.



Fonte: Autores, 2015.

As redes de associações criadas para os homônimos permitem identificar o sentido semântico empregado para o termo e extrair a informação semântica que a palavra indica no texto (Figura 3). O termo “mangueira” foi submetido a análise, nas redes de homônimos para esta palavra e por meio da relação de adjacência do homônimo no texto, *“mangueira – torneira”* pode-se classificar que este termo significa “mangueiraObjeto”. Da mesma forma que através da associação*“mangueira – tronco”* encontradano texto possibilita definir que o termo significa “mangueiraArvore”. O procedimento para identificação e classificação do termo “manga” segue o mesmo padrão identificando o uso do termo relacionado a fruta e a roupa. Esta abordagem permite tratar casos de ambiguidades em textos que utilizem termos homônimos e substituir a palavra adicionando a referência semântica. Após o processamento de identificação dos homônimos no texto (Figura 3), os mesmos são substituídos por um termo específico (Figura 6).

Figura 5: Texto com homônimos identificados.

Ao sentar-se abaixo de uma mangueiraArvore de tronco largo sentiu o cheiro da mangaFruta madura no pé e foi surpreendido com uma mangaFruta pobre caindo logo em sua roupa nova sujando a mangaRoupa de sua camisa. Saiu triste de perto da mangueiraArvore que o fazia sombra e foi em busca de água para se lavar, quando avistou ao lado da casa uma mangueiraObjeto encaixada na torneira, lavou a mangaRoupa de sua vestimenta e retornou a sombra da mangueiraArvore.

Fonte: Autores, 2015.

A substituição dos homônimos contidos no texto representa a classificação semântica identificada para a mesma com base no contexto. O processo representa uma tarefa de PLN importante e permite utilizar o resultado dessa abordagem para auxiliar diversos procedimentos de processamento textual. Assim, o modelo aqui proposto se baseia em Redes Complexas para agregar um maior grau de eficiência para aplicações PLN que demandam verificar um texto semanticamente, extrapolando as análises somente morfossintáticas. Logo, o modelo, associado com outras técnicas e

ferramentas da PLN, pode efetuar análise da qualidade de textos e correlacioná-los a outros textos de forma automatizada.

4 CONCLUSÃO

O processo de usar redes de palavras para permitir a atribuiçãodo significado semântico para homônimos selecionados e etiquetados igualmente representa uma técnica com caráter de eficácia satisfatório, pelo fato de tratar o processamento de linguagem natural baseado na identificação semântica, através da análise de contexto em que as palavras são usadas. Tal procedimento possibilita utilizar esse conceito de diversas formas, considerando a vasta área de atuação do PLN. De modo que ao verificar um texto, as repetições serão consideradas não apenas morfossintaticamente, mas também quanto a semântica, logo processadas com um maior grau de eficiência.

A análise de homônimos selecionados e etiquetados igualmente é um processo de entendimento, não apenas morfossintático, é acrescentado a semântica buscando analisar e julgar o sentido e o contexto que as palavras são empregadas em uma frase, parágrafo ou texto.

Considerando a vasta área de atuação do Processamento de Linguagem Natural (PLN), e a junção dos métodos morfossintáticos com o semântico descrito neste artigo é construída uma abordagem nova, aumentando o grau de eficiência e influenciando de maneira positiva a distinção de textos bons ou ruins, melhoria da compactação, e modelagem de textos.

Com o aperfeiçoamento do método e a junção de outros processos será possível construir uma aplicação de análise de textos que julgue a qualidade, o sentido que as palavras são empregadas de maneira a relacionar com os acontecimentos mundiais e fatos filosóficos que determinaram não só se um texto é bom ou ruim, mas o quão engajado, e o poder de argumentação, traçando perfis de escrita para cada autor.

REFERÊNCIAS

Aires, R. V. X., Aluísio, S. M., Kuhn, D. C. S., Andreeta, M. L. B. and Oliveira Jr., O. N. (2000). **Combining Multiple Classifiers to Improve Part of Speech Tagging**: A Case Study for Brazilian Portuguese, In: Proceedings of the 15th Brazilian Symposium on Artificial Intelligence (SBIA'2000), Atibaia, SP.

ANTIQUEIRA, L ; NUNES, M G V ; OLIVEIRA JR, O N ; COSTA , L. da. F. Complex networks in the assessment of quality text. **In Physics**, 2005.

BARABÁSI, A. L. **Linked**: How everything is connected to everything else and what it means for business, science and everyday life. New York: Plume, 2003.

BIBER, D. et all. **Corpus Linguistics**: Investigating Language Structure and Use. Cambridge: Cambridge University Press, 1998.

CALDEIRA, S. M. G. **Caracterização da rede de signos linguísticos**: um modelo baseado no aparelho psíquico de Freud. 2005. Dissertação (Mestrado). Fundação Visconde de Cairu, Salvador, 2005.

CAMPOS, M. L. de A.; CAMPOS, L. M.; MEDEIROS, J. da S. A Representação de Domínios de Conhecimento e uma Teoria de Representação: a ontologia de fundamentação. **Informação & Informação**, [S.l.], v. 16, n. 2, p. 140-164, dez. 2011. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/10389. Acesso em: 11 Nov. 2015.

CHISHMAN, R. L. de O. Integrandoléxicossemânticos e ontologias: uma aproximação a favor da Web Semântica. **Informação&Informação**, [S.l.], v. 14, n. 1esp, p. 103-124, dez. 2009. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/2159. Acesso em: 11 Nov. 2015.

