Patricia Nunes Gonçalves

CorrefSum: Revisão de Coesão Referencial em Sumários Extrativos

São Leopoldo

2008

Patricia Nunes Gonçalves

CorrefSum: Revisão de Coesão Referencial em Sumários Extrativos

Dissertação submetida a avaliação como requisito parcial para a obtenção do grau de Mestre em Computação Aplicada.

Orientador:
Renata Vieira

UNIVERSIDADE DO VALE DO RIO DOS SINOS
CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA INTERDISCIPLINAR DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

São Leopoldo

2008

Dedicatória

Dedico este trabalho:
aos meus pais, Hilário (in memorian) e Nina.

Agradecimentos

Agradeço primeiramente a Deus por ter-me guiado no desenvolvimento deste trabalho.

A minha orientadora, Professora Dra. Renata Vieira, pelo incentivo, ânimo, paciência, amizade, apoio e orientação, que foram imprescindíveis para a conclusão desta dissertação.

A minha co-orientadora, Professora Dra. Lucia Rino, pela recepção durante a visita a Universidade Federal de São Carlos, pela amizade e incentivo durante essa jornada.

Aos meus pais Hilário e Nina, por me ensinarem a viver com paixão e alegria. Por me incentivarem a vida toda a perseguir meus sonhos. Ao meu marido, Rodrigo pelo apoio e incentivo. Ao meu irmão Nilton e demais familiares pelo carinho.

Aos amigos e colegas do LEL, Sandrinha, Zeca, César, Jonatan, Mirian e especialmente para Patricia Pizzinato, minha fiel amiga que me ajudou muito nesta etapa final.

Ao NILC em especial aos professores Thiago Pardo, Graça Nunes, por me receberem de forma tão carinhosa no laboratório e por me fazer sentir em casa, mesmo estando tão longe.

Aos amigos de São Carlos que tive a oportunidade de conhecer, Ariane, Thiago Carbonel, Jorge, Gawa, especialmente para Élen que acabou se tornando uma amiga inseparável (apesar da distância) nos últimos meses.

Aos amigos que me apoaiaram nos momentos difíceis, Cristiane, Marcio, Rafael, Mara, Daniela, Cristiano, Nataniel, Leandro, Francisco, Luciana, Paula, Marcos Garcia, Luiz Acauã e Vivi.

E é claro, os colegas inseparáveis no mestrado Cony, Roberto, Luciano, Paulo, Pessin, Sérgio e Andressa, em especial para Luiz Carlos pela amizade e parceria nas viagens, festas, estudos e congressos.

A Capes pelo apoio financeiro e ao projeto Farol Procad-Capes por me proporcionar realizar um intercâmbio durante a realização deste trabalho.

Resumo

Com o avanço da Internet, cada vez mais convivemos com a sobrecarga de informação. É nesse contexto que a área de sumarização automática de textos tem se tornado uma área proeminente de pesquisa. A sumarização é o processo de discernir as informações mais importantes dos textos para produzir uma versão resumida. Sumarizadores extrativos escolhem as sentenças mais relevantes do texto e as reagrupam para formar o sumário. Muitas vezes, as frases selecionadas do texto não preservam a coesão referencial necessária para o entendimento do texto. O foco deste trabalho é, portanto, na análise e recuperação da coesão referencial desses sumários. O objetivo é desenvolver um sistema que realiza a manutenção da coesão referencial dos sumários extrativos usando como fonte de informação as cadeias de correferência presentes no texto-fonte. Para experimentos e avaliação dos resultados foram utilizados dois sumarizadores: GistSumm e SuPor-2. Foram utilizadas duas formas de avaliação: automática e subjetiva. Os resultados mostram o potencial dessa abordagem e indicam maneiras de avançar nesta pesquisa.

Palavras-chave: Processamento de Língua Natural, Sumarização Automática, Cadeias de Correferência, Coerência e Coesão Textual.

Abstract

With the advance of Internet technology we see the problem of information overload. In this context, automatic summarization is an important research area. Summarization is the process of identifying the most relevant information brought about in a text and on that basis to rewrite a short version of it. Extractive summarizers choose the most relevant sentences in a text and regroup them to form the summary. Usually the juxtaposition of the selected sentences violate the referential cohesion that is needed for the interpretation of the text. This work focuses on the analysis and recovery of referential cohesion of extractive summaries on the basis of knowledge about correference chains as presented in the source text. Some experiments were undertaken considering the summarizers GistSumm and SuPor-2. Evaluation was done in two ways, automatically and subjectively. The results indicate that this is a promising area of work and ways of advancing in this research are discussed.

Keywords: Natural Language Processing, Automatic Summarization, Coreference Chains, Coherence and Textual Cohesion.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ART	Anaphor Resolution Tool
DMSUMM	Discourse Modeling Summarizer
HTML	Hipertext Markup Language
ML	Machine Learning
MMAX	Multi-Modal Annotation in XML
NILC	Núcleo Interinstitucional de Lingüística Computacional
PLN	Processamento de Linguagem Natural
POS	Part-of-Speech
RA	Resolução de Anáforas
RI	Recuperação de Informação
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RST	Rhethorical Structure Theory
SA	Sumarização Automática
TFISF	Term Frequency Inverse Sentence Frequency
XML	eXtensible Markup Language

Sumário

Sumário

Capítulo 1

Introdução

1.1 Contextualização

Atualmente, com o advento da Internet, a informação está disponível de maneira rápida e em grande quantidade, ocasionando a sobrecarga de informação. Com isso, a necessidade de filtrar e discernir informação de maior relevância tem se tornado cada vez maior.

Conforme Pardo (PARDO, 2005a), é nesse contexto que a área de sumarização automática de textos tem se tornado uma área proeminente. Com o avanço da Internet, onde as pessoas se vêem em um mar de informação em constante expansão e atualização, um grande interesse acadêmico, comercial e governamental surgiu por essa área. A idéia de produzir um texto contendo apenas as informações centrais, a partir de um texto mais elaborado se harmoniza perfeitamente com esta tendência global.

A sumarização é uma atividade bastante comum. Quando uma pessoa narra um evento geralmente se utiliza de um resumo. Inconscientemente as pessoas estão sempre sumarizando. É também muito comum encontrar resumos na forma escrita, como previsões meteorológicas, chamadas em jornais e revistas, resenhas e *abstracts* de livros e teses.

A sumarização é o processo de seleção de informações mais importantes de um texto, que nesta área chamamos de texto-fonte (PARDO, 2002). Na área de sumarização automática existem duas principais abordagens, a superficial e a profunda, as quais caracterizam métodos distintos de sumarização automática. A abordagem superficial utiliza, sobretudo, métodos experimentais e estatísticos, enquanto a profunda está relacionada a teorias formais e lingüísticas.

O foco deste trabalho está relacionado à qualidade dos sumários extrativos, gerados pela abordagem superficial. Essa abordagem utiliza a escolha de sentenças de maior relevância do texto para compor o sumário.

A sucessão de palavras em um texto formam uma cadeia que vai muito além da simples seqüencialidade, pois deve existir um entrelaçamento significativo que aproxima as partes formadoras do texto. Uns dos mecanismos que estabelecem a conectividade e a coesão de um texto são os referentes textuais. Cada palavra escrita estabelece relações de sentido e significado tanto entre os elementos que a antecedem como os que o sucedem construindo uma cadeia textual significativa (KOCH, 2003). A coesão em um texto traz uma relação de unidade demonstrando que o texto trata de um assunto principal. É comum sumários extrativos não preservarem a coesão referencial original necessária para o entendimento do texto.

Conforme (KOCH; TRAVAGLIA, 1996), coesão referencial é um componente da superfície do texto que faz remissão a outro(s) elemento(s) nela presentes ou inferíveis a partir do universo textual. A construção de cadeias de correferência é parte do processo de estruturação coesa de um texto.

