Extração de Termos de Manuais Técnicos de Produtos Tecnológicos: uma Aplicação em Sistemas de Adaptação Textual

Fernando A. M. Muniz[1], Willian M. Watanabe[2], Carolina E. Scarton[1], Sandra M. Aluisio[1]

[1]NILC - ICMC - Universidade de São Paulo São Carlos - SP, Brasil

[2]Intermídia - ICMC - Universidade de São Paulo São Carlos - SP, Brasil

{fernando.muniz,watinha,carol.scarton}@gmail.com, sandra@icmc.usp.br

Abstract. *In tasks that require the use of technical documentation, the quality of the texts is critical. If the documentation is imprecise, incomplete or too complex, it increases the cost of realizing the tasks and might lead to higher accident rates. This work reports on how the procedural relationships (Goldman's procedural relations of Generation and Enablement) are represented in portuguese technical manuals and can be used to generate an automatic term extraction method, specifically in the context of technical manuals. We also present the adaptation of a simplified texts authoring system to assist the elaboration of technical manuals. The adaptation of the authoring system uses the list of candidate terms extracted to: (a) help identify words that should not be simplified by lexical simplification methods, and (b) elaborate the technical terms, to improve the understanding of texts.*

Resumo. *Em tarefas que exigem o uso de documentação técnica, a qualidade da documentação é um ponto crítico, pois caso ela seja imprecisa, incompleta ou muito complexa, o custo da tarefa ou até mesmo o risco de acidentes aumenta. Este trabalho estudou como as relações procedimentais entre ações gera e habilita, propostas por Goldman, são realizadas em manuais em português, para a criação de um método de extração automática de termos dedicado ao gênero procedimental, especificamente, manuais técnicos. Também propôs uma adaptação de um sistema de autoria de textos simplificados para atender este gênero textual. Esta adaptação usa a lista de candidatos a termos extraídos para: (a) auxiliar na identificação de palavras que não devem ser simplificadas pelo método de simplificação léxica, e (b) elaborar tais termos, facilitando o entendimento dos textos.*

1. Introdução

Tarefas operacionais, procedimentos de manutenção e diagnósticos de falhas em sistemas técnicos complexos requerem o uso de documentação técnica. Se a documentação está imprecisa, incompleta ou difícil de entender, o custo e o tempo da operação de reparo poderá aumentar muito, desta forma a qualidade dessa documentação é um ponto crítico. Até mesmo prejuízos a equipamentos caros ou acidentes com vítimas humanas podem ocorrer devido ao mau entendimento da documentação técnica (Eijk, 1997).

Textos de manuais são do gênero procedimental (ou instrucional). Estes textos deveriam conter uma sequência de instruções concebidas com certa precisão a fim de alcançar um objetivo. O leitor deveria ser capaz de seguir passo a passo cuidadosamente as instruções fornecidas pelo manual a fim de alcançar o objetivo (Fontan and Saint-Dizier, 2008).

O filósofo Alvin Goldman identificou duas relações procedimentais básicas, *generation* (gera) e *enablement* (habilita) (Goldman, 1970). A relação gera aparece entre duas ações e passa o sentido de que após a realização da ação "A", a ação "B" ocorrerá automaticamente, ou seja, "A" gera "B". No português, expressões linguísticas da relação gera geralmente envolvem o conectivo "para", primeiramente seguido por um infinitivo e, em ocasiões raras, seguido por um sintagma nominal. O seguinte trecho de um manual de instruções, em português, de uma serra elétrica exemplifica essa relação (Delin, 1994): *Para colocar a serra na posição de corte oblíquo, solte a porca borboleta e incline a sapata para o ângulo desejado* (Black&Decker).

A relação habilita ocorre quando a realização de uma ação "A" não resulta na realização automática da ação "B". Apesar do conectivo "para" também ser usado para a relação habilita, ele não foi encontrado em (Delin, 1994). Ao contrário, neste estudo, as relações habilita foram encontradas através de sinais de ordem temporal nas ações envolvidas, em orações consecutivas ou ligadas pela conjunção "e": *Desligue a serra da tomada antes de fazer qualquer ajuste* (Black&Decker).

Em Paris *et al.* (1995) uma análise de requisitos para uma ferramenta de suporte à escrita de documentos técnicos multilíngue confirmou que uma ferramenta de auxílio a escrita é mais útil do que uma ferramenta de geração automática que mantém o escritor longe do texto produzido. Paris *et al.* (1994) mostra que os manuais de instruções podem ter diferentes estilos; nem todas as instruções usam uma sequência de imperativos, como seria mais natural de se esperar, e que diferentes partes do manual frequentemente usam diferentes estilos. Aouladomar (2005) faz uma análise da estrutura de manuais e de perguntas relacionadas a textos procedimentais (por exemplo: "Como?" e "Por quê?") e mostra que perguntas e fragmentos de textos procedimentais podem ser combinados a fim de produzirem respostas para máquinas de busca.

O desenvolvimento contínuo de novas tecnologias e produtos, combinados com o fato de que grande parte da população (68%) tem um nível básico e rudimentar de letramento (INAF, 2009) torna clara a importância da adaptação de manuais técnicos para atender as necessidades de grande parte da população. Além disso, é um assunto interessante a ser estudado.