FIGUEIREDO, A.; FIGUEIREDO, F. **Gramática Comentada com Interpretação de Textos**. 2. ed. São Paulo: Elsevier, 2012. 515p.

FINATTO, M. J. B. **Sobre a Eficácia do MXPOST Etiquetador morfossintático para o português do Brasil**. Disponível em: http://www.ufrgs.br/textecc/porlexbras/porpopular/arquivos/Sobre.pdf. 2011.

GALLISON, R. **Dicionário de Didáctica das Línguas**. Coimbra: Livraria Almedina, 1983.

JÚNIOR, J. M. C. **Sobre o Conceito de Processamento de Linguagem Natural.** Universidade Estadual de Campinas, Faculdade de Tecnologia. Limeira, São Paulo, 2010.

METZ, J. ; CALVO, R. ; SENO, E. R. M. ; ROMERO, R. A. F. ; LIANG, Z. . Redes Complexas: conceitos e aplicações. Série de relatórios técnicos do ICMC - USP nº 290, 2007 (Relatório Técnico).

NELSON, D. L., MCEVOY, C. L., e SCHREIBER, T. A.The University of South Florida word association rhyme, and word fragment norms. Behav. Res. Methods Instrum. Comput, 36, 402–407, 2004. Disponível em: https://www.ncbi.nlm.nih.gov/pubmed/15641430 . Acesso em 2014.

Newman, M.E.J., **The Structure and Function of Complex Networks**. SIAM Review, 2003. 45(2): p. 167

OTHERO, G. de Á. **Linguística Computacional Teoria & Prática.** São Paulo: Parábola Editorial, 2005.

RATNAPARKHI, A. A Maximum Entropy Part-Of-Speech Tagger. In: the Proceedings of the Empirical Methods in Natural Language Processing Conference. Pennsylvania: University of Pennsylvania, 1996.

SOARES, L. D. **Abecedário Maluco.** Porto: Civilização Editora, 2004.

Title

Homonyms's Complex Networks to Semantic Analysis Textual

Abstract:
Introduction: Study centres in natural language processing already spread and the study have several applications. Relate with this research area, it is common the use technic for manipulation a text. These technic is be able to determine the word morphology and the word syntax. There are tools that do this work, however adding engines for semantic identification of the words is essential for increase the automatic understanding the used language. **Objective:** On the basis of that, This paper present the process of using complex networks as a comparative database to determine by context the meaning of words that express different positions. Moreover, they are classified as same morphology and syntax , as with some homonyms. **Methodology:** Through of a experimental methodology, the model proposed it is based in consolidate researches in Natural Language Processing for building a Complex Network that receives as vertices the words of a certain text and establishes its connections from the occurrence of adjacency between these terms. Therefore, observing the variations of network, it is identified how to textual namesakes are related and through an context analyzed how if be there, check whether it is used to express more than one meaning. **Results:** A generic process with stages of preprocessing, building of a Complex Network used to Natural Language Processing for the building of a network homonyms to extract semantic information textual. **Conclusions:** The analyze of homonyms selected and labeled is the process not only morphosyntatic, adding semantic in the phrase, paragraph or text where the words are applied. However, with Natural Language Processing an events and philosophical facts can be better analyzed through of a world written textually, for example, the power of argument and the writing of an author profile.

Keywords: Natural Language Processing. Complex Networks. Textual Processing. Semantic.

Titulo

Redes complejas homónimos a Análisis Semántico del Texto

Resumen:
Introducción: Estudios dirigidos al procesamiento del lenguaje natural están bien difundidos y tienen varias aplicaciones. En relación con esta línea de investigación, es muy común el empleo de tecnicas para manipular un texto haciendo determinación de la morfología y de la sintaxis en sus palabras. Existen herramientas que hacen el trato de los textos, sin embargo añadir mecanismos de indentificación semántica para las palabras es fundamental para el aumento de la autocomprensión del lenguaje utilizado. **Objetivo:** Basados en el contexto, el presente artículo presenta el proceso de utilización de las redes complejas como la base de datos comparativos para la determinación, a través del contexto, del significado de las palabras que expresan diferentes posicionamientos. Además son clasificados con la misma morfología y sintaxis, como ocurre con algunos homónimos. **Metodología:** A través de una metodología experimental, el modelo propuesto en el presente artículo se basa en investigaciones reconocidas acerca del Procesamiento del Lenguaje Natural para organizar uma red compleja que recibe como los vétices las palabras de un texto en particular y establece sus conecciones a partir de la ocurrencia de la adyacencia entre estos términos. De este modo observando las variaciones de la red, se reconoce como los homónimos de los texto están relacionados, y a través del análisis del contexto en que están, comprobrar si es utilizado para expresar más de uno significado. **Resultados:** Uno proceso genérico con etapas del preprocesamiento, montaje de las Redes Complejas utilizando el Procesamiento del Lenguaje Natural para la concepción de una red de homónimos para extracciones de las informaciones semántica textuales. **Conclusiones:** El análisis de los homónimos seleccionados y rotulados es un proceso que no es solo morfosintáctico, añada la semántica en una frase, párrafo o texto dónde las palabras son utilizadas. Por lo tanto, por el Procesamiento del Lenguaje Natural los acontecimientos mundiales y hechos filosóficos escritos por extenso puedem ser mejores analizados, por ejemplo, el poder de la argumentación y el perfil de escritura del autor.

Palabras clave: Procesamiento del lenguaje natural. Redes Complejas. Pruebas de procesamiento. Semánticos.

Recebido: 12.02.2016
Aceito: 25.03.2017