Introdução

"Nenhum homem realmente produtivo pensa como se estivesse escrevendo uma dissertação."

Albert Einstein

Mineração de textos é o processo de se extrair, dirigido pelos dados, conhecimento não previamente explícito, a partir de fontes textuais (correio, imprensa, transações, *websites*, *newsgroups*, fóruns, listas de correspondência, além dos mais diversos documentos no formato texto) úteis para tomar decisões cruciais de negócio (ZANASI, 2007). Minerar textos pode ser uma tarefa bastante complexa dependendo dos objetivos a serem alcançados e também do escopo com o qual se queira trabalhar. Processos automatizados por computadores necessitam de modelos capazes de representar o objeto com o qual estão tratando. No caso das linguagens naturais essa modelagem ganha contornos ainda mais complexos, pois esses modelos devem ser capazes de capturar, de certa maneira, aspectos semânticos, presentes em todas as linguagens naturais. Uma linguagem pode ser formalmente definida como sendo um conjunto de *strings* de símbolos de algum alfabeto. Uma s*tring* é uma seqüência finita de símbolos justapostos (HOPCROFT, 1969). No caso do português o alfabeto é o ocidental, composto de 26 letras e as *strings* são as palavras, ou seja, a concatenação dos símbolos do alfabeto. Muitos são os problemas envolvidos quando se lida com linguagens naturais e talvez o mais complexo seja o tratamento das ambigüidades. Com respeito ao substancial trabalho realizado ao longo do tempo não existe nenhum algoritmo que possa desambiguar completamente um texto (SHOLOM, 2005). No ramo da inteligência artificial a desambiguação de sentido de palavras (WSD – *Word Sense Desambiguation*) é tradicionalmente considerada um problema de difícil solução (*AI-hard problem*) e uma inovação nesse campo pode ter impacto significativo sobre muitas aplicações relevantes baseadas na *Web*, tais como recuperação de informações na *Web*, aumento de acesso a *Web services*, extração de informação e etc (NAVIGLI, 2005). De forma específica, pode-se dizer que desambiguação de sentido de palavra é o processo de atribuição de um significado a uma palavra baseado no contexto em que ela ocorre (ROSSO, 2003) ou ainda

PUC-Rio - Certificação Digital Nº 0621313/CA

um mecanismo baseado em lingüística para definir automaticamente o sentido correto de uma palavra dentro de um contexto (LEDO-MEZQUITA, 2006).

De maneira mais geral, ambigüidade se refere, normalmente, à propriedade, inerente às sentenças, de poderem ser interpretadas de mais de uma maneira e de que os indícios disponíveis sejam insuficientes para o entendimento ou otimização da interpretação (KOOJI, 1971). Discussões sobre ambigüidades em linguagem natural vêm desde a era Aristotélica, em obra do próprio Aristóteles, *De sophisticis elnchis*, onde expõe a ambigüidade de sentido nas disputas filosóficas. Um exemplo de ambigüidade de entendimento apresentado na obra é a sentença "lei da natureza". Sendo uma lei entendida como um conjunto de regras estabelecido por alguém, subentende-se que alguém tenha escrito as regras da natureza também. No século XVII a ambigüidade já era uma preocupação do lingüista francês Vaugelas na sua obra *équivoque*. Ele se refere às construções duvidosas (*constructions louches*) como a encontrada na sentença "*la fille du fermier qui nous vend des légumes*", onde o pronome relativo *qui* pode estar se referindo tanto a *fille* como a *fermier*. Em sua obra, Vaugelas prescreve algumas maneiras de se evitar ambigüidades na língua escrita e falada (KOOJI, 1971).

Como comparação entre as linguagens naturais, (JESPERSEN, 1922) encara ambigüidade como sendo uma propriedade inerente de qualquer linguagem natural, mas não procura excluir a possibilidade de que alguma língua possa ser mais adequada que outra com respeito à existência de ambigüidades. (BALLY, 1944) também entende que a ambigüidade esteja presente em todas as línguas, mas é cético quanto à tentativa de provar que uma língua possa ser mais clara que outra, ou seja, conter menos ambigüidades.

A ambigüidade pode ser encarada como um obstáculo para a comunicação, mesmo que, na maioria das vezes, possa ser resolvida automaticamente pelo contexto ou pela situação, em geral, em que se procede a comunicação. Porém, pelo ponto de vista de (EMPSON, 1965), em seu estudo sobre poesia, pode ser um aspecto positivo do texto, no sentido de deixar o leitor entretido com a incerteza. De certa forma a ambigüidade pode ser utilizada, de maneira proposital, como um recurso lingüístico de estilo. A ambigüidade, como recurso, pode ser usualmente encontrada em textos humorísticos e publicitários.

PUC-Rio - Certificação Digital Nº 0621313/CA

Com respeito ao auxílio durante o processo de tomada de decisões é importante que a existência de ambigüidades seja minimizada ou até mesmo eliminada se for possível. O tratamento de ambigüidades, de forma automatizada, demanda estudos em diversos campos desde análise estatística até aprendizado de máquina, passando por técnicas de inteligência computacional e processamento de linguagem natural.

Este trabalho pretende analisar o problema de ambigüidade dentro do contexto de mineração de textos. Para isso, está organizado da seguinte forma: o capítulo 1 faz uma introdução ao universo de mineração de textos, expondo alguns problemas relacionados e também o ferramental atualmente empregado na tentativa de solucioná-los; o capítulo 2 estabelece uma organização estruturada em etapas (*framework*) para o processo de mineração de textos; o capítulo 3 discorre sobre a área de recuperação de informações, apresentando os modelos mais empregados para se recuperar informações textuais; o capítulo 4 define formalmente o problema de desambiguação de sentido de palavras e apresenta as abordagens mais utilizadas para solução desse problema de forma específica, discutindo e comparando suas práticas; o capítulo 5 apresenta uma contextualização do emprego da desambiguação dentro dos processos de mineração de textos e faz ainda um exame dos trabalhos atuais correlatos; o capítulo 6 procura unir tudo que foi apresentado e propõe um modelo para solução para a melhoria de precisão durante a recuperação da informação, por meio da identificação e tratamento de termos ambíguos, baseada em agrupamento de documentos. Além disso, demonstra a capacidade potencial de processamento da aplicação; e finalmente o capítulo 7 apresenta um estudo de caso em universo reduzido, de maneira a exemplificar o processo de melhoria de precisão durante a recuperação da informação.

PUC-Rio - Certificação Digital Nº 0621313/CA