Correção de Palavras em Chats: Avaliação de Bases para Dicionários de Referência

Gustavo Piltcher[1] **, Thyago Borges**[1] **, Stanley Loh**[1 2] **, Daniel Lichtnow**[1] **, Gabriel Simões**[1]

[1]Universidade Católica de Pelotas (UCPEL) – Grupo de Pesquisa em Sistemas de Informação
R. Félix da Cunha, 412 – 96010-000 Pelotas, RS

[2]Universidade Luterana do Brasil (ULBRA) – Faculdade de Informática
R. Miguel Tostes, 101 – 92420-280 Canoas, RS

piltcher@gmail.com, thyago@ucpel.tche.br, sloh@terra.com.br

lichtnow@ucpel.tche.br, gsimoes@vetorial.net

Abstract. *This work presents an analysis of a ortographic corrector for chat environments. To aid this corrections, it was used a combination of metrics for a similarity function development. Some experiments had been made using different databases for the reference dictionary. The reference dictionary is composed by terms compared through the similarity function, with the purpose to point necessary corrections in this chat.*

Resumo. *Este trabalho apresenta a análise de um corretor ortográfico para ambientes de chat. Foi utilizada uma combinação de métricas para criação de uma função de similaridade visando auxiliar as correções. Foram feitos experimentos com diferentes bases para o dicionário de referência. O dicionário de referência é composto por termos que são comparados através da função de similaridade, com a finalidade de indicar as correções necessárias neste chat.*

1. Introdução

O presente trabalho trata da correção de palavras dentro de um ambiente de *chat* (salas de bate-papo). Esta área costuma ser explorada com maior freqüência por editores de texto que tentam auxiliar os usuários na escrita. Porém dentro de um sistema crítico - que dependa da exatidão - qualquer erro na escrita poderá apresentar resultados imprecisos ou indesejados.

Costumam ser encontrados em salas de *chat* usuários com um perfil atípico em relação aos que utilizam um software para edição de texto; eles costumam utilizar uma linguagem informal, suas mensagens são concisas, formadas por apenas um período, sem a preocupação de utilizar pontuações e acentuações corretamente, além de ser comum o uso de abreviações e símbolos específicos deste ambiente. Pode-se observar também o excesso de erros de digitação e a falta de preocupação em corrigi-los, ocorre ainda a inversão de letras, substituição de fonemas semelhantes, utilização de termos abreviados e muitas outras formas de variação lingüística.

No presente trabalho são feitos experimentos com algumas técnicas combinadas e aplicadas para correção individual de palavras, através de um corretor ortográfico automático destinado a ambientes de *chats* na Web. Trata-se de um problema teórico que possui muitas outras aplicações na Ciência da Computação e áreas relacionadas: reconhecimento de padrão sintático [Fu, 1982], geometria fractal [Edgar, 1990] [Blanc-Talon, 1992], *text-to-speech systems* [Liberman and Church, 1991]. As técnicas

de correção apresentadas neste trabalho estão relacionadas à análise de cada palavra separadamente, não sendo avaliados aspectos relacionados à concordância verbal ou nominal. Cabe salientar também que estratégias de processamento de linguagem natural não foram consideradas, pois demandaria de uma maior complexidade de implementação para avaliações. A abordagem utilizada é probabilística/estatística já que esta não requer a utilização de *parsers*.

O corretor ortográfico em questão foi desenvolvido para dar suporte ao *SisRec* (Sistema de Recomendação para apoio a Colaboração), um sistema que consiste em um *chat* Web onde os usuários trocam mensagens e recebem recomendações conforme o assunto identificado nas mensagens. Para tanto, é necessário identificar os assuntos sendo discutidos no chat. A identificação do assunto é realizada através de técnicas de *text mining* apoiadas em uma ontologia de domínio desenvolvida para o sistema, [Kickhöfel and Loh, 2004]. Uma ontologia de domínio (*domain ontology*) é uma descrição de coisas que existem ou podem existir em um domínio [Sowa, 2002] e descreve o vocabulário relacionado ao domínio em questão [Nicola, 1998].