O projeto PorSimples[1] (Aluísio and Gasperin, 2010) aplicou as duas abordagens de Adaptação Textual, a Simplificação e a Elaboração, para ajudar leitores com baixo nível de alfabetização e letramento a compreender documentos disponíveis na Web em português do Brasil, principalmente textos jornalísticos. O trabalho apresentado nesse artigo também se dedicou à elaboração textual, mas o foco foi o gênero de textos instrucionais. Foram estudadas as realizações das relações procedimentais entre ações gera e habilita em manuais de instruções em português, para dar base para a criação de um método de extração de termos dedicado ao gênero de textos instrucionais, especificamente, aos manuais técnicos. Além disso, também propomos uma adaptação

[1] http://caravelas.icmc.usp.br/wiki/

do sistema de autoria de textos simplificados criado no projeto PorSimples – o SIMPLIFICA – para atender a escrita de manuais. O SIMPLIFICA adaptado usa a lista de candidatos a termo gerada pelo método proposto para executar duas funções: (a) auxiliar na identificação de palavras que não devem ser simplificadas pelo método de simplificação léxica baseado em sinônimos, e (b) receber uma elaboração léxica para facilitar o seu entendimento.

Vale destacar que este trabalho tem como objetivo facilitar o entendimento e compreensão de textos instrucionais por pessoas com baixo nível de alfabetização, utilizando técnicas computacionais de processamento de língua natural. Dessa forma, a proposta contribui para o processo de adquirir conhecimentos pelos usuários, sem discriminá-los pela sua condição social ou nível de alfabetização, atuando junto ao tema do SEMISH 2011 - "Computação para todos: No caminho da Evolução Social", mais especificamente no quarto Grande Desafio de Pesquisa da Computação de "Acesso participativo e universal do cidadão brasileiro ao conhecimento".

Na próxima seção, descrevemos trabalhos relacionados ao nosso nas áreas de extração automática de termos e adaptação textual. Na Seção 3, o projeto NorMan – Normalização de Manuais – que compreende o método de extração de termos dedicado a manuais técnicos e a aplicação do resultado da extração em sistemas de adaptação textual, é apresentado. A Seção 4 detalha os experimentos de avaliação intrínseca e extrínseca realizados e, na Seção 5, apresentamos nossas conclusões e trabalhos futuros.

2. Trabalhos Relacionados

2.1. Extração Automática de Termos para o Português

Teline (2004) fez uma avaliação de três abordagens de extração automática de termos: linguística, estatística e híbrida. Na abordagem estatística foram usadas as medidas de Frequência para unigramas, Frequência, Informação Mútua, Log-Likelihood e Coeficiente Dice para bigramas e para trigramas foram usadas Frequência, Informação Mútua, Log-Likelihood. Em seguida foi feita uma análise manual da lista de unigramas e bigramas candidatos a termos. Esta intervenção foi feita com o intuito de eliminar palavras e siglas da língua geral, marcas publicitárias, nomes próprios e símbolos especiais. Neste caso, o método estatístico com intervenção humana é considerado um método semiautomático. Na abordagem lingüística, foi feito um pré-processamento no corpus para permitir a realização de consultas sobre o mesmo. O primeiro tipo de consulta realizada no corpus é a busca por expressões e indicadores estruturais, que são expressões lingüísticas, que geralmente vêm acompanhados de definições, descrições e outros tipos de orações que concentram termos. Em seguida é feita uma busca padrões morfossintáticos (Ex. substantivo + adjetivo). Além disso, nesta abordagem é feito o uso de uma *stoplist*, que é uma lista contendo palavras da língua geral. Na abordagem híbrida, foi feito o processamento do corpus e a buscas por expressões e indicadores estruturais, em seguida foram aplicados métodos estatísticos (cálculo de frequência para unigramas, bigramas e trigramas e informação mútua para bigramas). Após essas etapas, foi feita uma intersecção com a lista de padrões morfossintáticos. O método implementado que teve a melhor precisão foi o estatístico com interação humana. A melhor revocação foi obtida pelo método linguístico e a medida F com maior valor também foi obtida pelo método estatístico com interação humana.

Ribeiro Jr (2008) trabalhou com construção de ontologias e fez o uso de um método de extração automática de termos, utilizando uma abordagem híbrida. Os conhecimentos linguísticos utilizados foram a classe gramatical, sintagmas nominais e padrões morfossintáticos. Em seguida, foram aplicados cálculos de relevância de frequência, tf-idf (Medida que considera relevantes os termos que possuem alta frequência de ocorrência em numero limitado de documentos) e NC-Value, que estão descritos em (Ribeiro Jr, 2008). Na extração de unigramas, o melhor resultado obtido foi com o uso das classes gramaticais e o núcleo do sintagma nominal combinado com o cálculo NC-Value usando tf-idf como parâmetro de frequência que obtiveram os melhores índices de precisão (14,7%), cobertura (49,96%) e medida F (22,39%). Para extração de bigramas, o melhor método foi o uso de padrões morfossintáticos com o cálculo de relevância de frequência, obtendo o índice de precisão de 5,7%, cobertura de 41,91% e medida F de 10,04%. Na extração de trigramas, a melhor estratégia foi o uso de padrões morfossintáticos com o cálculo de relevância tf-idf, obtendo índice de precisão de 2,9%, cobertura de 46,77% e medida F de 5,46%.

