5
Ambigüidades no contexto de Mineração de Textos

"Nem tudo que se enfrenta pode ser modificado, mas nada pode ser modificado até que seja enfrentado."

Albert Einstein

Mineração de textos é o processo de se extrair, dirigido pelos dados, conhecimento não previamente conhecido, a partir de fontes textuais. Para obtenção desse conhecimento é necessário mitigar os enganos introduzidos por construções textuais ambíguas. Desambiguação de sentido de palavra é o processo de atribuição de um significado a uma palavra baseado no contexto em que ela ocorre. Considerando a mineração de texto como um suporte à tomada de decisões é importante que a existência de ambigüidades seja minimizada, ou até mesmo eliminada, se for possível.

5.1
Tipos de ambigüidades

Todo enunciado que possa ser interpretado de diferentes maneiras pode ser considerado ambíguo. A ambigüidade, dentro do contexto lingüístico, pode ser estrutural ou polissêmica. O primeiro caso se prende a problemas de construção da sentença; o segundo deve-se à possibilidade de os vocábulos apresentarem mais de um significado. No caso da ambigüidade estrutural, também denominada gramatical, as causas são muitas e as possibilidades de eliminá-la variam conforme o problema que a origina. Dentre variados motivos pode-se destacar:

a) pelo posicionamento de certos complementos ou adjuntos, como:

Pedi o livro de João.

Nesse caso, não é possível afirmar sobre a propriedade do livro, o que poderia ser esclarecido com as seguintes construções:

Pedi o meu livro a João Carrilho.

Pedi que João me emprestasse o seu livro.

Pedi o livro cujo autor é João.

b) devido à posição do adjunto adverbial, conforme:

Crianças que comem doces freqüentemente têm cáries.

Nesse exemplo, a ambigüidade está associada ao uso do termo "freqüentemente", que pode estar associado ao ato de comer doces ou a ter cáries. No primeiro caso a sentença poderia ser mais bem escrita da seguinte maneira:

Crianças que, com freqüência comem doces, têm cáries.

Já o segundo sentido poderia ser mais bem expresso pela sentença:

É muito freqüente o aparecimento de cáries em crianças que comem doces.

c) em orações adjetivas, como:

Procuro a chave do cofre que estava em meu quarto.

Pelo exemplo apresentado não é possível afirmar categoricamente o que estava em meu quarto, o cofre ou a chave. As duas sentenças a seguir são mais claras na representação das duas idéias possíveis.

Procuro em meu quarto a chave do cofre.

Procuro a chave daquele cofre que fica em meu quarto.

d) Nas orações reduzidas:

Vi o prefeito passeando no centro da cidade.

Quem exatamente estava passeando, eu ou o prefeito? As diferentes idéias podem ser expressas pelas seguintes sentenças:

Enquanto eu passeava no centro da cidade, vi o prefeito.

Vi o prefeito quando ele passeava pelo centro da cidade.

e) Devido ao uso de pronomes:

O advogado disse ao réu que o juiz acreditaria em suas palavras.

Afinal o juiz acreditaria nas palavras de quem, do réu ou do advogado? O pronome "suas" não possui uma referência clara a quem está associado.

Já a ambigüidade derivada da polissemia do vocábulo, também conhecida como ambigüidade lexical, ou ainda como ambigüidade do sentido da palavra pode ser evitada, em grande parte das vezes pelo esclarecimento extraído do contexto, ou pela substituição do vocábulo polissêmico por outro de sentido equivalente. Como exemplo tem-se:

O cadáver foi encontrado próximo ao banco.

Nesse caso, o conhecimento do contexto pode esclarecer a dúvida relativa ao tipo de banco ao qual a sentença se refere, que poderia ter o sentido de instituição financeira ou banco com o sentido de assento. No caso específico do exemplo esse contexto não é revelado dentro da frase e por isso deveria ser obtido com o restante do documento.

O presente trabalho se preocupa com os casos de ambigüidade polissêmica e propõe uma abordagem para melhoria de precisão na recuperação da informação baseada no tratamento dessas ambigüidades.