A necessidade de um módulo de correção foi detectada após um determinado período de uso do sistema, pois foram encontradas algumas deficiências na análise das mensagens dos usuários. Muitas delas deixavam de ter assuntos corretamente relacionados devido a equívocos cometidos na escrita, impedindo que o módulo de *text mining* pudesse compará-las com os termos presentes na ontologia. Um corretor interativo se tornaria impreciso e ineficaz neste ambiente, pois uma outra característica dos usuários de *chat* é a pressa. Assim, embora mais complexo, esta necessidade seria melhor atendida mediante o uso de um corretor automático.

O objetivo deste trabalho é apresentar uma análise sobre o grau de sucesso obtido na correção automática de palavras dentro do ambiente de *chat*. Serão comparados ainda os diversos resultados obtidos a partir de diferentes bases de dados utilizadas como dicionário de referência: documentos textuais, históricos de sessões e ontologia.

A seção 2 deste artigo aborda as técnicas e métricas propostas em outros trabalhos que foram combinadas para criar o corretor, a seção 3 faz considerações finais a respeito do corretor e das métricas combinadas, a seção 4 explica o funcionamento do *SisRecCol* e como o corretor automático foi integrado ao mesmo, a seção 5 discute as avaliações feitas e os resultados obtidos, e a seção 6 apresenta as conclusões e contribuições.

2. Trabalhos Correlatos

Muitos trabalhos já exploraram esta área, inúmeros softwares são providos de dicionários para auxiliar a correção ortográfica. Entretanto, partem do princípio de que os usuários dominam o conteúdo escrito, possuem o conhecimento técnico adequado para utilização de termos muito específicos em determinada área ou que dispõem de tempo para revisões. Este tipo de abordagem caracteriza os corretores interativos, necessitando sempre de intervenção humana para que suas correções estejam contextualizadas e sem erros [Kukich, 1992]. Enquanto isso, os corretores automáticos se caracterizam por operarem de maneira transparente para o usuário, eles se encarregam de gerar a lista de possíveis palavras candidatas à correção e avaliam qual a melhor escolha a ser feita. O usuário pode continuar sua tarefa dentro do ambiente sem se preocupar com a escrita e revisão.

Em [Gentner and Grudin, 1983] foi apresentado um experimento em que 58% de todos os erros de substituição foram causados por teclas adjacentes no teclado. Além disso, as taxas de erros devido a posição das teclas giram em torno de 0.4 à 1.9% em pessoas com hábito de digitar e 3.2% naquelas que não o fazem com freqüência. Também se

observou que o tamanho das palavras é um fator determinante para que o usuário cometa um erro ao digitar e grande parte destes erros na ortografia se resumem a no mínimo 2 e no máximo 3 caracteres em relação à forma de escrita correta.

Outros estudos foram feitos por [Pollock and Zamora, 1983] [Yannakoudakis and Fawthrop, 1983] e colaboraram para afirmar que erros de digitação raramente ocorriam na primeira letra de uma palavra. Uma das primeiras descobertas nesta área foi apontada por [Damerau, 1964], desde então tais descobertas tiveram inúmeras aplicações nas técnicas de correção. Ele observou que aproximadamente 80% de todas as palavras digitadas incorretamente, continham pelo menos uma instancia dos quatro erros a seguir: inserção, remoção, substituição e transposição.

Optou-se pela utilização da técnica proposta por [Levenshtein, 1966], pois ela expressa exatamente o que foi descrito acima. A distância de *Levenshtein* exibe a similaridade local entre duas palavras, retornando o número que representa o custo de transformar uma palavra A em uma B. Esta técnica parte do princípio que objetos podem ser representados por palavras e as transformações geométricas nestes objetos são traduzidas em operações entre os caracteres, uma seqüência numérica resultante pode ser vista como o custo destas operações [Luzeaux, 1992]. O algoritmo é flexível o bastante para permitir que se possa especificar os custos de cada operação: substituição/transposição, inserção e remoção. Para o experimento se utilizou os custos 4 (substituição), 1(inserção) e 2(remoção), já que eles apresentaram resultados satisfatórios em testes preliminares.