Lopes *et al*. (2010) apresentou uma avaliação de termos compostos para um corpus do domínio de pediatria. A extração de termos foi feita usando três métodos diferentes e sua performance foi avaliada através do cálculo da precisão, revocação e medida F, com base numa lista de referência de termos compostos feita a mão. A extração automática de termos foi realizada em três etapas: anotação linguística do corpus, utilizando o parser PALAVRAS (Bick, 2000); extração de termos usando a ferramenta OntoLP (Ribeiro Jr, 2008) e a seleção de termos pelo ponto de corte. A extração de termos simples foi feita através da classe gramatical. Na extração de termos compostos, três métodos foram usados. O primeiro método foi puramente estatístico. O segundo método utilizou padrões morfossintáticos e o terceiro método utilizou sintagmas nominais para tentar identificar termos que poderiam estar presentes em seu núcleo. Após a extração, a lista de candidatos a termos foi ordenada através do calculo de relevância pela frequência. Os experimentos realizados usaram corte absolutos e cortes relativos. Os resultados do experimento que usou corte absoluto mostram que na extração de bigramas o método de padrões morfossintático foi o melhor (medida F = 0.5684 em 6E-6), seguido pelo método estatístico (medida F = 0.5478 em 6E-6) e pelo método de sintagmas nominais (medida F = 0.4437 em 4E-6). Para trigramas, o melhor método foi o estatístico (medida F = 0.5211 em 6E-6), seguido pelo método de padrões morfossintáticos (medida F = 0.4612 em 6E-6) e pelo método de sintagmas nominais (medida F = 0.3671 em 4E-6). Os resultados do experimento que usou ponto de corte relativo mostraram que na extração de bigramas, o método de padrões morfossintáticos foi o melhor (medida F = 0.5547 em 10%), seguido pelo método estatístico (medida F = 0.5233 em 10%) e pelo método de sintagmas nominais (medida F = 0.4393 em 10%). Para trigramas, o método estatístico obteve o melhor resultado (medida F = 0.4831 em 5%), seguido pelo método de padrões morfossintáticos (medida F = 0.4539 em 5%) e pelo método de sintagmas nominais (medida F = 0.3510 em 10%)

2.2. Adaptação Textual

Adaptação textual (AT) é uma atividade comum de professores para facilitar a compreensão de conteúdos e também usada no cenário do ensino de novas línguas (Burstein, 2009). Seus benefícios são vários, por exemplo, ajudar aprendizes de língua e crianças aprendendo a ler diferentes gêneros textuais e também audiências com

necessidades especiais. Tais audiências podem ser leitores com baixo nível de letramento, adultos em fase de alfabetização, pessoas engajadas em cursos de Educação à Distância, para os quais a compreensão é de grande importância. Surdos que se comunicam com a língua de sinais e desejam aprender línguas como português ou inglês também se beneficiam da AT (Aluísio and Gasperin, 2010).

Os estudos na área de AT tentam responder duas questões: *O que é modificado?* e *Como é modificado?*. Para responder à primeira questão as pesquisas investigam modificações nos diferentes níveis linguísticos: fonológico, léxico, sintático e discursivo. Já para a segunda, existem duas grandes abordagens (ou tipos) de adaptações: Simplificação Textual (ST) e Elaboração Textual (ET) (Young, 1999; Urano, 2000). A primeira pode ser definida como qualquer tarefa que reduza a complexidade de um texto (por exemplo, no nível léxico ou sintático), enquanto tenta preservar o significado e informação (Siddharthan, 2003). A Elaboração Textual, por sua vez, é definida como um conjunto de técnicas para acrescentar material redundante em textos, sendo tradicionalmente usadas a adição de definições, sinônimos, antônimos, ou qualquer informação externa com o objetivo de auxiliar na compreensão do texto por meio dessa informação complementar (Rahimi, 2011). É importante dizer que a simplificação e a elaboração são fortemente relacionadas; enquanto a simplificação aumenta a inteligibilidade de um texto (torna ele mais fácil de ser lido), a elaboração melhora a compreensão do texto, isto é, facilita o entendimento de conceitos nos textos. Abaixo, relatamos os trabalhos sobre simplificação léxica próximos ao nosso, pois é este o nível da língua tratado no projeto NorMan.

Elhadad (2006) estudou como melhorar o acesso de pessoas sob tratamento ou cuidados médicos à literatura médica, com foco na terminologia da área médica. Um método para a predição em um texto de quais termos um leitor comum pode não entender foi proposto, usando definições tomadas da Web para melhorar a compreensão de tais termos não familiares. A predição é realizada via contagem de frequência de termos em corpora de textos médicos dedicado a pessoas comuns, para achar o corte na frequência ideal para separar palavras comuns de não familiares. Os autores classificam as abreviações como palavras complexas.

Devlin and Unthank (2006) apresentam o projeto HAPPI (Helping Aphasic People Process Information) em que usam um método para simplificar palavras que os leitores não entendem por palavras mais comuns com o mesmo significado e, além da simplificação léxica, imagens e áudio com versões faladas dos conceitos são apresentados aos usuários.

Existem vários estudos sobre elaboração textual para o inglês e um trabalho recente para o português, apresentados abaixo.

A ferramenta ATA (The Automated Text Adaptation Tool) (Burstein *et. al.*, 2007) é uma aplicação da área de Processamento de Língua Natural (PLN) para propósitos educativos, que é usada por aprendizes do inglês em aulas para o ensino fundamental nas quais os professores a usam para gerar adaptações de textos similares àquelas feitas manualmente. São usados sinônimos para palavras de baixa frequência, antônimos usando a WordNet de Princeton[2] e listas de falsos cognatos (homógrafos com significados diferentes), além de notas marginais (tipo de resumo) traduzidas para o espanhol.

[2] http://wordnet.princeton.edu/

Urano (2000) investigou os efeitos da simplificação e elaboração léxicas para compreensão e aquisição incidental de vocabulário por falantes japoneses aprendendo o inglês como segunda língua. As modificações realizadas foram a substituição de palavras de baixa frequência por sinônimos de alta frequência, e a adição de sinônimos seguindo as palavras desconhecidas. Os resultados deste estudo sugerem que ambas as abordagens ajudam a melhorar a compreensão no nível sentencial, entretanto, a elaboração léxica resulta em aquisição de vocabulário incidental, enquanto que a simplificação não.

Watanabe *et .al.* (2010) aplica seus estudos em pessoas com baixo letramento que navegam na Web em busca de informação. Foram usadas elaboração léxica via sinônimos mais simples, além de definições curtas da Wikipédia para definir entidades nomeadas que aparecem nos textos, além de enfatizarem estas. O conjunto de entidades nomeadas foi o proposto na avaliação conjunta de sistemas de reconhecimento de entidades nomeadas do português (HAREM[3]) (i.e., pessoa, organização, local, artefatos, cosias, eventos, entidades abstratas, quantidades, tempo, coisas produzidas por pessoas). Além disso, figuras para as entidades nomeadas da classe pessoa foram usadas.