5.2
Desambiguação de sentido de palavra

Desambiguação de sentido de palavra é talvez a mais crítica tarefa na área de lingüística computacional e, tradicionalmente, considerada um problema de difícil solução (*AI-hard problem*) no ramo da inteligência artificial (NAVIGLI, 2005). Uma inovação nesse campo pode ter impacto significativo sobre muitas aplicações relevantes, tais como recuperação de informações (na Internet ou locais), aumento de acesso a *Web services*, extração de informação e etc. Para um exame dos trabalhos na área consulte (IDE, 1998). Muitas abordagens para atacar o problema de desambiguação têm se valido de conhecimento semântico obtido por codificação manual ou extraído automaticamente de recursos léxicos. Dentre esses recursos, disponíveis na *Web* destaca-se principalmente a **WordNet**[24], discutida em (FELLBAUM, 1998). É importante destacar a dificuldade de um computador em explorar os recursos de conhecimento léxico de um dicionário convencional para solucionar ambigüidades, visto que a informação contida em tais dicionários está organizada apenas por verbetes, o que gera uma grande demanda de trabalho manual para aprendizado semântico. (SOARES, 2008) utiliza métodos de varredura na *Web* e *parsing* das informações obtidas para montagem de um dicionário cuidadosamente estruturado na forma de uma base de

[24] A base de dados e programas para sua utilização podem ser obtidos através do sítio: http://www.cogsci.princeton.edu/obtain.shtml .

dados. A WordNet é também uma base de dados em inglês, que agrupa substantivos, verbos, adjetivos e advérbios em conjunto de sinônimos cognitivos, os *synsets*, expressando um conceito distinto. Os *synsets* são interligados por significados conceituais semânticos e por relações léxicas, resultando numa rede de conceitos e palavras relacionadas. De forma mais específica, pode-se dizer que o termo atômico usado pela WordNet, o *synset*, é uma estrutura que representa um significado específico para o termo, que inclui a palavra, uma breve explanação e seus sinônimos, que representam as conexões semânticas entre os termos. O significado específico de uma palavra sob uma mesma classe de gramática é chamado de sentido. Cada sentido de uma palavra está em um diferente *synset*. Da mesma forma pode-se ter muitas palavras relacionadas a um mesmo *synset* como por exemplo, *night*, *nighttime* e *dark* com o seguinte sentido: *the time after sunset and before sunrise while it is dark outside*. *Synsets* se interconectam através de relações semânticas explícitas. Algumas dessas relações são específicas para uma classe gramatical como: hipernímia e hiponímia, para substantivos; hipernímia e troponímia para verbos; holonímia e meronímia para substantivos. Porém a WordNet não fornece nenhuma informação sobre diferentes significados dado o contexto sob o qual a palavra aparece (FELLBAUM, 1998).

5.3
Pesquisas na área de desambiguação

Muitos trabalhos têm tirado proveito das valiosíssimas contribuições dadas pela WordNet, no âmbito de desambiguação de sentido de palavras. Em (DIOU, 2006), o problema de desambiguação de sentido de palavra é formulado como um problema de associações imprecisas entre palavras e sentido de palavras dentro de um contexto textual. A abordagem utilizada possui duas partes principais. Inicialmente, considera-se que para cada sentido, um conjunto nebuloso provê o grau de associação entre palavras e sentidos. Um algoritmo classifica os sentidos de uma palavra em um texto baseado na informação dos conjuntos nebulosos, conduzindo efetivamente para a desambiguação de sentido da palavra. Na segunda parte, é desenvolvido um método baseado na *WordNet*, que constrói os conjuntos nebulosos para os sentidos (independente de um documento textual). Os

resultados experimentais são satisfatórios e mostram que a modelagem de desambiguação de sentido de palavras como um problema de associações imprecisas é promissor, e dessa forma, pode ser tratado através de relações *fuzzy*.