Existem casos que não podem ser tratados utilizando apenas a métrica anterior, tanto no idioma Inglês como no Português é comum a substituição de caracteres por outros cujo som é igual ou se assemelham. No trabalho apresentado por [Dittenbach and Merkl, 2003] verifica-se esta constatação, eles também propuseram a utilização de métricas que prevêem os padrões sonoros de uma palavra. Seu trabalho apresentou índices de sucesso em aproximadamente 60%, utilizando a métrica conhecida como *Metaphone*, dentro de um sistema de *queries* que eram relacionadas a entidades de um banco de dados SQL. A mesma métrica foi descrita por [Philips, 1990], onde são criadas chaves de caracteres para cadeias com pronúncia semelhante. A desvantagem desta técnica é a necessidade de criar variações de acordo com o idioma no qual ela será aplicada.

Uma outra métrica vista em [Knuth, 1973], chamada de *Soundex*, consiste em um esquema de indexação fonética que transforma qualquer palavra em uma cadeia de 4 caracteres, compostos por 1 letra e 3 números entre zero e seis, onde somente as consoantes das palavras desempenham um papel para a formação da chave. A utilização desta técnica não é tão eficiente como o *Metaphone*, pois seu algoritmo não conhece nenhuma regra de pronúncia do idioma. Entretanto, seu uso é justificado para detecção de palavras onde uma delas é derivada de outra. Nestes casos o *Metaphone* é considerado muito rígido e não apresentaria os resultados esperados. Técnicas como o *stemming* foram evitadas para não comprometer a precisão do sistema.

3. O Corretor Ortográfico para Chat

A função de similaridade presente no corretor automático utilizou um hibrido das métricas *Levenshtein*, *Metaphone* e *Soundex*, para tentar minimizar qualquer tipo de erro que pudesse comprometer a exatidão do sistema de recomendação. Será feita uma análise para cada uma destas métricas e para a função de similaridade como um todo. Esta função também apresenta a necessidade de um limiar, para determinar quando uma palavra precisará ser corrigida. Este valor interfere diretamente na precisão e abrangência dos resultados avaliados. O algoritmo utilizado na função de similaridade dividiu-se em três

etapas:

- Custo da *Levenshtein Distance* entre palavra analisada e palavra do banco.
- Custo da *L.D.* entre a *Metaphone Key* de ambas as palavras.
- Diferença de *bits* entre a *Soundex Key* gerada a partir de cada palavra.

Quanto maior o valor obtido, menores serão as chances das palavras se equivalerem. Utilizaram-se técnicas de programação dinâmica através da criação de *logs* de correção, evitando a necessidade de calcular novamente a similaridade caso apareça a mesma palavra. Também foi utilizado *hashing* para criação de chaves a partir da primeira letra das palavras, adquirindo assim uma maior performance para localiza-las. As palavras relacionadas ao banco de dados foram indexadas em função de sua primeira letra, pois se considerou que estes tipos de erros são raros e não comprometeriam os resultados esperados.

Foram utilizadas três bases distintas e confiáveis presentes no SisRec explicadas detalhadamente em [5]. Para cada uma delas há uma lista de termos assumidos como corretos e que serão utilizados posteriormente para comparação através da função de similaridade. Estas bases se encontram em constante expansão, ao contrário dos dicionários fixos utilizados em editores de texto que integram corretores interativos. A vantagem desta abordagem é a automatização do sistema de correção, o usuário além de não precisar se preocupar com sugestões de correção também não encontrará a necessidade de especificar novas palavras para o dicionário.

4. Funcionamento do Corretor

No SisRec é utilizada uma ontologia de domínio para classificar os documentos na biblioteca digital, traçar o perfil de usuários e identificar assuntos nas mensagens. A ontologia é apresentada como uma hierarquia de conceitos e cada conceito contém uma lista de termos associados e seus respectivos pesos, o que ajuda a identificar o assunto nas mensagens, no decorrer do trabalho utilizou-se novas técnicas como as propostas em [Salton and McGill, 1983] para evitar possíveis discrepâncias nos pesos.

No momento é utilizada apenas uma ontologia referente à área da Ciência da Computação, contendo termos em Inglês e Português para permitir discussões em ambas as línguas. Os pesos associados aos termos determinam a importância relativa ou a probabilidade de um determinado termo identificar um assunto em um texto. O processo de criação desta ontologia é descrito com maiores detalhes em [Loh, 2004].