Uma cadeia de correferência é o conjunto de todas as menções a uma determinada entidade encontradas no texto. Por exemplo, observe o texto na Figura 1.

Neste exemplo, temos a cadeia de correferência formada pelas palavras: “o agrônomo Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina)”, “Guerra”e “o agrônomo”. A hipótese considerada neste trabalho é de que a coesão refe-

A discussão sobre a biotecnologia nacional está enviesada, pois está sendo entendida como sinônimo de transgenia. A opinião é d**o agrônomo Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina)**. **Guerra** citou a micropropagação de vegetais (produção de mudas em laboratório, feita para evitar doenças e selecionar vegetais saudáveis) como exemplo de biotecnologia de baixo custo. Com ela, aumentou-se a produção de moranguinho, no sul do país, de 3,2 kg para 60 kg por hectare. Para **o agrônomo**, o Brasil deve buscar o desenvolvimento de transgenias que tentem melhorar as condições da agricultura local, como o cultivo de plantas com a capacidade de captar certos elementos presentes na terra.

Figura 1: Exemplo de Cadeias de Correferência

rencial dos sumários extrativos pode ser melhorada se a informação a respeito da cadeia de correferência do texto-fonte for levada em consideração.

Suponha que o sumário extrativo resulte em: *"Para o agrônomo, o Brasil deve buscar o desenvolvimento de transgenias que tentem melhorar as condições da agricultura local, como o cultivo de plantas com a capacidade de captar certos elementos presentes na terra"* não seria possível o leitor recuperar o referente pretendido para a expressão "o agrônomo". Se a cadeia for levada em consideração ela poderia ser utilizada e a expressão "o agrônomo" poderia ser substituída por "o agrônomo Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina)".

Dado que a pesquisa na área de resolução de correferência tem se desenvolvido nos últimos anos, consideramos que a informação sobre as cadeias pode ser obtida por algoritmos desse tipo. Neste trabalho, no entanto, teremos à disposição um corpus com informações de cadeias anotados manualmente, contando com informações mais precisas sobre as cadeias.

Logo em seguida apresentamos os objetivos desse trabalho relacionados ao problema de coesão referencial em sumários extrativos automáticos, e a utilização da informação das cadeias como uma solução ao problema.

1.2 Objetivo do Trabalho

1.2.1 Objetivo Geral

Este trabalho tem como objetivo geral investigar e propor processos de enriquecimento de sumários extrativos procurando observar e melhorar a coesão referencial dos sumários através das informações das cadeias de correferência. Nesse sentido, será desenvolvida uma ferramenta que, a partir de cadeias de correferência, verifica problemas de coesão em sumários extrativos produzidos automaticamente e propõe correções.

1.2.2 Objetivos Específicos

- Estudar o fenômeno das cadeias de correferência nos textos da Língua Portuguesa com base em corpus anotado;
- Gerar sumários para os textos do corpus utilizando os sumarizadores extrativos GistSumm (PARDO, 2005a) e SuPor-2 (LEITE; RINO, 2006a);
- Estudar e implementar métodos de análise dos sumários extrativos para verificar coesão referencial com base na informação das cadeias de correferência;
- Projetar e implementar uma interface para a ferramenta, com opção de intervenção humana no processo de correção;
- Avaliar os sumários modificados com base na análise das cadeias de correferência.

1.3 Organização da Dissertação

Este trabalho está organizado da seguinte forma, o capítulo 2 apresenta uma introdução aos principais conceitos deste trabalho: coerência e coesão textual; sintagma nominal e suas classificações; sumarização automática.

O capítulo 3 apresenta uma visão geral dos materiais e métodos da pesquisa: corpus e ferramentas. São apresentadas as ferramentas PALAVRAS, MMAX, GistSumm, SuPor-2 e Rouge utilizadas nesta pesquisa.

No capítulo 4, o sistema CorrefSum, desenvolvido neste trabalho, é apresentado e cada um dos seus módulos detalhados. No capítulo 5, são apresentados os resultados dos experimentos realizados neste trabalho usando o GistSumm e o SuPor-2. Este capítulo apresenta, ainda, as avaliações que cercam esses experimentos. Foram realizadas duas diferentes avaliações: avaliação automática e avaliação subjetiva.

Por fim, no capítulo 6 são feitas as considerações finais da dissertação, apresentamos as contribuições e limitações deste trabalho.

Capítulo 2

Fundamentação Teórica

O objetivo deste capítulo é apresentar conceitos importantes relacionados a este trabalho, tais como os conceitos de coerência, coesão textual e cadeias de correferência. Uma visão geral da área de sumarização automática também será apresentada.

2.1 Coerência e Coesão

2.1.1 Coerência

A Coerência de um texto está apoiada na compreensão do sentido do texto. O sentido do texto deve resultar em uma compreensão global e não apenas superficial e local. Um texto compreendido apenas em parte, em geral, não é coerente. Observe o seguinte exemplo: *"João tinha terminado de estudar para a prova quando chegamos, mas ainda estava estudando"*. Nesse exemplo, observamos que, se a frase for lida por partes, ela faz sentido, mas, quando terminamos de ler, ela perde o sentido e, conseqüentemente, a coerência (KOCH; TRAVAGLIA, 1990).

Para que o texto seja coerente deve haver uma unidade de sentido no texto. A coerência ocorre na continuidade e linearidade de sentido entre as expressões do texto.

Um texto sem continuidade é considerado um amontoado de palavras e frases sem sentido e sem coerência.

Muitas vezes, para entender um texto, é necessário que o leitor ative seu conhecimento de mundo. Ao ativar esse conhecimento, a pessoa estabelece ligações não explícitas entre os termos componentes, fazendo com que o texto torne-se coerente.

A coerência de um texto também depende da sua boa formação em termos de interlocução comunicativa. O sentido do texto, geralmente, é compreendido no seu objetivo, pois o escritor utiliza um argumento básico na sua construção. Existe uma intenção por parte do escritor ao redigir um texto (KOCH, 2000).

Segundo (KOCH; TRAVAGLIA, 1990), a coerência subdivide-se em 4 grandes grupos: coerência semântica, coerência sintática, coerência estilística e coerência pragmática.

Coerência Semântica

A coerência semântica refere-se à relação entre significados das frases, levando em consideração a seqüencialidade em um determinado texto. Além disso, está presente nas relações de sentido entre os termos componentes de um texto, por exemplo, hiponímia e hiperonímia. Considere o seguinte exemplo como um caso de problema de coerência semântica: *"João possui um belo* ***veículo****. É um* ***cavalo*** *árabe puro sangue."* Nesse exemplo, temos os termos "cavalo" e "veículo", mas o termo "cavalo" não é um hipônimo de "veículo", tornando o texto incoerente.

Coerência Sintática

A coerência sintática refere-se aos meios sintáticos para expressar a coerência. Um exemplo disso é o uso de conectivos, de pronomes ou de sintagmas nominais definidos e indefinidos. Observe o seguinte exemplo: *"Maria foi ao baile, entretanto ele não fora*

convidada." Nesse exemplo, temos o emprego do pronome "ele" que poderia somente se referir ao substantivo "baile". Todavia, a palavra "baile", dentro de um senso comum, não pode concordar com ser "convidada", tornando, assim, a frase incoerente. Esse exemplo traz uma clara idéia de um problema de coerência sintática.

Coerência Estilística

A coerência estilística refere-se ao estilo lingüístico de escrita. Um exemplo de quebra de estilo lingüístico em um texto é o uso de gírias e termos inapropriados em artigos acadêmicos e textos formais. Por exemplo, *"Este artigo apresenta resultados preliminares, pois o treco ainda não está pronto"*. Nesse exemplo, considerando que a frase esteja num texto acadêmico formal, a palavra "treco" não deveria ser utilizada, pois o termo é inapropriado para o contexto e pode provocar estranhamento no leitor/interlocutor, que não espera o uso de um termo totalmente informal neste tipo de produção.