3. O Projeto NorMan

3.1. O Método de Extração de Termos Dedicado a Manuais Técnicos

Durante a tarefa de simplificação léxica de manuais técnicos, é preciso tomar um cuidado especial com relação aos termos técnicos presentes no texto. Manuais técnicos contêm termos que não podem ser excluídos ou trocados por possíveis sinônimos da língua geral. Caso um termo técnico seja erroneamente suprimido do texto ou trocado por um mais simples durante o processo de simplificação léxica, o sentido final da sentença poderá ficar seriamente comprometido, prejudicando o entendimento do leitor. Para evitarmos esse tipo de situação, é proposto um novo método de extração de termos, sensível ao gênero de instruções. Este método foi baseado nas relações gera e habilita, que são comumente encontradas em manuais. Essas relações são de suma importância, pois elas demonstram explicitamente o que o leitor deve fazer para alcançar seu objetivo com segurança e livre de erros. Sabendo que as relações gera e habilita contêm as instruções propriamente ditas, conclui-se que as sentenças que possuem tais relações tem maior chance de conter candidatos a termos.

O método de extração de termos sensível ao gênero de instruções é dividido em cinco etapas, detalhadas a seguir. A Figura 1 ilustra todo o processo de extração.

ETAPA 1: A primeira etapa do método de extração é o pré-processamento dos textos[4] pelo parser PALAVRAS (Bick, 2000) para a extração do conhecimento morfossintático dos textos. O conhecimento extraído do parser é utilizado para a identificação das relações gera e habilita.

ETAPA 2: O arquivo resultante do pré-processamento pelo parser PALAVRAS é então usado como entrada pela etapa de extração na qual este arquivo é percorrido e são reconhecidas as sentenças de acordo com a marcação feita pelo parser PALAVRAS.

3 http://www.linguateca.pt/HAREM/

4 Neste trabalho, utilizamos as seções de instalação e instrução, pois nelas a frequência das relações gera e habilita é maior do que nas outras seções.

Assim que uma sentença é reconhecida, a mesma é submetida a uma busca por uma das formas gramaticais descritas por Moura (2008) e Delin *et. al*. (1994). As formas gramaticais usadas para a relação gera são as seguintes: "**Para + Infinitivo**", "**Se + Subjuntivo**" e "**Para + Frases**". Para a relação habilita, foram usadas as seguintes formas gramaticais: "**Sequencia**", "**condição Antes**" e "**condição Depois**".

ETAPA 3: Assim que uma relação gera ou habilita é identificada, a sentença então é filtrada por uma lista de padrões morfossintáticos, gerando a lista final de candidatos a termos. Os padrões morfossintáticos para extração de termos compostos são os mesmos usados por Baségio (Baségio, 2006 apud Ribeiro Jr, 2008), por exemplo, "substantivo adjetivo" e "substantivo preposição substantivo". Para o método de extração implementado neste projeto, o maior tamanho de termo composto aceito é três, isto é, são gerados uni, bi e trigramas. Para termos simples, são utilizados os mesmos padrões de Teline (2004) que extrai substantivos, adjetivos e verbos.

ETAPA 4: Nesta etapa, foram removidos da lista final de candidatos a termos todos aqueles n-gramas que estão presentes na *stoplist* que foi compilada para o projeto e-Termos[5] com algumas adaptações para atender o gênero de manuais de instruções.

ETAPA 5: Nesta última etapa, a lista de candidatos a termos é ranqueada pelo C-value (Frantzy and Ananiadou, 1997). O C-value é uma medida estatística usada para extração de termos multi-palavras. De posse dessa medida, o sistema faz o raqueamento dos candidatos a termos.

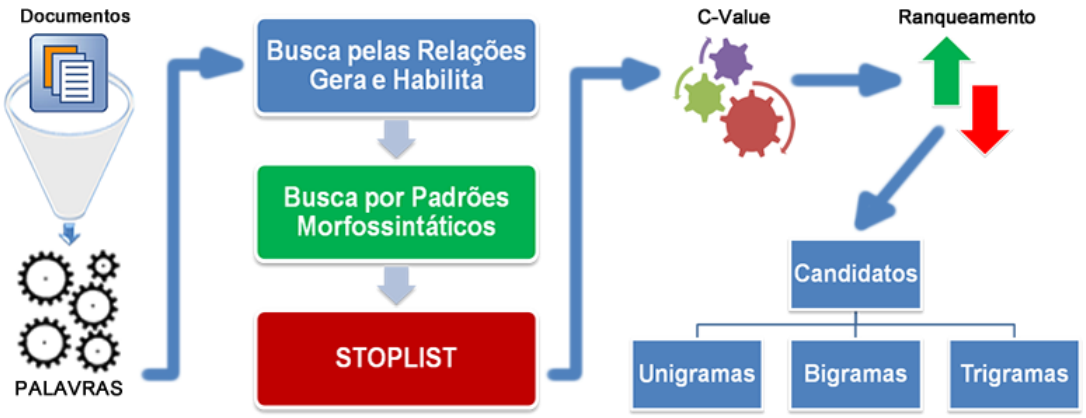

Figura 1. Processo de Extração de Termos para Textos Instrucionais

3.2. Aplicação da Extração de Termos em Sistemas de Adaptação Textual

O SIMPLIFICA é um sistema de autoria para apoiar a produção de textos simplificados, em que textos originais recebem simplificação automática com possível pós-edição pelos autores. Foi customizado para simplificar cada fenômeno sintático descrito em (Specia *et al.*, 2008) e mais o tratamento de adjuntos adverbiais longos, para atender as necessidades de alfabetizados de nível rudimentar e básico (Gasperin *et. al.*, 2010). Possui também o módulo de Simplificação Léxica e um avaliador automático da complexidade de textos (Aluisio *et. al*., 2010), baseado nas classes de alfabetização do INAF (2009), isto é, classifica um texto de acordo com três classes: adequado a

5 http://www.etermos.cnptia.embrapa.br/

alfabetizados em nível pleno, básico e rudimentar e pode indicar a necessidade de mais simplificação léxica ou sintática, fornecendo a possibilidade de simplificar via um processo cíclico que pode ser avaliado a cada momento que o usuário autor desejar.