A ferramenta de *text mining*, utilizada para identificação de assuntos, atua como um *sniffer* examinando cada uma das mensagens enviadas pelos usuários participantes do *chat*. A identificação é feita através da comparação dos termos que aparecem nas mensagens com aqueles relacionados aos conceitos da ontologia. Não é utilizada nenhuma técnica de Processamento de Linguagem Natural, mas sim técnicas probabilísticas que obtiveram índices de acerto superiores a 60% como foi apresentado em [Loh and Wives, 2000].

O corretor foi implementado para atuar como um *middleware* entre o módulo de *chat* e *text mining*, cada mensagem enviada é separada em *tokens* que serão repassados para corretor e comparados, através de uma função de similaridade cujas técnicas foram apresentadas em [2], com os termos presentes em uma base de dados variável. Caso a função de similaridade aponte a necessidade de correção, um novo *token* será enviado para o módulo de *text mining*.

5. Avaliação do Método

O método de correção automática foi avaliado utilizando três bases comparativas. Para todas elas foram utilizadas como entrada os erros de sessões do SisRec. Tratou-se como sessão uma discussão *online* no sistema, através de um ambiente de *chat*, entre pessoas da área de Computação discutindo sobre assuntos relacionados. Esta sessão era composta por inúmeras palavras, onde foram selecionadas todas aquelas que não estavam presentes na ontologia do sistema, criando assim uma lista de termos assumidos como erros de sessões. Os termos presentes nesta lista caracterizam o universo de correção, ou seja, palavras candidatas à correção.

Seguindo este critério obteve-se 7652 palavras, entre elas estavam palavras digitadas incorretamente, novas palavras ou termos sem significado. Apenas aquelas com grafia incorreta interessavam para este experimento. Foi feita uma análise manual por pessoas da área de Computação obtendo 2976 (38%) palavras realmente incorretas entre as 7652 selecionadas, as demais estavam corretas e foram separadas porque não estavam presentes na ontologia. Este processo findou na criação, também manual, de uma lista de *correções esperadas* para cada uma das 2976 palavras constatadas como incorretas. Esta lista serviu como modelo comparativo para afirmar se a correção feita de maneira automática estava correta ou não.

Os experimentos de correção utilizaram três diferentes bases de dados, descritas a seguir, como dicionário de referência. O dicionário de referência foi composto por palavras, assumidas como corretas e extraídas de uma base confiável. Elas foram utilizadas individualmente para comparar sua similaridade em relação a cada erro de sessão.

A função de similaridade recebe um par de palavras (A, B) e retorna o percentual de similaridade entre elas. A comparação feita utilizou como entrada para a função de similaridade o par de termos (erro de sessão, palavra dicionário). Para cada erro de sessão (7652) fizeram-se K comparações, onde K representa o total de palavras do dicionário de referência. Alocou-se para este dicionário apenas as palavras cuja letra inicial era idêntica a da palavra incorreta a ser corrigida, reduzindo assim o total de comparações necessárias e tempo de processamento.

Foram armazenados todos os percentuais obtidos durante o processo comparativo, tornando possível localizar qual termo, presente no dicionário de referência, possuía maior similaridade com o erro de sessão em questão.

O termo do dicionário que possuísse a maior similaridade em relação ao erro de sessão e que estivesse acima de um limiar estipulado (72% e 84%), foi assumido como a correção ideal. Em seguida, comparou-se o *termo corrigido* com a *correção esperada* para saber se a correção feita de maneira automática estava correta ou não.

Segue abaixo a descrição das três bases utilizadas como dicionário de referência:

- **Históricos de sessões:** todas as mensagens enviadas ao ambiente de Chat são gravadas, compondo assim um histórico de sessões de todos os usuários. Assumiu-se que, se um termo aparecia com freqüência dentro do *Chat*, então ele estava grafado corretamente e assim poderia ser utilizado como base para as comparações dos termos tidos como incorretos.
- **Documentos textuais:** descreveu-se anteriormente que o SisRec possuía uma Biblioteca Digital. O universo destes documentos é composto por artigos científicos e livros técnicos da área de Computação. Um dos dicionários de referência foi formado por termos extraídos destes documentos, pois se presume que são fontes confiáveis de termos grafados corretamente. Foram testados termos de um grupo de 35 e outro de 110 artigos. Também se fez teste com termos vindos de ape-

nas um documento, e neste caso, foram utilizados dois livros técnicos (*MySQL Manual*, *Gurps Book*) nas versões inglês e português.