Coerência Pragmática

Este tipo de coerência é mais pertinente ao ato de fala do que ao texto escrito. A coerência pragmática diz respeito a uma seqüência comunicativa dada uma situação específica. Por exemplo, pessoa A pergunta: *"Você me empresta sua caneta?"* e pessoa B responde: *"Hoje comi chocolate o dia todo."* Podemos observar que, nesse exemplo, a resposta da pessoa B não tem relação de sentido com a pergunta realizada pela pessoa A, portanto essa seqüência, no ato de fala, torna-se incoerente. Se o foco fosse a escrita das frases, poderíamos afirmar que ambas estão semântica e sintaticamente corretas, dado o motivo deste tipo de coerência ser mais pertinente ao ato de fala do que ao texto escrito. Podemos observar que nesse exemplo a resposta da pessoa B não tem relação de sentido com a pergunta realizada pela pessoa A. Portanto essa seqüência no ato de fala torna-se incoerente.

Considerando os tópicos descritos nessa seção, os casos de maior relevância no contexto deste trabalho são os de coerência sintática e semântica, pois estão mais diretamente ligados à questão da correferência.

2.1.2 Coesão

Segundo (KOCH; TRAVAGLIA, 1996), a coesão de um texto é a relação que se estabelece entre os elementos do texto. Ela utiliza marcas lingüísticas, também chamadas de elementos coesivos, que se encontram conectados às palavras do texto permitindo uma seqüência linear. Esses elementos coesivos devem obedecer a uma ordem gramatical e são totalmente dependentes desta ordem.

A coesão está no nível superficial do texto e é imprescindível em um texto bem formado. É o primeiro passo para entender o sentido e subdivide-se em dois grandes grupos (KOCH; TRAVAGLIA, 1996): coesão referencial e coesão seqüencial.

Coesão Referencial

A coesão referencial estabelece a coesão entre dois ou mais elementos do texto, esses elementos remetem-se a um mesmo referente, isto é, um elemento do universo textual.

Segundo (JURAFSKY; MARTIN, 2000), existem duas formas de fazer remissão a um elemento: anáfora e catáfora.

- Anáfora: Faz remissão a um elemento já introduzido em um discurso. Por exemplo: *"**João** foi a festa. **Ele** se divertiu muito."* Nesse exemplo, temos o termo "Ele", que se refere a "João", que já havia sido introduzido no discurso. O termo "Ele" é anafórico em relação à "João".

- Catáfora: Traz um termo no texto que não consegue se resolver até que encontra

seu referente na seqüência do texto. Isto é, a remissão encontra-se à frente do referente. Por exemplo: *"**Ele** ainda não chegou, **meu irmão** sempre se atrasa."* Nesse exemplo, temos o termo "Ele", que se refere a "meu irmão", termo que só aparece na seqüência do texto.

Além da utilização de anáforas e catáforas, a coesão referencial (KOCH; TRAVAGLIA, 1996) ainda faz uso do mecanismo de reiteração, utilizando o emprego de sinônimos, meronímia, hiperonímia e nomes genéricos, conforme exemplos a seguir.

- Emprego de sinônimos: *"**Um garoto** estava correndo. **O menino** estava apavorado"*. Nesse exemplo "Um garoto" e "O menino" são sinônimos.
- Meronímia: *"**O carro** roubado foi encontrado. **Os pneus** não estavam no veículo."* Nesse exemplo "Os pneus" fazem parte do "carro".
- Hiperonímia: *"Dentre **os mamíferos** estudados para essa pesquisa[...]**A vaca** foi escolhida para a pesquisa."* Nesse exemplo, temos uma relação de hiperonímia entre "a vaca" e "os mamíferos".
- Nomes Genéricos: *"Todos ouviram o barulho da **moto**. Olharam para o fim da rua e viram **a coisa** chegando rápido."* Nesse exemplo, "a coisa" está substituindo "a moto" como um nome genérico.

Coesão Seqüencial

A coesão seqüencial diz respeito à progressão textual. Neste caso, existem elementos no texto que se unem para dar a idéia de seqüencialidade e continuidade da idéia central do texto.

Num texto coeso, suas partes são interdependentes, sendo elas de máxima importância para a compressão geral do texto. Chamamos isso de progressão textual.

A coesão seqüencial faz uso de dois procedimentos: coesão seqüencial por recorrência e coesão seqüencial por progressão:

- Coesão Seqüencial por Recorrência: é obtida pelos mecanismos de recorrência de termos e estruturas, de conteúdos semânticos e de recursos fonológicos (ritmo, rima e eco). Por exemplo: *"O homem nadava, nadava e nadava buscando salvar sua vida."*

- Coesão Seqüencial por Progressão: é utilizada para possibilitar manutenção temática e encadeamentos.

    1. A manutenção temática faz uso de termos com a mesma contigüidade semântica, por exemplo, *"O incêndio no edifício provocou sérios* ***acidentes****. Várias* ***ambulâncias*** *foram chamadas para realizar o atendimento às* ***vítimas*** *e transportá-las a um* ***hospital****."* Através dos termos que estão destacados neste exemplo, é possível que o leitor ative seu esquema cognitivo, desfazendo ambigüidades e avançando na perspectiva do texto.

    2. O encadeamento permite estabelecer relações semânticas entre orações, enunciados ou seqüências do texto. Por exemplo, *"****Primeiramente*** *trarei informações sobre coerência e coesão,* ***a seguir*** *falarei sobre sumarização automática e* ***finalmente*** *trarei informações sobre materiais e métodos que serão utilizados."* Nesse exemplo, temos os termos destacados realizando a função de ordenação e encadeamento numa determinada linha de tempo.

Dentre esses conceitos, os de coesão referencial e manutenção temática são os de maior relevância para este trabalho.

2.1.3 Relação entre Coerência e Coesão

Como visto anteriormente nesta seção, a coerência relaciona-se com a linearidade e sentido do texto, diferentemente da coesão que está na parte superficial do texto e utiliza mecanismos coesivos para realizar a conexão entre termos e frases. A coesão utiliza marcas explícitas no texto que são fáceis de identificar (KOCH; TRAVAGLIA, 1990).

A relação entre a coerência e a coesão existe porque a coerência é estabelecida a partir da seqüencialidade na leitura do texto e, por sua vez, a coesão fornece pistas para tornar o texto coerente utilizando os mecanismos coesivos. Portanto, a coesão é o ponto de partida para se estabelecer a coerência de um texto.

A interpretação entre a coerência e a coesão, conforme (KOCH, 2000), nos diz que a coerência está na profundidade de um texto e a coesão na parte superficial do mesmo.

Embora a coesão auxilie na compreensão do sentido do texto, ela não é suficiente para estabelecer a coerência. Entretanto, um texto sem coesão traz incoerências locais de fácil identificação.

Sumários extrativos (por eliminarem partes do texto) podem facilmente corromper a coesão de um texto e conseqüentemente sua coerência.

2.2 Correferência e Anáfora

Tradicionalmente, a anáfora se define como toda retomada de um elemento anterior em um texto, mantendo-se a identidade referencial. Quando uma entidade é mencionada pela primeira vez no texto, se faz o processo de `evocação` da entidade. Durante a leitura, na seqüência do texto, quando essa entidade é novamente mencionada, temos a realização do `acesso` a essa entidade. A expressão que faz o acesso é dita como `anafórica` e a expressão a quem ela se refere no texto é dita como seu `antecedente`. A relação entre essas duas expressões (anáfora e antecedente) é dita como `relação de correferência` (JURAFSKY; MARTIN, 2000). Na Figura 2 temos uma representação visual deste conceito.