A primeira etapa da simplificação léxica é a marcação das palavras consideradas complexas. Para isso, são usados três dicionários que foram desenvolvidos para o Projeto PorSimples. Um dicionário é composto por palavras comuns para crianças, outro composto por palavras frequentes e outro composto por palavras concretas. O sistema percorre o texto e verifica cada palavra. Se a palavra não está no dicionário de palavras simples e também não é um nome próprio, então ela é considerada uma palavra complexa. Para fazer a busca da palavra no dicionário de palavras simples, a mesma precisa estar lematizada e para isso foi utilizado o dicionário Unitex-PB[6] para descobrir o lema de cada palavra.

A busca por um lema em um dicionário pode se tornar um problema quando processamos palavras ambíguas, como por exemplo, a palavra "canto", que possui dois significados: o ponto ou área em que linhas e superfícies se encontram e formam um ângulo, ou então pode ser o verbo "cantar" conjugado na primeira pessoa do singular do presente do Indicativo. Para tratar essas ambiguidades eventualmente encontradas nos textos, o sistema utiliza o etiquetador gramatical MXPOST POS tagger (Ratnaparkhi, 1996) treinado com o conjunto de etiquetas NILC tagset[7], que tem como objetivo etiquetar o texto identificando a categoria gramatical de cada palavra automaticamente.

Após a etiquetação das palavras, todas aquelas que não são substantivos, preposições ou numerais são selecionadas e então a busca no dicionário de palavras simples é feita utilizando a informação da categoria gramatical de cada palavra. Como se trata de uma operação automática, ela não é totalmente precisa e algumas palavras podem não constar no dicionário. Além disso, se o sistema não foi capaz de categorizar gramaticalmente a palavra, a busca no dicionário é feita utilizando apenas o lema e mesmo que a palavra não seja encontrada ela é marcada como complexa. Em seguida, todas as palavras que foram marcadas são associadas com sinônimos mais simples. Para isso, dois recursos são utilizados: o tesauro TeP 2.0[8] e o PAPEL[9]. Esse procedimento é executado quando o usuário clica na palavra marcada, disparando uma consulta por sinônimos no dicionário de palavras simples.

Sabendo-se que em manuais de instrução os termos técnicos não podem sofrer substituição, pois a substituição de determinado termo pode alterar a interpretação de determinada instrução no manual e, em razão disso, causar acidentes ou prejuízos, foram implementadas algumas adaptações no método de simplificação léxica para atender o gênero de manuais de instruções.

Sabendo-se dos resultados positivos que a utilização de elaborações textuais traz para a compreensão do texto (Watanabe *et. al.*, 2010), decidiu-se aplicar esta estratégia para elaborar os termos técnicos durante a simplificação léxica de manuais de instrução. Desta forma, estes receberam uma definição adicional, com o uso da primeira oração da Wikipédia que relata o termo.

---

6 http://www.nilc.icmc.usp.br/nilc/projects/unitexpb/web/dicionarios.html
7 http://www.nilc.icmc.usp.br/nilc/TagSet/ManualEtiquetagem.htm
8 http://www.nilc.icmc.usp.br/tep2/
9 http://www.linguateca.pt/PAPEL/

O primeiro passo na adaptação da simplificação léxica foi a geração de um dicionário a partir da lista de termos extraídos e seu respectivo C-Value pelo NorMan. Este dicionário foi então inserido como um novo dicionário de palavras simples no sistema de simplificação léxica. Dessa forma, o sistema passa a considerar os termos técnicos como palavras simples, evitando que os mesmos sejam marcados como palavras complexas, sendo estes associados com definições da Wikipédia, sempre que houver uma, que podem ser usadas como explicações acompanhando os termos.

Tendo em vista que o método de extração de termos sensível ao gênero de manuais de instruções extrai unigramas, bigramas e trigramas, o sistema foi adaptado para percorrer o texto em três etapas e não apenas uma vez. Na primeira etapa, o sistema percorre o texto formando trigramas e verificando cada um deles no dicionário gerado pelo NorMan. Caso haja intersecção entre dois trigramas, aquele com o maior C-Value é escolhido para ser elaborado por uma definição da Wikipédia. Na segunda etapa, o sistema percorre o texto novamente formando bigramas e verificando cada um deles. Caso haja intersecção entre dois bigramas, vence aquele com maior C-Value, assim como ocorre com os trigramas, porém caso haja intersecção com trigramas, o trigrama tem preferência. Por último, o texto é percorrido da forma tradicional, buscando cada palavra no dicionário e caso haja intersecção com bigramas ou trigramas, os bigramas e trigramas têm prioridade sobre unigramas.

Após identificar todos os termos técnicos do dicionário do NorMan no texto, tais termos são marcados para receber informações adicionais. Quando o usuário clica em um termo técnico marcado, uma consulta é feita na Wikipédia e é apresentado para o usuário um link como uma inf**o**rmação adicional (Figura 2), contendo as duas primeiras orações da Wikipédia para aquele termo técnico.