Em (ROSSO, 2003) o problema de desambiguação de sentido de palavra é caracterizado como o processo de atribuição de um significado a uma palavra baseado no contexto em que ela ocorre. Ele considera a ausência do rótulo do sentido em dados de treinamento como um problema real para a tarefa de desambiguação. É apresentado um método para a resolução de ambigüidade léxica que recai sobre o uso da taxonomia de nomes, completamente coberta pela *WordNet* e a noção da distância conceitual entre conceitos, capturada por fórmula de densidade conceitual desenvolvida para esse propósito. Esse método integralmente automático não requer codificação manual de entradas lexicais, rotulação manual de texto nem nenhuma espécie de processo de treinamento. Os resultados experimentais foram automaticamente avaliados sobre o *SemCor*, o *Brown Corpus* em sua versão com rótulos para os sentidos das palavras e se mostraram bastante consistentes para o tratamento do problema.

Outros trabalhos têm se focado nos aspectos ambíguos atribuídos a problemas de tradução. Em (OLIVEIRA, 2005) é apresentado um método estatístico de desambiguação de sentido de palavra com aplicação em um sistema de tradução automática português-chinês. Devido à limitada disponibilidade de recursos português-chinês na forma de *corpora* digital ou rotulado com respeito à estrutura sintática, um aprendizado supervisionado e um *corpus* bilíngüe não-alinhado é utilizado. O método proposto primeiro identifica palavras relacionadas a cada uma das palavras-base ambíguas em seu universo de palavras e distância relativa. Um modelo é então aplicado na identificação de muitos sentidos adequados de uma palavra ambígua em termos de suas palavras relacionadas. Todos os sentidos descobertos são convertidos em um conjunto de regras e armazenados em uma base de conhecimento para mais tarde ser usado nos processos de tradução e desambiguação. Resultados experimentais mostram um aumento de 6% na atribuição correta da tradução correspondente em comparação com o método base.

Já outros pesquisadores da área preferem tratar o problema com independência da língua. Consideram a diversidade de fontes de informação e o

explosivo crescimento da Internet evidências irrefutáveis de uma necessidade para recuperação de informação que possa ir além dos limites da linguagem. Ambigüidades por falha na tradução das consultas é uma das maiores causas para grande perda de efetividade de desempenho mono lingual, para métodos baseados em dicionário para recuperação de informação em linguagem cruzada. Em (SADAT, 2002) o enfoque em cima de tradução de consultas e desambiguação, para aumentar a efetividade da recuperação de uma informação e para reduzir erros comuns nas abordagens normalmente utilizadas. Um método de desambiguação estatística combinada antes e depois da tradução é proposto para evitar problema de seleção errada da tradução alvo. A efetividade do método de desambiguação proposto é testada para uma aplicação de recuperação de informação Francês-Inglês e comprovamos a efetividade do método de desambiguação proposto.

Alguns pesquisadores mantêm a abordagem de utilizar redes semânticas, assim como a WordNet, porém preferem criar suas próprias estruturas para tratamento do problema. Em (NAVIGLI, 2005) é apresentado um método chamado de *structural semantic interconnections* (SSI), que cria especificações estruturais de possíveis sentidos para cada palavra no contexto e seleciona a melhor hipótese de acordo com uma gramática, descrevendo relações entre especificações de sentido. Especificações de sentido são criadas de diversos recursos léxicos disponíveis, integrados parte manualmente, parte com auxílio de procedimentos automáticos. O algoritmo SSI é aplicado a diferentes problemas de desambiguação semântica, como população automática de ontologias, desambiguação de sentenças em textos genéricos e desambiguação de palavras em definições de glossários. Experimentos de avaliação são executados sobre domínios específicos de conhecimento (turismo, redes de computadores, etc.), bem como sobre conjunto de testes padrão de desambiguação.