Segundo (KOCH, 2003), as anáforas possuem um papel importante na construção da coerência de um texto. Não apenas na coerência, mas também na compreensão global e sentido do texto. Durante a leitura, ocorre o processamento textual, em que fica claro que existem representações de entidades no texto. A partir daí, o leitor faz uso do encadeamento referencial para resolver qual das entidades descritas no texto deve ser sele-

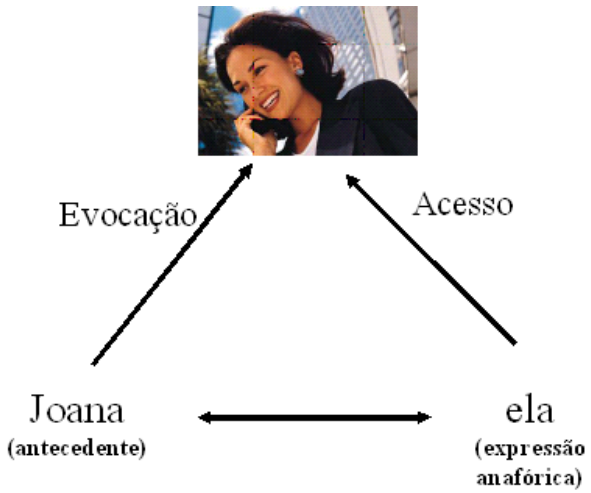

Figura 2: Relação de Correferência

cionada para interpretação do sentido do texto. Essas entidades são geralmente evocadas ou acessadas por sintagmas nominais.

A seguir apresentamos o conceito de sintagmas nominais, a sua classificação e as cadeias de correferência.

2.2.1 Sintagmas Nominais

Os sintagmas consistem num conjunto de elementos que representam uma unidade significativa dentro da oração e que mantém entre si relações de dependência e ordem. Os sintagmas costumam organizar-se em torno de um núcleo que, por si só, pode ser considerado um sintagma (KOCH; SILVA, 2002).

A natureza de um sintagma está na presença de seu núcleo. Se o núcleo for um verbo, o sintagma é verbal, se for um substantivo é um sintagma nominal. Além de substantivo, o sintagma nominal pode ainda apresentar núcleos como pronome pessoal, pronome demonstrativo, pronome indefinido, pronome interrogativo e pronome possessivo, logo abaixo temos exemplos com diferentes núcleos.

- Núcleo nome próprio: *"**William Eberhard** descobriu que as larvas provocam mudanças no comportamento da hospedeira."*

- Núcleo substantivo comum: *"**Pesquisas** em camundongos foram realizadas."*
- Pronome: *"**Ela** surgiu a partir de células isoladas da vaca Vitória."*

Geralmente, os sintagmas nominais são as expressões utilizadas para evocação e acesso de entidades mencionadas em um texto.

De acordo com (PERINI, 2003), o sintagma nominal pode se tornar uma estrutura bem complexa, pois pode apresentar grandes diferenças estruturais, como, por exemplo, apresentar determinante(s) e/ou modificador(es). Os determinantes antecedem o núcleo e os modificadores podem aparecer antes e depois do núcleo. Observe esses elementos em alguns exemplos de sintagmas nominais com uso de determinantes e modificadores (em destaque nos exemplos abaixo):

- Determinantes: O uso de determinantes é muito comum em sintagmas nominais, podem ser artigos definidos, indefinidos, adjetivos entre outros. *"**Os pingüins** são acostumados a mar aberto."*

- Modificadores
  - Pré-modificadores: aparecem antecedendo o núcleo. *"**O pequeno astro** vai passar a uma certa distância do Sol."*
  - Pós-modificadores: aparecem após o núcleo. *"**Amostras celulares de animais ameaçados de extinção** foram coletadas."*

2.2.2 Classificação dos Sintagmas Nominais

Os sintagmas nominais são expressões lingüísticas usadas para referenciar entidades mencionadas nos textos. Conforme trabalhos de (VIEIRA, 1998) e (COLLOVINI; VIEIRA, 2006b), os sintagmas nominais são divididos em 4 classes distintas: novas no

discurso, anáforas diretas, anáforas indiretas e associativas.

Novas no Discurso

Quando um sintagma nominal introduz um novo referente no discurso (evocação), sem apresentar parte de seu sentido, ancorado em uma expressão anterior, definimos esse sintagma como novo no discurso. As expressões dadas como novas no discurso não são anafóricas, já que são mencionadas pela primeira vez.

É muito comum que as expressões novas sejam mencionadas no início dos textos. Durante a seqüencialidade do discurso, outras expressões poderão ser utilizadas fazendo referência a uma entidade mencionada anteriormente. As expressões novas no discurso podem servir de antecedentes para as anáforas.

Anáfora Direta

A anáfora direta possui uma relação de identidade com seu antecedente e sua expressão lingüística apresenta o mesmo nome-núcleo do antecedente. Por exemplo: *"Um grupo que reúne 13 sociedades científicas nacionais enviou* ***uma carta*** *ao Senado Federal para pedir mudanças no projeto da nova Lei de Biossegurança. N****a carta*** *os cientistas falam sobre células-tronco."* Nesse exemplo, temos em destaque o sintagma nominal *"uma carta"*. Na segunda vez em que ele é mencionado, o termo torna-se anafórico direto, possuindo mesmo nome-núcleo mencionado anteriormente.

Anáfora Indireta

A anáfora indireta é também caracterizada pela relação de identidade com o antecedente, mas não possui o mesmo nome-núcleo. Vejamos o exemplo: *"Os EUA foi um*

dos últimos países a assinar ***a Declaração de Helsinque****.* ***O texto*** *traça diretrizes para ética em pesquisas...”*, nesse exemplo, o termo “O texto” está referindo-se ao “a Declaração de Helsinque”. Como podemos notar, eles não possuem o mesmo nome-núcleo, mas os dois termos referem-se à mesma entidade.

Segundo (TEIXEIRA, 2007), anáforas indiretas não são explicáveis por simples processos de associação, mas por complexos processos cognitivos. Como, por exemplo, processos inferenciais nos quais o leitor ativa a representação da informação armazenada em sua memória. A classe anafórica indireta possui, portanto, vários tipos. Abaixo, temos alguns exemplos:

- Relação de nome próprio e nome comum: Aplicação de um nome comum para referenciar um nome próprio dito anteriormente. Veja o exemplo: *“William Eberhard descobriu que as larvas d****a Hymenoepimecis*** *provocam mudanças no comportamento da hospedeira. A aranha modifica o formato da teia para que o casulo d****a vespa*** *possa se desenvolver.”*

- Relação de Sinonímia: Utilização de sinônimos para o mesmo referente. Por exemplo: *“Isso quer dizer que os* ***camundongos*** *transgênicos reduziram a gordura de seu corpo. Os* ***ratos*** *estudados[...]”*

- Nominalização de verbos: Aplicação de um substantivo para referenciar um verbo. Por exemplo: *“O presidente da Comissão Nacional de Ética em Pesquisa* ***propôs*** *na 52ª Reunião Anual da Sociedade Brasileira para o Progresso da Ciência....* ***A proposta*** *foi discutida pelos cientistas[...]”*

- Hiponímia/Hipernímia: Relações de hiponímia e hipernímia também são utilizadas para refereciar entidades já mencionadas. Por exemplo: *“As mudanças nas populações de* ***pingüins*** *também serviram como indicativo do problema climático.* ***Os animais*** *usavam geleiras para se abrigar e procriar.”*

- Numerais: Utilização de numerais como termo anafórico. Por exemplo: *“****As***

moléculas DM43 e a DM64 *parecem especificamente talhadas para neutralizar os principais efeitos do veneno das serpentes.* ***As duas*** *têm essa função antiofídica."*

- Relações pronominais: Inserção de pronomes para identificar um referente no texto. Esse caso é o mais comum de anáforas indiretas, pois o emprego de pronomes evita a repetição de um grupo nominal. Há algumas formas de pronominalização, vejamos alguns exemplos:

    1. Pronomes pessoais: *"Carlos Nobre participou do debate Cenários da Amazônia... a defesa feita por* ***ele*** *foi contra o desmatamento da floresta..."*
    2. Pronomes demonstrativos: *"Os dados preliminares sugerem que o animal pode tanto ter sido um placentário quanto um marsupial. Se* ***essa*** *hipótese for verdadeira, os pesquisadores..."*
    3. Pronomes possessivos: *"Benjamin Wolozin escreveu em* ***seu*** *artigo que obteve a primeira reconstrução de múltiplos genomas diretamente de uma amostra natural."*
    4. Pronomes indefinidos: *"As propostas foram discutidas entre pesquisadores e governo, entretanto* ***nenhuma*** *foi aceita."*

Anáforas Associativas

De acordo com (VIEIRA, 1998), a anáfora associativa introduz um novo referente no discurso. Entretanto, seu significado está fortemente ancorado em uma expressão anterior. A anáfora associativa pode ser de vários tipos, vejamos alguns exemplos:

- Relação conjunto/sub-conjunto: *"Adalberto Veríssimo, da ONG Imazon, apresentou estudo segundo o qual* ***as cidades em regiões amazônicas*** *ocupadas de forma predatória duram por volta de 23 anos. Ele citou como exemplo* ***as cidades de Paragominas (PA), Açailândia (MA) e Humaitá (AM)****."*

- Relação grupo/membros: *"Um tratamento para a obesidade que faz você perder peso e reduzir a taxa de gordura do corpo é o que sugere um estudo realizado por* ***um grupo de cientistas britânicos*** *será publicado hoje na revista Nature.* ***Um dos cientistas, John Clapham****, diz que esse é um alvo viável para remédios contra a obesidade."*

- Relação objeto/substância: *"****Uma estrela*** *é composta de* ***gás hidrogênio*** *condensada pela gravidade."*

- Relação entidade/atributo: *"O mecanismo que faz* ***as pessoas*** *sentirem falta de ar em regiões montanhosas...Cientistas descobriram que esses gases atuam na regulação respiratória, fazendo com que* ***os vasos sanguíneos e vias respiratórias*** *dilatem."*

- Relação parte/todo: *"As larvas ao parasitar* ***a aranha*** *provocam mudanças no comportamento da hospedeira. A relação espúria começa no* ***abdome****."*

2.2.3 Cadeias de Correferência

O conceito de correferência foi dado no início desta seção, entretanto, cabe observar que nem todas as anáforas são correferentes. Como as anáforas associativas introduzem um novo referente no discurso, elas não são exatamente correferentes com seus antecedentes. Apenas as anáforas diretas e indiretas são correferentes.

Uma cadeia de correferência pode ser definida como o conjunto de todas as menções (expressões referenciais) a uma determinada entidade (referente) encontrada em um texto. Este conjunto é responsável pela construção coesa de um texto. Vejamos um exemplo: considere o texto mostrado na Figura 3, onde foram encontradas 8 cadeias de correferências, detalhadas a seguir:

O Ibama (Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis) aplicou, desde maio passado, multas de cerca de R$ 50 milhões a pelo menos 200 fazendeiros cujas terras estão às margens do Taquari. Segundo o Ibama, nas propriedades multadas houve desmatamento nas margens do rio, com a finalidade de abrir espaço à agricultura ou à criação de gado. Em conseqüência, a erosão despejou toneladas de terra no Taquari. O coordenador do Programa Pantanal do Ministério do Meio Ambiente, Paulo Guilherme Cabral, recebeu o estudo da Embrapa sobre o Taquari no início de julho. Ele disse que a ministra Marina Silva criou um grupo de trabalho para analisar as propostas. Existem, segundo Cabral, projetos polêmicos dentro do estudo, como a dragagem (retirada de terra do rio). "Onde vamos colocar essa terra no Pantanal?", questiona. A Sema (Secretaria do Meio Ambiente de Mato Grosso do Sul) informou que defende a dragagem e estuda a contratação de uma empresa para o serviço. A situação do rio Taquari motivou a criação do Programa Pantanal, quando o então presidente Fernando Henrique Cardoso, em 1995, sobrevoou a região e ficou impressionado com o rio assoreado, invadindo fazendas.

Figura 3: Texto de exemplo para cadeias de correferência retirado de corpus (CIENCIA_2005_6515.txt)

- Cadeia 1: O Ibama (Instituto Brasileiro do MeioAmbiente e dos Recursos Naturais Renováveis) - o Ibama
- Cadeia 2: o Taquari - o rio - o Taquari - o Taquari - o rio Taquari
- Cadeia 3: terra - terra do rio
- Cadeia 4: O coordenador do Programa Pantanal do Ministério do Meio Ambiente, Paulo Guilherme Cabral - Ele - Cabral
- Cadeia 5: o estudo da Embrapa - o estudo
- Cadeia 6: a dragagem (retirada de terra de o rio) - a dragagem - o serviço
- Cadeia 7: o Pantanal - a região
- Cadeia 8: A situação de o rio Taquari - o rio assoreado

Podemos observar que as cadeias de correferência são componentes textuais complexos e demonstram a unidade de sentido em um texto. Segundo (KOCH, 2003), as expressões referenciais que compõem as cadeias de correferências não têm apenas a função de referir. Pelo contrário, funcionam como expressões multifuncionais dentro de um texto,

pois contribuem para elaborar o sentido global do texto e indicam pontos de vista, sinalizando dificuldades de acesso ao referente e recategorizando os objetos presentes na memória do leitor.

Nesse sentido, consideramos que as cadeias podem servir de subsídios para análise da qualidade dos sumários extrativos.

2.2.4 Ferramentas de Resolução de Anáfora e Correferência

Nessa seção, são descritos os trabalhos desenvolvidos nessa área considerando a língua portuguesa. O desenvolvimento de algoritmos que realizam a resolução de anáfora é útil em diversas tarefas e aplicações na área de PLN, como, por exemplo, recuperação e extração de informações, sumarização automática, tradutores, entre outros. A pesquisa do processamento automático de textos utilizando a língua portuguesa, em especial o processamento de resolução de anáfora e correferência, ainda é muito recente. Há um grande esforço no meio acadêmico no desenvolvimento de algoritmos desse tipo. Vejamos a seguir alguns trabalhos referentes a esse assunto.

Em (GASPERIN; GOULART; VIEIRA, 2003), temos uma ferramenta desenvolvida para resolução anafórica de descrições definidas chamada ART-*Anaphor Resolution Tool*. Para entrada da ferramenta, são fornecidos arquivos XML processados pelo PALAVRAS (BICK, 2000). Em um dos arquivos é fornecida a informação de sintagmas. Então, os sintagmas nominais são escolhidos e organizados na forma de um conjunto de sintagmas. Os autores consideraram todas as descrições definidas[1] do texto como anáforas e todos os sintagmas nominais como seus possíveis antecedentes. A heurística para identificação de anáforas diretas funciona da seguinte forma:

1. Seleciona o primeiro candidato a anáfora e realiza a extração do núcleo.

2. Localiza os núcleos dos sintagmas nominais.

[1]Sintagmas nominais com artigo definido ("o"/"a") como determinante.

3. Verifica se os núcleos são iguais. Se forem iguais, é indicada uma relação anafórica entre eles.

Como saída, a ferramenta gera um arquivo XML no padrão utilizado pelo MMAX (MüLLER; STRUBE, 2001) com um apontamento para o sintagma nominal antecedente.

O trabalho desenvolvido por (COELHO, 2005) foi de implementação do algoritmo de Lappin e Leass (LAPPIN; LEASS, 1994) para resolução anafórica pronominal em textos da língua portuguesa. Para esse trabalho, ele usou um corpus anotado com informações morfológicas e sintáticas. Baseado nessa informação o algoritmo procura pronomes em um texto e procura reconhecer seu antecedente. O algoritmo trabalha com os textos anotados no formato XML e foi desenvolvido utilizando a linguagem Java[2] na implementação.