Figura 2. Trecho de manual com termos técnicos marcados para receber informações adicionais da Wikipédia

4. Avaliação Intrínseca e Extrínseca

A avaliação intrínseca foi realizada através da comparação dos resultados da extração de termos feita pelo NorMan com outros quatros métodos que foram descritos no trabalho de (Teline, 2004): o método linguístico, estatístico e dois métodos híbridos, sendo que o Híbrido-A utilizou indicadores estruturais no texto para ajudar na identificação de termos e o método Híbrido-B não utilizou estes indicadores. Essa avaliação teve como objetivo verificar se o método NorMan trazia resultados diferentes dos demais métodos, o que significaria que o método trouxe novidades. O corpus utilizado nesta avaliação

continha 111.164 palavras e era composto pelas seções de "instruções" e "instalação" de 50 manuais de produtos tecnológicos. A comparação entre os métodos foi feita através do cálculo da intersecção dos candidatos a termos extraídos pelos métodos devido a falta de uma lista de referência para o cálculo de precisão, cobertura e medida F. Também foi calculado a similaridade entre os métodos através do cálculo do coeficiente Jaccard (J), que é uma estatística usada para comparar a similaridade e diversidade de conjuntos de amostra.

Para unigramas, a intersecção entre o NorMan e os quatros métodos foram de 71,6% para método linguístico (J=0,44), 77,3% para o método estatístico (J=0,52), 42,4% para o método Híbrido-A (J=0,37) e 51,8% para o Híbrido-B (J=0,44). Na extração de bigramas, o método NorMan teve uma intersecção de 42,5% com o método Linguístico (J=0,15), 36,3% com o método estatístico (J=0,03), 10,8% com o método Híbrido-A (J=0,09) e 12,6% com o Híbrido B (J=0,10). Na extração de trigramas, o método NorMan teve uma intersecção de 46,9% com o método Linguístico (J=0,17), 22,2% com o método estatístico (J=0,03), 10,6% com o método Híbrido-A (J=0,09) e 15,1% com o Híbrido-B (J=0,13). Estes resultados demonstram que, embora na extração de unigramas tenha havido intersecção de até 77,3%, para bigramas e trigramas a intersecção foi bem mais baixa, demonstrando que o método trouxe novidades, já que foi capaz de recuperar termos diferentes dos métodos de extração usados com textos científicos. O cálculo do coeficiente de similaridade (J) também demonstrou que apesar de a similaridade entre os métodos de unigramas ter variado entre 0,37 e 0,44, para bigramas e trigramas a similaridade entre os métodos foi bem mais baixa, variando de 0,03 a 0,15, reforçando que o método trouxe novidades.

Também foi feita uma avaliação do método através do cálculo da estatística Kappa, que é comumente utilizada para medir a concordância entre diferentes anotadores em tarefas de PLN. Nesta avaliação, o corpus foi limitado a somente as seções de "instalação" e "instruções" dos manuais de televisores. Essa escolha se deu pelo fato de o produto ser extremamente popular e, portanto, facilitaria a tarefa dos avaliadores. Não houve recursos financeiros disponíveis para a contratação de técnico especialista em TVs, portanto os avaliadores foram membros do próprio laboratório onde a pesquisa foi realizada. Foram selecionados os 150 primeiros termos extraídos da lista de uni, bi e trigramas. A lista de unigramas foi avaliada por três avaliadores diferentes, sendo que o avaliador 1 identificou 104 (69,33%) termos entre os 150 candidatos da lista de unigramas. O avaliador 2 identificou 126 (84%) e o avaliador 3 identificou 105 (70%). O valor da medida Kappa resultante dessa avaliação foi de 0.507. Este valor representa que houve um índice moderado de concordância entre os avaliadores. A lista de bigramas foi avaliada por dois avaliadores diferentes, sendo que o Avaliador 4 identificou 91 (60,66%) termos entre 150 candidatos da lista. O avaliador 5 identificou 75 (50%) termos. O valor da medida Kappa resultante dessa avaliação foi de 0.069. Este valor representa um índice de concordância pequeno/fraco. A lista de trigramas foi avaliada por outros dois avaliadores, sendo que o avaliador 6 identificou 99 (66) termos entre 150 candidatos da lista. O avaliador 7 identificou 116 (77,3%) termos. O valor da medida Kappa resultante dessa avaliação foi de 0.061. Este valor representa um índice de concordância pequeno/fraco. Os valores de concordância encontrados nesta etapa da avaliação podem ser um reflexo de que os avaliadores não são especialistas e, além disso, evidencia a dificuldade da avaliação manual da tarefa de extração automática de termos neste gênero de textos.

A avaliação extrínseca tinha como objetivo verificar se adaptação da simplificação léxica para o gênero de manuais de instrução ajudaria pessoas que trabalham em funções técnicas a entender melhor um manual de instrução. Esta avaliação foi feita através da submissão de um trecho de manual de instrução de televisão no sistema SIMPLIFICA já adaptado para simplificação léxica de manuais de instruções. O trecho escolhido para avaliação continha 16 termos distintos marcados como termos para elaboração léxica. Dentre estes 16 termos, 12 foram reconhecidos pelo sistema como termo técnico, pois pertencia às listas de candidatos a termo geradas pelo método NorMan e os quatro restantes, como palavras complexas. 15 voluntários participaram da avaliação, que consistia na leitura do trecho do manual e sua interação com os termos marcados. Os voluntários eram funcionários da Prefeitura do Campus da USP em São Carlos, com faixa etária entre 33 e 68 anos, escolaridade entre ensino fundamental incompleto e pós-graduação (um único funcionário) e com funções que podem ser divididas entre auxiliares e técnicos. Após a leitura, os voluntários eram questionados sobre o conhecimento prévio de cada termo marcado e a utilidade das informações extras fornecidas pela Wikipédia. Por fim, cada voluntário respondeu 16 questões de múltipla escolha, referente a cada termo destacado no trecho. A contabilização das respostas (Figura 3) mostrou que os voluntários responderam que já conheciam 67,7% dos termos técnicos e 100% das palavras complexas. Os voluntários consideraram 62,19% das informações da Wikipédia como sendo úteis contra apenas 6,6% das palavras complexas. Este valor de apenas 6,6% de utilidade nas palavras complexas é reflexo da qualidade do recurso léxico disponível, pois os sinônimos apresentados como opções de substituição fugiam do contexto do texto e, portanto, não foram considerados úteis. A média de acerto no questionário foi de 86,6% tanto para termos técnicos como para palavras complexas. O fato de os voluntários responderem que tinham conhecimento prévio sobre 67,7% dos termos técnicos e mesmo assim considerarem 62,19% das informações da Wikipédia como sendo úteis demonstra que o sistema, de fato, atingiu seu objetivo, que era auxiliar os voluntários a compreenderem melhor o manual de instrução.