- **Ontologia:** considera-se a ontologia como mais confiável em relação as anteriores porque foi criada de modo supervisionado por humanos. Cada um de seus conceitos possui uma lista de termos relacionados e todos eles integram o dicionário de referência. Os termos da ontologia foram detectados através de aprendizado supervisionado, onde um especialista na área selecionou documentos científicos sobre o conceito e uma ferramenta automática extraiu os termos mais significativos. Acredita-se que a utilização desta base possa apresentar os resultados mais satisfatórios, visto que, ao contrário dos dois itens anteriores, todas as palavras que vierem a ser corrigidas serão necessariamente termos desta ontologia. Como foi dito anteriormente, técnicas de *stemming* não foram utilizadas, nem mesmo para integrar termos na ontologia. Para efeito comparativo se utilizou duas ontologias presentes no SisRec. A primeira foi gerada de modo supervisionado e não teve intervenção humana. A segunda ontologia é a primeira com intervenção humana, isto é, especialistas na área de Computação revisaram os termos de cada conceito, eliminando os muito genéricos ou rebaixando seu peso e acrescentando termos importantes que não estavam presentes e também acrescentando variações lingüísticas de gênero, número e conjugações verbais importantes ou mais freqüentes.

5.1. Experimentos

Para melhor compreensão dos resultados deve ficar claro que a entrada utilizada em todos os experimentos era composta por 7652 palavras e que 2976 delas precisavam de correção.

A interpretação das figuras deve ser feita da seguinte forma: *termos extraídos* representam a quantidade de palavras - não repetidas - que compuseram o dicionário de referência a partir de uma das três bases explicadas anteriormente; *termos corrigidos* são o total de palavras, entre as 7652, que sofreram a correção automática, seja ela correta ou não; *termos corretos* é a intersecção entre o conjunto de *termos corrigidos* e a lista de *correções esperadas* (2976), ou seja, o total de correções exatas; *precisão* é obtida pela razão entre *termos corretos* e *termos corrigidos* e *abrangência* através da razão entre *termos corretos* e *correções esperadas*.

Freqüência \ Termos	Extraídos	Corrigidos	Corretos	Precisão	Abrangência
maior ou igual a 1	7910	392	126	0,32	0,04
maior ou igual a 3	1800	375	102	0,27	0,03
maior ou igual a 5	1092	369	98	0,27	0,03
maior ou igual a 8	717	352	94	0,27	0,03

Figura 1: Experimento com Histórico de Sessões (limiar: 72%)

Freqüência \ Termos	Extraídos	Corrigidos	Corretos	Precisão	Abrangência
maior ou igual a 1	7910	127	95	0,75	0,03
maior ou igual a 3	1800	126	95	0,75	0,03
maior ou igual a 5	1092	103	84	0,82	0,03
maior ou igual a 8	717	88	71	0,81	0,02

Figura 2: Experimento com Histórico de Sessões (limiar: 84%)

5.2. Avaliação com Histórico de Sessões

Nesta avaliação foi definido um limiar L representando a freqüência mínima que um termo necessitava possuir para que fosse considerado grafado corretamente e, por conseqüência, adicionado ao dicionário de referência. A análise consistiu na atribuição de diferentes valores para L=1,3,5,8, verificados em (Figura 1) e (Figura 2) juntamente com seus respectivos resultados.