Os trabalhos apresentados em (COLLOVINI, 2005) e (COLLOVINI; VIEIRA, 2006b) tiveram por objetivo classificar de forma automática as descrições definidas em quatro classes: novas no discurso, anáforas diretas, anáforas indiretas e associativas. Para essa tarefa, foram extraídas 16 *features* morfológicas e sintáticas para o aprendizado de máquina. O algoritmo utilizado foi o J48, implementado no pacote Weka (WITTEN; FRANK, 2000).

O trabalho apresentado em (COELHO et al., 2006) traz um estudo de corpus sobre resolução das descrições definidas utilizando a informação semântica fornecida pelo parser PALAVRAS (BICK, 2000). O objetivo foi melhorar a performance da resolução das anáforas associativas e indiretas.

O trabalho apresentado em (COLLOVINI; VIEIRA, 2006a) propõe a técnica de balanceamento de corpus por repetição de exemplos com objetivo de melhorar os resultados alcançados em (COLLOVINI, 2005).

Já o trabalho (RIBEIRO-JUNIOR et al., 2007) propõe uma combinação das duas técnicas apresentadas nos trabalhos de (COLLOVINI; VIEIRA, 2006a) e (COELHO et al., 2006), utilizando tanto as informações semânticas para classificação das expressões nas

[2]http:sun.java.com

4 classes, quando a técnica de balanceamento de corpus. Foram implementadas as 16 *features* apresentadas em (COLLOVINI; VIEIRA, 2006a), mais duas novas *features* baseadas nas informações semânticas.

Em (CHAVES, 2007), a autora desenvolveu uma adaptação do algoritmo de (MITKOV, 2002) para a língua portuguesa. Essa abordagem busca resolver anáforas pronominais com o foco de pronomes pessoais de 3ª pessoa. Esse trabalho não utiliza aprendizado de máquina nem conhecimento semântico. Para cada candidato à antecedente o sistema realiza, baseado em regras de natureza multilíngüe, um sistema de pontuação. Dentre os candidatos é selecionado o antecedente que tiver o maior número de pontos.

O trabalho desenvolvido por (SOUZA, 2007) busca extrair e montar cadeias de correferência de um texto. É o primeiro a considerar todos os tipos de sintagmas nominais. O sistema utilizada aprendizado de máquina para esta tarefa. Para o treinamento do sistema foram utilizado textos do gênero jornalístico. Maiores detalhes sobre o sistema podem ser encontrados neste trabalho na seção 4.5.

Com o desenvolvimento da área de resolução de anáforas e correferências, podemos começar a pensar em aplicações de tais sistemas em outras tarefas de PLN, o que é o caso neste trabalho. Aqui procuramos integrar resultados de resolução de anáforas em sumarização automática.

2.3 Sumarização

Sumarização é um processo muito comum. Hoje vemos sumários de vários tipos: notícias em jornais, noticiários, trailler de filme, sinopse de livros, guias de televisão, entre outros. Um sumário pode ser gerado de várias formas: um pequeno filme, um trecho de áudio, mas a forma mais comum encontrada na área de pesquisa é textual.

Sumarização automática de textos é um campo de pesquisa que tem chamado a atenção da área de PLN nos últimos anos. Em parte, isso se dá pelo fato de que a

sumarização incorpora muitos aspectos de conhecimento de linguagem natural e, também, pelo próprio fato de que um sumarizador gere um texto também em linguagem natural.

Segundo (MANI, 2001), a sumarização é o processo de seleção das informações mais relevantes de um texto, com o objetivo de produzir uma versão resumida do mesmo. Um sumário deve conter o conteúdo principal do texto fonte que beneficie um leitor ou uma tarefa. Por exemplo, um sumário pode ser gerado para uma finalidade específica. A geração do sumário vai depender da finalidade do leitor ou grupo de leitores. Quando um sumário tem o foco no usuário é chamado de *user-focused*. Entretando, os sumários mais pesquisados nessa área são os sumários genéricos. Os sumários genéricos independem de usuário leitor. Neste trabalho consideramos sumários genéricos.

Um item importante dentro da área de sumarização automática é o conceito de taxa de compressão. O tamanho do sumário pode ser dimensionado em função do tamanho do texto fonte. Para definir a taxa de compressão podemos considerar que o texto fonte tem 100% e o sumário pode ser gerado usando, por exemplo, 30% de taxa de compressão. Neste caso, o sumário gerado terá 30% do total das palavras do texto fonte.

Um sumário pode ser um *abstract* ou extrato. Um extrato é um sumário onde as sentenças que o compõe são copiadas do texto fonte. Já um *abstract* pode ser formado por trechos do texto fonte ou até mesmo por geração de segmentos textuais a partir do texto original (MANI, 2001).

O processo de sumarização é dividido em três etapas (JONES, 1999), como mostra a Figura 4.

- Análise: é a interpretação do texto para criação de uma representação conceitual.
- Transformação: é a transformação da representação interna do texto original em uma representação interna do sumário.
- Síntese: é a geração do sumário em linguagem natural observando a representação gerada no passo anterior.

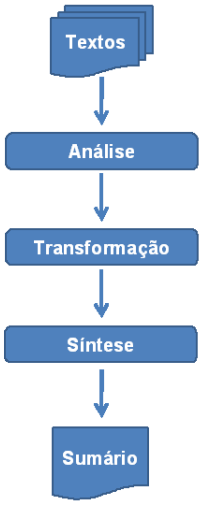

Figura 4: Etapas da Sumarização

A área de sumarização é, ainda, aplicada em diferentes dimensões, como, por exemplo, o sumário pode ser gerado a partir de um único texto fonte ou até mesmo de múltiplos documentos (RADEV, 2004).

A área de sumarização segue duas principais abordagens:

- Abordagem Superficial: Utiliza, geralmente, métodos combinados com técnicas estatísticas para comporem sumários. Alguns exemplos desses métodos são:
  - Palavra-Chave (LUHN, 1958): A idéia principal do texto é expressa pelas palavras que mais aparecem no texto. Algumas abordagens optam pela escolha de palavras que aparecem no título do texto-fonte.
  - Localização da sentença (BAXENDALE, 1958): Para seleção das sentenças mais relevantes de um texto é levada em consideração sua localização. Acredita-se que a primeira e/ou a última sentença do parágrafo podem vir a serem as mais importantes.
  - Palavras sinalizadoras (PAICE, 1981): Neste método, um dicionário é previamente montado para servir de base de consulta para seleção de sentenças relevantes. As sentenças serão consideradas relevantes se conterem uma ou mais palavras desse dicionário.

  - Frase auto-indicativa (PAICE, 1981): Utiliza algumas frases previamente selecionadas, como por exemplo: *"O objetivo deste trabalho é..."* ou *"O foco deste artigo é..."*, na sumarização de trabalhos acadêmicos.

- Abordagem Profunda: Utiliza conhecimento lingüístico para realizar a tarefa de sumarização. Alguns exemplos de conhecimento lingüístico, que podem ser aplicados nesse tipo de abordagem, são (PARDO, 2005b):

  - Relações Semânticas: Capturam a relação sobre a forma como os conhecimentos descritos no texto se relacionam.
  - Relações Intencionais: Como visto na seção 2.1.1 deste trabalho, todo texto possui uma intenção que é transmitida pelo escritor ao escrevê-lo. Esse método visa analisar as relações entre as intenções descritas no texto.
  - RST (*Rhetorical Structure Theory*): É uma das mais importantes teorias discursivas utilizadas em sumarização automática. A RST realiza a análise para descobrir como um texto está organizado funcionalmente, ou seja, qual a função de suas partes para que o objetivo do texto seja satisfeito.