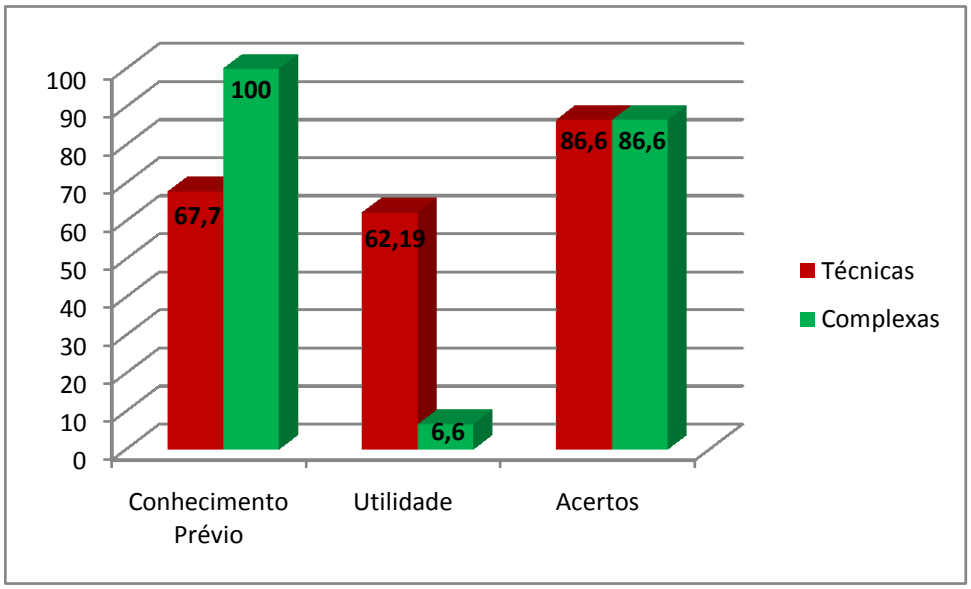

Figura 3. Comparação de conhecimento prévio, utilidade e acertos entre as palavras complexas e os termos técnicos

Foram feitas análises de correlações para verificar a existência de relações entre duas ou mais variáveis através do cálculo estatístico qui-quadrado. Estes cálculos mostraram que o cargo exercido voluntário teve influência no índice de utilidade das informações trazidas pela Wikipédia e dicionário de sinônimos (qui-quadrado = 9.3548, 0.01>P>0.001, Grau Liberdade (G.L.) = 1). Além disso, o conhecimento prévio dos termos técnicos estava correlacionado ao nível de acerto dos termos técnicos sinônimos (qui-quadrado = 4.0074, 0.05>P>0.02, G.L. = 1). Também foi verificado que a idade (qui-quadrado = 6.4071, 0.02>P>0.01, G.L. = 1) e o nível de escolaridade (qui-quadrado = 25.3215, P<0.001, G.L. = 1) influenciaram o nível de acerto e a idade também influenciou a utilidade (qui-quadrado = 4.0625, 0.05>P>0.02, G.L. = 1).

5. Considerações Finais

O objetivo principal do projeto NorMan foi o desenvolvimento de um método de extração de termos de manuais de instruções de produtos tecnológicos que pudesse extrair candidatos a partir de um conjunto pequeno de textos. Este cenário difere da aplicação da extração de termos de textos científicos na qual o tamanho do corpus é bem grande, dada a disponibilidade deste gênero textual. Como estudo de caso da aplicação do método, foi adaptado o sistema de autoria de textos simplificados SIMPLIFICA, para realizar a extração de termos de manuais técnicos e apresentar uma definição extraída da Wikipédia para os termos. As avaliações intrínsecas reportaram que o sistema produz resultados distintos dos métodos apresentados na literatura científica, recuperando termos específicos para o contexto, que seriam negligenciados pelos métodos propostos para o gênero de textos científico.

Pela avaliação extrínseca pode ser observado o aumento de reconhecimento de utilidade dos mecanismos de elaboração textual apresentado para os termos técnicos em comparação com as palavras complexas, ou seja, a informação textual elaborada (extraída da Wikipédia) apresentou maior utilidade pelos usuários. No teste realizado com os usuários também foi reportado um aumento de entendimento adquirido a respeito dos termos técnicos, depois de se interagir com os mecanismos que apresentavam a definição extraída da Wikipédia. Esses resultados destacam que a extração de termos, combinada com a elaboração textual, impactou positivamente na compreensão dos textos pelos usuários do experimento. Como trabalhos futuros, indicamos a adaptação sintática para o mesmo contexto do sistema (manuais técnicos), criação de listas de referências de termos do gênero de manuais de instruções para produtos tecnológicos e aprofundamento do estudo de uso de recursos de linguísticos, como a Wikipédia, disponíveis livremente para criação de outros métodos de elaboração textual.

Referências

Aluisio, S. M. and Gasperin, C. (2010). Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplification of Portuguese Texts. In the Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, June, 2010, Los Angeles, California, Association for Computational Linguistics, p. 46-53.