Documento \ Termos	Extraídos	Corrigidos	Corretos	Precisão	Abrangência
Mysql Manual Inglês	13500	1921	473	0,25	0,16
Mysql Manual Português	18568	2135	701	0,33	0,24
Gurps Book Inglês	11351	1811	462	0,26	0,16
Gurps Book Português	14692	1794	512	0,29	0,17
Biblioteca Digital (35 docs)	8944	1633	631	0,39	0,21
Biblioteca Digital (110 docs)	12857	2253	984	0,44	0,33

Figura 3: Experimento com Documentos (limiar: 72%)

Documento \ Termos	Extraídos	Corrigidos	Corretos	Precisão	Abrangência
Mysql Manual Inglês	13500	849	451	0,53	0,15
Mysql Manual Português	18568	1276	657	0,51	0,22
Gurps Book Inglês	11351	763	398	0,52	0,13
Gurps Book Português	14692	1275	493	0,39	0,17
Biblioteca Digital (35 docs)	8944	907	608	0,67	0,20
Biblioteca Digital (110 docs)	12857	1318	912	0,69	0,31

Figura 4: Experimento com Documentos (limiar: 84%)

Ontologia \ Termos	Extraídos	Corrigidos	Corretos	Precisão	Abrangência
com intervenção humana	2275	928	448	0,48	0,15
sem intervenção	7751	2891	903	0,31	0,30

Figura 5: Experimento com Ontologias (limiar: 72%)

Ontologia \ Termos	Extraídos	Corrigidos	Corretos	Precisão	Abrangência
com intervenção humana	2275	687	411	0,60	0,14
sem intervenção	7751	2056	1074	0,52	0,36

Figura 6: Experimento com Ontologias (limiar: 84%)

Termos corrigidos	1018
Termos corretos	437
Precisão	0,43
Abrangência	0,15

Figura 7: Experimento com Editor de Texto MS-Word 2003

Métricas \ Termos	Extraídos	Corrigidos	Corretos	Precisão	Abrangência
Levenshtein	7751	1658	604	0,36	0,20
Soundex	7751	254	103	0,41	0,03
Metaphone	7751	351	201	0,57	0,07

Figura 8: Experimento com métricas isoladas

Pode-se observar que o limiar L influenciou diretamente a quantidade de palavras extraídas dos históricos: 7910 termos (L=1); 1800 termos (L=3); 1092 termos (L=5) e 717 termos (L=8). Embora este último represente 9% dos termos do primeiro, sua abrangência (proporção de termos corretos em relação àqueles que deveriam receber uma correção) continuou praticamente a mesma, variando apenas 1% tanto na utilização do limiar de similaridade em 72% como também em 84%. O limiar L=5 demonstrou o melhor resultado, isto significa que ele seria o valor ideal para determinar quais termos iriam compor o dicionário de referência.

É importante salientar que neste segundo limiar a precisão manteve uma proporção superior ao primeiro, na ordem de 33% a 42%. Este fato pode levar a conclusão de que o limiar de similaridade em 84% apresenta resultados muito mais satisfatórios já que sua precisão aumenta sem afetar a abrangência.

Entretanto, não era esperada uma abrangência tão baixa (entre 2% e 4%). Isto demonstra que esta base não é tão confiável como se acreditava, os erros de grafia e abreviações ali presentes tendem a ser repetidos muitas vezes, confundindo o corretor que por sua vez pensava estar tratando de algo grafado corretamente. Entre as poucas correções feitas, estavam palavras utilizadas com grande freqüência e cujos erros não eram tão comuns, ou seja, a base não contribuiu suficientemente para que correções importantes fossem feitas.

5.3. Avaliação com Documentos Textuais

Através da utilização de um livro técnico e outro não pertencente à área da Computação, foi possível avaliar se os erros de sessões refletiam a falta de conhecimento da grafia de termos técnicos. Foram utilizadas duas versões de cada livro, uma no idioma português e outra no inglês, pois como foi visto em [4] as discussões no *chat* podem ser em ambos os idiomas.

Os resultados obtidos em (Figura 3) demonstram que o livro técnico tem uma abrangência 8% superior em relação ao não técnico, quando este é utilizado como dicionário de referência. Isto deixa clara a necessidade de um dicionário de referência composto não só por dois idiomas, mas também por palavras técnicas comuns aos tópicos em discussão.

Esta propriedade está presente nos documentos da Biblioteca Digital. Tomando como base a (Figura 4) onde a precisão é até 47% superior à anterior, será verificado que a quantidade de documentos afetará diretamente na abrangência do corretor enquanto a precisão sofrerá uma variação insensível. A prova disto é que depois de aumentar a quantidade de documentos em 75 (314%), a precisão variou apenas 2% enquanto a abrangência deu um salto de 11%.