Os sumários podem ser de três tipos:

- Indicativo: Uma síntese do texto, que serve como ponto de partida para a leitura do texto principal, é apresentada.
- Informativo: O sumário gerado é auto-suficiente a ponto de não necessitar uma leitura complementar do texto-fonte.
- Crítico: Uma avaliação crítica, ou opinião do texto-fonte, é expressa. Há, portanto, informação extra.

Nesse contexto de sumarização automática, os sumários informativos são os mais estudados. Os sumários que são utilizados nesta dissertação são desse tipo.

2.3.1 Avaliação de Sumários

A avaliação, na área de sumarização, é notoriamente uma tarefa árdua. Geralmente, o problema envolve juízes humanos na avaliação de sumários. O problema na avaliação humana ocorre quando os juízes não concordam em relação a avaliação do sumário. Além dos problemas de concordância entre juizes, a avaliação humana torna-se muito cara de ser produzida em função do tempo e custo.

A dificuldade em se avaliar sumários está na elaboração dos dados de referência, pois as métricas de avaliação são de comparação entre sumários ideais e sumários automáticos. Os sumários ideais, preferencialmente, não devem ser gerados por ferramentas automáticas, pois sistemas automáticos não garantem qualidade textual, uma vez que a mesma estaria ligada à própria qualidade do sistema de sumarização. A opção mais comum é a geração manual de sumários ideais. Os sumários ideais podem ser construídos pelo próprio autor do texto (pois se considera que existe um domínio sobre o assunto principal descrito no texto) ou, até mesmo, por pessoas que desenvolvem sumários profissionais (RINO; PARDO, 2006).

Os sumários podem ser avaliados sob dois aspectos:

1. Nível de Informatividade: Refere-se à preservação da idéia central do texto-fonte. Com essa avaliação é possível medir o quanto o sumário é informativo e correspondente ao texto-fonte.

2. Qualidade: Essa avaliação diz respeito à construção do sumário enquanto texto. É avaliado o quanto o sumário preserva a coesão e a coerência textual, gramaticalidade, pontuação, entre outros.

Atualmente, há um acréscimo nas pesquisas sobre diferentes métricas para avaliação de sumários automáticos. Os principais métodos adotados de avaliação de sumários são[3]:

[3]Considere **Ni** número de sentenças do sumário ideal, **Na** número de sentenças do sumário automático

- *Precision*: Indica o índice de sentenças relevantes no sumário automático, segue sua fórmula: (P=Nc/Na).

- *Recall*: Indica a representatividade do sumário em relação a todos os dados considerados relevantes, fórmula: (R=Nc/Ni).

- *F-Measure*: Utiliza as medidas de *Precision* e *Recall* para produzir uma única medida de eficiência, observe a fórmula: (Fm=2*R*P/R+P).

- *Relative Utility (RU)*: Apresenta uma medida que busca identificar a utilidade do sumário(trabalho desenvolvido em (RADEV; JING; BUDZIKOWSKA, 2000)). A medida é obtida manualmente, sendo que juízes humanos fornecem uma nota para cada sentença do texto-fonte, indicando sua importância. A partir dessa pontuação, é formado o sumário ideal. As sentenças do sumário produzido automaticamente também recebem notas. Então, são comparadas as notas entre o sumário ideal e o sumário automático. Se as medidas forem próximas (entre o sumário ideal e o automático), o sumário automático é considerado suficientemente informativo para ser útil.

- *Rouge - Recall-Oriented Understudy for Gisting Evaluation*: É um pacote de medidas para determinar a qualidade dos sumários automáticos comparando-os a sumários ideais. As medidas ROUGE utilizam, principalmente, a co-ocorrência de n-gramas, de forma completamente automática (LIN, 2000) Maiores detalhes sobre a ROUGE podem ser encontrados na seção 4.6.

Todas as medidas, descritas acima, são utilizadas para avaliar a informatividade do sumário. Para avaliar sua qualidade a respeito de coesão e coerência, ainda são realizadas avaliações humanas. Como, por exemplo, no trabalho desenvolvido em (PARDO; RINO, 2002), em que se solicitou que os juízes humanos verificassem a textura do sumário e sua legibilidade.

e **Nc** número de sentenças iguais entre o sumário ideal e o automático.

Outros critérios que podem ser utilizados na avaliação da qualidade do sumário, são: presença de referência anafórica não resolvida, falta de coesão entre as sentenças, presença de palavras diferentes que expressam o mesmo pensamento/idéia central, ortografia, gramática, compreensão, siglas seguidas de suas expressões completas entre outras. Esses critérios estão ligados a avaliação da qualidade do sumário.

Neste trabalho, a medida de qualidade é a mais relevante. No entanto, outras medidas poderão ser também consideradas.

2.3.2 Ferramentas de Sumarização Automática

Essa seção é dedicada à ferramentas de sumarização desenvolvidas e testadas para língua portuguesa.

O sistema DMSumm (PARDO, 2002) - *Discourse Modeling Summarizer* é um sumarizador baseado em modelagem discursiva e utiliza a abordagem profunda na sumarização. A entrada no sistema é feita por um arquivo com informações referentes ao discurso anotado manualmente. O DMSumm implementa em seu sumarizador as seguintes abordagens: semântica, relacional e retórica. O objetivo desse sumarizador é a construção de sumários coerentes. O DMSumm é baseado no modelo de discurso Problema-Solução (JORDAN, 1980), que é um modelo bastante utilizado em diversos gêneros e domínios textuais.

O sumarizador NeuralSumm (PARDO; RINO; NUNES, 2003) utiliza Redes Neurais como técnica de *Machine Learning* para seleção das sentenças mais relevantes em um texto para compor o sumário, sendo assim um sumarizador extrativo. São utilizadas *features* que servem para identificar as sentenças em essenciais, complementares e supérfluas. As sentenças essenciais são as primeiras selecionadas para composição do sumário, as complementares serão ou não acrescentadas no sumário, dependendo da taxa de compressão determinada pelo usuário, e as sentenças supérfluas são descartadas.

O GistSumm (PARDO, 2005a) é um sumarizador automático de texto que utiliza abordagem superficial para selecionar as sentenças que irão compor o sumário. O GistSumm procura simular a forma de sumarização humana, buscando a sentença que melhor expressa a idéia do texto (sentença-gist). Além da sentença-gist o sumarizador busca outras sentenças que complementam a sentença-gist. Maiores detalhes sobre a ferramenta estão disponíveis na seção 4.3 deste trabalho.

O RheSuma-2 (RINO; CARBONEL, 2006) é uma ferramenta de sumarização automática que utiliza a abordagem profunda. O RheSuma-2 foi concebido a partir de duas outras ferramentas: DiZer (PARDO; NUNES, 2006) e o RheSumaRST (SENO, 2005). O DiZer é responsável pela análise do texto e geração das informações retóricas. O RheSumaRST é um gerador de sumários a partir das informações RST (MANN; THOMPSON, 1987). Além da teoria discursiva RST o sistema implementa a teoria de veias (CRISTEA; IDE; ROMARY, 1998) como uma solução para quebras de cadeias de correferência.

O SatSumm (NETO; B; GOMES, 2007) é uma ferramenta que implementa um sumarizador utilizando abordagem superficial. O foco dessa ferramenta é um sumarizar textos jornalísticos por conterem sempre uma idéia central bem definida. Alguns passos são realizados antes do processamento dos sumários como, por exemplo: *casefold, stemming e remoção de stopwords*. A técnica estatística utilizada no SatSumm é o TF-ISF-*Term Frequency Inverse Sentence Frequency*. O SatSumm apresenta uma interface gráfica para uma melhor manipulação da ferramenta pelo usuário.

O trabalho de Carbonel (CARBONEL, 2007) traz um estudo aprofundado dos fenômenos textuais que ocorrem nos sumários gerados automaticamente, sendo o foco dele o estudo sobre as quebras dos elos correferenciais. Além desse estudo, esse trabalho propõe e implementa um sumarizador automático usando a Teoria das Veias (CRISTEA; IDE; ROMARY, 1998). Essa teoria propõe um mapeamento do fenômeno referencial a partir da construção retórica baseada na RST. Esse trabalho foi a reimplementação do sistema Rhesuma-RST.