Algumas pesquisas consideram que as abordagens para desambiguação de sentido de palavras, baseadas em técnicas de representação do conhecimento, devem ser substituídas por aprendizado de máquina robusto e técnicas estatísticas. Alguns resultados de avaliações comparativas entre modelos baseados em aprendizado de máquina e representações do conhecimento tentam mostrar que esses últimos possuem limitações inerentes. Em outras palavras, o aumento da

disponibilidade, em larga escala, de recursos de conhecimento léxico introduzem um novo desafio para as abordagens baseadas em conhecimento. (CHAN, 1998) defende que o entendimento de linguagem natural envolve consideração simultânea de um grande número de diferentes fontes de informação e diz que métodos tradicionais empregados em análise de linguagens têm focado sobre o desenvolvimento de formalismos poderosos para representar estruturas sintáticas ou semânticas com regras para transformação de linguagens dentro desses formalismos. (CHAN, 2005) considera, contudo, esses métodos fazem uso de um pequeno subconjunto do conhecimento somente e por isso descreve como o usar a escala inteira da informação através de uma arquitetura neurosimbólica que é uma hibridização de uma rede simbólica e vetores de sub-símbolos gerados de uma rede conexionista. Partindo de uma rede simbólica com conhecimento prévio, os vetores de sub-símbolos são usados para realçar as capacidade de desambiguação do sistema e prover flexibilidade no entendimento da sentença. O modelo captura uma diversidade de informações incluindo associações de palavras, restrições sintáticas, regras semânticas e contexto e alcança processamento altamente interativo por representação do conhecimento em uma rede associativa em que inferências semânticas reais são executadas. Um uso integrado das sentenças previamente analisadas no entendimento é outra importante característica do modelo. O modelo seleciona dinamicamente uma hipótese entre muitas. Essa idéia é apoiada por três simulações que mostram o grau de desambiguação sobre uma quantidade de regras lingüísticas e informação semântico-associativa disponível para apoiar o processo de inferência no entendimento de linguagem natural. Diferentemente de sistemas similares, esse sistema híbrido parece mais sofisticado no tratamento do problema de desambiguação de linguagem pelo uso de indícios de fontes díspares bem como efeito de modelagem de contexto em análise de sentenças. (CHAN, 2005) defende que isso é potencialmente mais poderoso que um sistema que confia em um paradigma de pré-processamento, como é o caso do framework apresentado em (ARANHA, 2007).

Outros trabalhos preferem atacar apenas um subconjunto do problema considerando que a redução do escopo facilitaria o trabalho de métodos de aprendizado de máquina. Como exemplo dessa abordagem pode ser citado (SONG, 2005), que trabalha sobre as abreviações e propõe uma abordagem para

desambiguação que utiliza representação semântica de símbolos e termos numéricos adicionados às palavras em documentos clínicos. Enquanto a maioria dos trabalhos relacionados trata símbolos e palavras numéricas como *stopwords*. (SONG, 2005) tenta mostrar que todas as palavras têm um papel importante especialmente em documentos "crus", ou seja, sem tratamento, como documentos clínicos, que contém jargões, símbolos e abreviações escritas por médicos. Para a tarefa de desambiguação de abreviações utiliza um classificador e compara diversas variações da abordagem empregada com o método tradicional de **saco de palavras**[25]. Os resultados mostram que o sistema, usando abreviação semântica de símbolos e termos numéricos, pode aumentar a acurácia consideravelmente com a utilização de um classificador SVM (*Support Vector Machine*).

Pelos inúmeros trabalhos na área e pela diversidade e, em alguns casos, controvérsia nos resultados, é plausível dizer que não exista ainda uma linha bem definida de abordagem para tratamento de ambigüidades e dessa forma se tem hoje um grande campo aberto para novas pesquisas.