O resultado já era esperado, pois estes documentos costumam ser revisados por seus autores diversas vezes, evitando assim erros de grafia. Entretanto, foi descoberto que se comparados os termos de um documento técnico A com um não técnico B, todos aqueles exclusivos de A são em sua maioria palavras que poderiam fazer parte da ontologia.

5.4. Avaliação com Ontologias

Assim como nas duas avaliações anteriores, o limiar de similaridade em 84% demonstrou ser o ideal para utilização final no sistema, ele apresentou resultados mais satisfatórios que o de 72%. A verificação pode ser feita na (Figura 5) obteve-se uma precisão de 48% para ontologia com intervenção humana, 31% para a sem intervenção e abrangência de 15% e 30% respectivamente. Enquanto os experimentos demonstrados na (Figura 6) apresentaram precisões de 60% para ontologia com intervenção humana e 52% na sem intervenção, suas abrangências foram 14% e 36% (a mais alta dos experimentos, como o esperado).

A utilização da ontologia sem intervenção resultou em 7751 compondo o dicionário de referência, ajudando na correção bem sucedida de 1074 termos. A Biblioteca Digital utilizou 60% de termos a mais (12857), mas obteve sucesso em 912 deles (8% a menos que a ontologia). Esta observação confirma um detalhe despercebido nos experimentos anteriores, que a abrangência esta diretamente ligada à qualidade dos termos adotados para o dicionário, mas que a quantidade contribui para a melhoria da precisão caso não haja a revisão dos termos.

5.5. Avaliações Comparativas

Foram feitos dois experimentos com caráter comparativo e utilizando os mesmos critérios de avaliação dos anteriores. O primeiro tem como finalidade demonstrar a precisão e abrangência de um corretor interativo, explicado em [2], utilizado freqüentemente em editores de texto. Pela popularidade do software e qualidade de correção, o MS-Word 2003 foi escolhido para o experimento.

Na (Figura 7) foi apresentado que seu corretor atingiu a precisão de 43%, para este resultado foi utilizada a opção 'corrigir todas' do software. Utilizando este método

foi evitada a necessidade de intervenção por parte do usuário, pois ele poderia afetar diretamente a precisão obtida - também impedindo a comparação interativa X automática - já que neste caso a escolha não seria feita de maneira automática.

A abrangência considera que o software tenha feito a correção mais adequada para as palavras, mas não houve possibilidade de uma avaliação mais detalhada pelo fato dele possuir código fechado. Um total de 437 (15% de abrangência) palavras, entre as 1018 correções, foram corrigidas coincidindo com as *correções esperadas*. Isto demonstra que este corretor interativo apresenta resultados similares ao da ontologia com intervenção, mas 41% inferior ao da ontologia sem intervenção, o motivo disto é a deficiência de termos técnicos no dicionário do software. Sua precisão ficou em 43%, um valor satisfatório visto que nenhum usuário indicou qual correção deveria ser feita.

O segundo experimento demonstrado na (Figura 8) utilizou como limiar de similaridade o 84% (eleito o ideal) e a base de dados foi a ontologia sem intervenção, pois obteve a maior abrangência. Sua finalidade foi demonstrar que os resultados obtidos anteriormente no segundo item da (Figura 6) teriam uma queda significativa na qualidade (precisão e abrangência) se as métricas que compõe a função de similaridade fossem utilizadas separadamente.

A técnica de Levenshtein apresentou a precisão e a abrangência 16% inferiores em relação aos resultados obtidos com a função de similaridade. As demais técnicas, mesmo tendo uma boa precisão, obtiveram uma abrangência pouco significativa (entre 3% e 7%).

6. Conclusão

Este trabalho avaliou a correção de palavras dentro de um ambiente de *chat*. Para tanto foi implementada uma função de similaridade combinando três diferentes métricas. A dependência de um dicionário de referência levou a avaliação e comparação de: três bases de dados, um editor de texto e três métricas.