Aluísio, S. M.; Specia, L.; Gasperin, C.; Scarton, C. E. (2010). Readability Assessment for Text Simplification. In: NAACL 5th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-2010), 2010, Los Angeles.

Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications. New York : ACL, 2010. v. 1. p. 1-9.

Aouladomar, F. (2005). Towards Answering Procedural Questions, Workshop KRAQ05, IJCAI 05, Edinburgh, p. 21-30. Disponível em: http://www.irit.fr/recherches/ILPL/kraq05V1.pdf

Bick, E. (2000). The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Tese (Doutorado) – Arhus University, 2000.

Burstein, J., Shore, J., Sabatini, J., Lee, Y.W., Ventura, M. (2007). The automated text adaptation tool. In NAACL '07: Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations on XX, p. 3-4.

Burstein, J. (2009). Opportunities for Natural Language Processing Research in Education. In the Proceedings of CICLing, p. 6-27.

Delin, J.; Hartley, A.; Paris, C., Scott, D.; Vander Linden, K. (1994). Expressing Procedural Relationships in Multilingual Instructions, Proceedings of the Seventh International Workshop on Natural Language Generation, p. 61-70.

Devlin, S. and Unthank, G. (2006). Helping aphasic people process online information. Proceedings of ASSETS, 2006, p. 225-226.

van der Eijk, P. (1997). Controlled languages in technical documentation, Proceedings of the Computational Linguistics in the Netherlands, 1997.

Elhadad, N. (2006). Comprehending technical texts: predicting and defining unfamiliar terms. AMIA Annu Symp Proc. 2006:239-43.

Fontan, L.; Saint-Dizier, P. (2008). Analyzing the explanation structure of procedural texts: dealing with Advices and Warnings. In: International Symposium on Text Semantics (STEP 2008), Venise, Johan Bos (Ed.), Association for Computational Linguistics (ACL) p. 115-127.

Frantzy, K. T.; Ananiadou, S. (1997). Automatic Term Recognition using Contextual Cues. Manchester Metropolitan University. Third Delos Workshop Cross-Language Information Retrieval Zurich, 5-7 March 1997. Disponível em: http://www.ercim.eu/publication/ws-proceedings/DELOS3/Frantzi.pdf

Gasperin, C. Maziero, E. and Aluísio, S.M. (2010). Challenging Choices for Text Simplification, In: Proceedings of PROPOR 2010, p. 40-50, António Branco, Aldebaro Klautau, Renata Vieira, Vera Lúcia Strube de Lima (Eds.): Computational Processing of the Portuguese Language, Springer 2010, v. 6001. p. 40-50.

Goldman, A. (1970). A Theory of Human Action, Prentice-Hall, 230 pp.

INAF (2009). Instituto P. Montenegro e Ação Educativa. INAF Brasil - Indicador de Alfabetismo Funcional - 2009. Disponível em: http://www. ibope. com.br/ipm/relatorios/relatorio_inaf_2009.pdf

Lopes, L.; Vieira, R.; Finatto, M. J.; Martins, D. (2010). Extracting compound terms from domain corpora. Journal of the Brazilian Computer Society (Impresso), p. 1-13, 2010.

Moura, A. B. N. (2008). Tipologia Textual e Ativação de Terminologia: um Estudo em Manuais Técnicos de Produtos Tecnológicos. Tese de Doutorado em Letras - Universidade Federal do Rio Grande do Sul, p. 287.

Paris, C.; Scott, D. (1994). Stylistic variation in multilingual instructions. *In Proceedings of the Seventh International Workshop on Natural Language Generation*, Kennebunkport, MN, 21--24 June 1994, pages 45--52.

Paris, C.; Vander Linden, K., Fischer, M.; Hartley, A.; Pemberton, L.; Power, R.; Scott, D. (1995). A support tool for writing multilingual instructions. *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, p. 1398--1404, Montreal, Canada, 1995.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In Brill, E. & Church, K., editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 133–142. Nova Jérsei, EUA.

Ribeiro Jr., L. C. (2008). OntoLP : Construção Semi-Automática de Ontologias a partir de Textos da Língua Portuguesa. Dissertação (Mestrado) - Programa de Pós-Graduação em Computação Aplicada, Universidade do Vale do Rio dos Sinos, 2008.

Siddharthan, A. (2003): Syntactic Simplification and Text Cohesion. PhD thesis, University of Cambridge (2003).

Specia, L.; Aluísio, S. M.; Pardo, T. A. S. (2008). Manual de Simplificação Sintática para o Português. Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional (NILC-TR-08-06), 27 p., Junho 2008, São Carlos-SP. Disponível em: http://caravelas.icmc.usp.br/wiki/index.php/Publications.

Teline, M. F. (2004). Avaliação de métodos para extração automática de terminologia de textos em português. ICMC-USP, São Carlos, 2004. Dissertação de Mestrado.

Urano, K. (2000): Lexical Simplification and Elaboration: Sentence comprehension and incidental vocabulary acquisition. Unpublished master's thesis, University of Hawai'i at Manoa, Honolulu (2000). Disponível em: http://www.urano-ken.com/research/thesis.pdf

Watanabe, W.M., Candido Jr, A., Amancio, M.A., Oliveira, M., Pardo, T.A.S., Fortes, R.P.M., Aluísio, S.M. (2010). Adapting web content for low-literacy readers by using lexical elaboration and named entities labeling. In the Proceedings of the W4A-7th International Cross-Disciplinary Conference on Web Accessibility 2010, (2010). Nova York: ACM Press, v. 1, 1--9 (2010)

Young, D.N. (1999). Linguistic simplification of SL reading material: Effective Instructional Practice? The Modern Language Journal, 83(3), 350--366 (1999)

Rahimi, M. Y. (2011). Use of Syntactic Elaboration Techniques to Enhance Comprehensibility of EST Texts. English Language Teaching, Vol. 4, No. 1, 11-17.