5.4
Modelo geral para tratamento de ambigüidades

Entender a mente humana é uma tarefa bastante complexa e por meio dos computadores tem sido possível modelar e simular algumas das habilidades e comportamentos cognitivos. A linguagem, em particular, tem sido considerada um dos principais aspectos do comportamento humano e talvez a mais desafiadora manifestação da complexidade da mente humana. Uma linguagem é o veículo através do qual o ser humano expressa idéias e modelos de processos cognitivos e precisa compreender um léxico que funcione como um universo dentro do qual um conceito possa ser representado, no caso, por um termo. Porém, algumas vezes, um discurso pode não representar, de maneira clara, os conceitos que foram pensados pela mente de quem o originou. Nesse momento percebe-se que uma o entendimento de um discurso precisa consistir não somente do conhecimento simples das palavras que o compõe, mas também do contexto sob o qual elas ocorrem.

[25] Do inglês, *bag of words*.

De forma geral as técnicas para detecção e tratamento de ambigüidades costumam partir do mesmo princípio para solucionar o problema, obter informações dentro do contexto onde se localiza a ambigüidade. A diferença está na forma como essas informações de contexto serão buscadas. Pode-se propor um modelo geral para tratamento conforme a figura 18. Esse modelo traduz a maneira mais geral de tratar o problema e não se diferencia em nada da forma na qual o ser humano utiliza quando lê um texto e se depara com situações ambíguas.

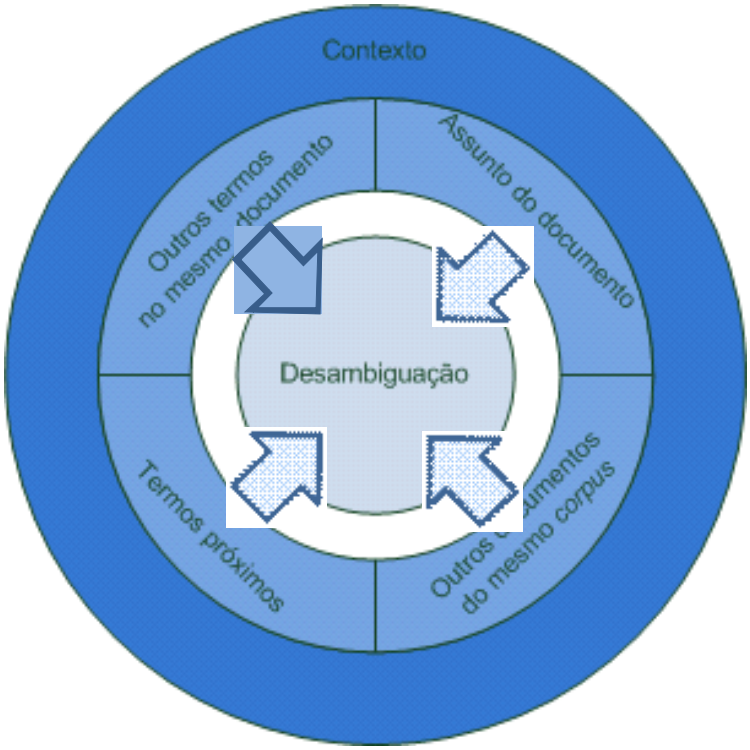

Figura 18 – Modelo geral para tratamento de ambigüidades

Quando o ser humano lê um trecho e fica com dúvidas em relação ao entendimento, devido à ambigüidade lexical, ele busca informações ao redor que possam elucidar suas dúvidas. Quando se vai ao redor, ou seja, no contexto, pode-se estender seu escopo tanto quanto seja necessário para a captura do entendimento desejado. É costume partir de um contexto mais próximo e ir expandindo-o. Na grande maioria das vezes a ambigüidade pode ser resolvida com termos próximos, porém como será mostrado no projeto proposto (veja capítulo 6), ambigüidades podem ser detectadas com informações de um contexto mais ampliado, chegando a ser resolvida pelo assunto ao qual o texto trata, que

por sua vez é obtido por intermédio das palavras mais freqüentemente encontradas nele. É importante destacar que não existe a exigência de percorrer um caminho fixo ou ordenação do contexto mais próximo para o mais distante. Na realidade qualquer informação adicional dentro do texto ou sobre o texto pode acabar auxiliando a resolução de uma ambigüidade.