A contribuição do artigo (e também sua diferença para outros trabalhos já comentados) está na avaliação do grau de sucesso obtido em correções sem que haja a necessidade de intervenção do usuário. Também é demonstrado que por si só, uma métrica não apresenta condições de detectar todos os erros de grafia cometidos por usuários em um sistema crítico (como é o caso do ambiente de *chat*). Por esta razão, foi proposto um método com o objetivo de contornar/identificar erros de digitação, substituição de fonemas e derivação de palavras. Método este, que obteve resultados mais satisfatórios que os utilizados em corretores interativos convencionais.

Outra constatação importante é que se um erro de sessão for corrigido por um termo abrangido pela ontologia, então ele contribuirá para a identificação do assunto através do módulo de *text mining*, já que obrigatoriamente este termo fará parte dela. Por outro lado, os termos obtidos com correções feitas a partir de uma Biblioteca Digital ou de um corretor interativo, poderão não fazer parte da ontologia e por conseqüência não irão contribuir para identificação de assuntos.

A aplicação dos resultados deste trabalho poderá melhorar o processo de identificação de assuntos em mensagens de *chat*. Isto tem conseqüências diretas sobre sistemas que analisam discussões em *chats*, tais como:

- Sistemas de Identificação de Especialistas ou Análise de Expertise: para encontrar pessoas autoridades em determinado assunto ou simplesmente identificar quem conhece algo sobre algum assunto;
- Sistemas de Recomendação: para indicar itens de forma sensível ao contexto (ofertas personalizadas conforme interesse de cada pessoa que participa do chat);

7. Agradecimentos

O presente trabalho foi realizado com o apoio do CNPq, uma entidade do Governo Brasileiro voltada ao desenvolvimento científico e tecnológico.

Referências

Blanc-Talon, J. (1992). How can grammars generate fractal curves? *Technical Report, CSIRO-CSIS-DIT, Canberra ACT, Australia*.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Commun. ACM*, (7):171–176.

Dittenbach, M. and Merkl, D. (2003). A natural language query interface for tourism information. *In Proc of the 10th Int. Conf. on Information Technologies in Tourism (ENTER 2003)*, pages 152–162.

Edgar, G. (1990). Measure, topology and fractal geometry. *UTM, Springer Verlag*.

Fu, K. (1982). *Syntactic Pattern Recognition and Applications*. Prentice-Hall.

Gentner, D. and Grudin, J. (1983). Studies of typing from the lnr typing research group. *Cognitive Aspects of Skilled Typewriting, Springer Verlag, New York.*

Kickhöfel, R. and Loh, S. (2004). Tecnologias de softwre livre para análise de mensagens em chats e recomendações online de documentos eletrônicos. *Anais do 5° Workshop sobre software livre*, pages 31–34.

Knuth, D. (1973). *The Art of Computer Programming*. Addison-Wesley, 3 edition.

Kukich, K. (1992). Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24(4):377–439.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8).

Liberman, M. and Church, K. (1991). Text analysis and word pronunciation in text-to-speech synthesis. *Advances in Speech Signal Processing, Marcel Dekker, New York*.

Loh, S. (2004). Investigação sobre a identificação de assuntos em mensagens de chat. *Workshop de TI e Linguagem Humana - XXIV Congresso da Sociendade Brasileira de Computação, Salvador*.

Loh, S. and Wives, L. (2000). Concept-based knowledge discovery in texts extracted from the web. *ACM SIGKDD Explorations*, 2(1):29–39.

Luzeaux, D. (1992). String distances. *Distancia Conference, Rennes, France*.

Nicola, G. (1998). Formal ontology and information systems. *International Conference on Formal Ontologies in Information Systems*, pages 3–15.

Philips, L. (1990). Hanging on the metaphone. *Computer Language Magazine*, 7(12).

Pollock, J. and Zamora, A. (1983). Collection and characterization of spelling errors in scientific and scholarly text. *J. Amer. Soc. Inf. Sci.*, 34(1):51–58.

Salton, G. and McGill, M. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Sowa, J. (2002). Building, sharing, and merging ontologies. *AAAI Press / MIT press*, pages 3–41.

Yannakoudakis, E. and Fawthrop, D. (1983). The rules of spelling errors. *Inf. Process Manage*, 19(12):101–108.