O sistema de sumarização automático SuPor-2 descrito em (LEITE; RINO, 2006a), utiliza aprendizado de máquina para seleção das informações do texto para geração do sumário. Esse sumarizador é um sistema que utiliza o método extrativo na composição dos sumários. Maiores detalhes desse sistema, podem ser encontrados na seção 4.4 desta dissertação.

Neste capítulo, foram discutidos importantes conceitos para esta proposta. Foi apresentada a conceitualização de coerência e coesão e sua importância para a compreensão de um texto. Vimos, também, correferência e anáforas, assim como, sintagmas nominais com suas classificações e alguns sistemas que utilizam essa informação para processamento de linguagem natural. Os sintagmas nominais são objetos de estudo neste trabalho.

Finalizamos o capítulo com a conceitualização da área de sumarização automática e com algumas ferramentas que foram desenvolvidas para a língua portuguesa.

Capítulo 3

Trabalhos Relacionados

Na última década, a pesquisa tem crescido na área de automatização de processos de textos em forma eletrônica. Um dos problemas, que têm-se estudado nessa área, é a resolução de correferência. Ultimamente, temos visto o surgimento de trabalhos que relacionam resolução de correferência e sumarização, área em que se insere nesta dissertação.

3.1 Trabalhos de Resolução de Correferência

Nesta seção, são descritos os trabalhos relacionados à resolução de correferência.

O trabalho de Mitkov (MITKOV, 1998) apresenta uma proposta para resolução de correferência para pronomes. Mitkov propõe um algoritmo de baixo custo computacional e de rápida execução. A abordagem evita a análise sintática complexa e análise de discurso. O trabalho de Mitkov utiliza um *tagger* para extrair informações de *part-of-speech* (*pos*), um simples identificador de sintagmas nominais e um localizador de antecedentes candidatos. O algoritmo funciona da seguinte forma: a) processa os textos os textos são processados para selecionar os sintagmas nominais e etiquetar com as informações de *pos*; b) realiza uma busca na sentença corrente e outras duas anteriores, procurando os sintagmas nominais; c) verifica e seleciona os sintagmas nominais que concordam em gênero e

número com o pronome anafórico, formando, então, um conjunto de candidatos; d) aplica um algoritmo que irá indicar o antecedente com base em 10 *features*, os sintagmas nominais recebem uma pontuação e é escolhido como antecedente o sintagma nominal com a maior pontuação.

No trabalho de Amo (AMO et al., 1999) foi desenvolvido um algoritmo para resolução de correferência com foco somente em nomes próprios. Foi analisada a relação denominada por eles de "replicância", que é a relação entre nomes próprios baseados na ortografia. Para o desenvolvimento do algoritmo, foi utilizada a linguagem Prolog[1]. O algoritmo aprende a partir de uma base de exemplos, no qual ele aplica uma rápida análise sobre pares de substantivos. O objetivo, nesse trabalho, é o reconhecimento da existência da resolução de correferência utilizando a ortografia das palavras. Vejamos alguns exemplos adotados por esse trabalho, correferência entre os seguintes nomes: José Luiz Martinez, J L Martinez, J. Martinez, Martinez, Luiz Martinez ou então, a correferência baseada na utilização de abreviaturas, como por exemplo: União Européia - UE e Boletim Oficial do Estado - BOE.

Luo e outros pesquisadores em (LUO et al., 2004), utiliza *Machine Learning* para resolução de correferência. Ele define que a resolução de correferência é o particionamento das menções para uma entidade. Uma menção é uma instância do referente para um objeto no mundo real. Já uma coleção de menções, que se referem a um mesmo objeto em um documento, forma uma entidade. A técnica de ML, utilizada nesse trabalho, foi o *Bell Tree*. Maiores detalhes desta técnica podem ser encontrados no trabalho mencionado. Foram extraídas 17 *features* para o aprendizado. Nos testes e na avaliação dos resultados foi utilizado o corpus MUC6 (MUC-6, 1995).

Para resolução de correferência no trabalho de Ponzetto (PONZETTO; STRUBE, 2006), os autores utilizaram o cálculo de entropia como técnica de *machine learning*. A resolução de correferência é considerada uma tarefa de classificação em que dado um par de expressões ele deve ser categorizado como referente ou não. Para realizar essa tarefa

[1]http://www.swi-prolog.org

de classificação, foi necessário um pré-processamento para etiquetação do texto, com informações de *part-of-speech*, reconhecimento de entidades mencionadas e um *chunker* que delimita os sintagmas. Para análise de correferência são analisadas 12 *features*.

O trabalho de Nicolae (NICOLAE, 2006) destaca a importância de reconhecimento de entidades mencionadas em diversas áreas como: a tradução automática, recuperação de informação e na sumarização automática. O objetivo desse trabalho é a detecção de entidades mencionadas (selecionar todas entidades mencionadas em um texto) e agrupá-las em classes (classes que representam o mundo real). Foram utilizados alguns recursos da Wordnet (MILLER, 1995), como, por exemplo, informações de sinonímias e hiponímias para ser utilizado como parâmetros na análise. Para o desenvolvimento do algoritmo foram extraídas *features* que foram utilizadas no processamento de resolução da cadeia de correferência.

O objetivo do trabalho de Ng descrito em (NG, 2007) é propor novas *features* lingüísticas para resolução de correferência e realizar um comparativo entre os resultados obtidos com essas novas features propostas com outros sistemas e algoritmos que utilizam conhecimento semântico. Esse trabalho segue diversos outros, do mesmo autor, que possuem o mesmo propósito de identificar cadeias de correferência de forma automática (NG, 2005b), (NG, 2005a) e (NG, 2003).

3.2 Trabalhos de Resolução de Correferência e Sumarização

Os trabalhos, citados na seção anterior, destacam-se pelas diferentes formas de resolução de correferência. Esta tarefa é de grande importância para área de PLN e suas sub-áreas, como a sumarização automática. A seguir, destacamos os trabalhos que utilizam a resolução de correferência especificamente na sumarização automática.

A pesquisa descrita por Azzam em (AZZAM; HUMPHREYS; GAIZAUSKAS, 1999)

descreve o uso de cadeias de correferência para a produção de sumários. Diferentemente desta dissertação, o trabalho de Azzam usa as informações de cadeias de correferência para seleção das sentenças que irão compor o sumário, pois considera que a "melhor" cadeia é o tópico mais relevante do texto. Para a seleção da "melhor" cadeia foram utilizados os seguintes critérios:

- Tamanho da Cadeia: A cadeia que contém a maior quantidade de expressões é considerada a "melhor" cadeia. Em caso de empate, outros critérios são avaliados.
- Propagação da Cadeia: Envolve um cálculo da distância, pois considera que a cadeia que mais se expande no texto deve ser considerada a melhor.
- Início da Cadeia: Como último critério, é utilizada uma medida que considera a cadeia que contém a expressão no primeiro parágrafo do texto ou, até mesmo, no título.

No trabalho desenvolvido por Nenkova em (NENKOVA; SIDDHARTHAN; MCKEOWN, 2005), destaca-se a pesquisa em aprendizado automático do status cognitivo do ouvinte/-leitor, na área de sumarização automática para múltiplos documentos. Essa é uma área rica para pesquisas, pois sumários gerados a partir de diferentes documentos podem conter, na maioria das vezes, muita informação, pouca informação ou até mesmo repetição de informação sobre seu referente. O foco do trabalho é na modelagem de referentes para pessoas (nomes próprios). Esse trabalho está apoiado na seguinte premissa: Se o referente é desconhecido para o ouvinte/leitor, exatamente no ponto em que ele é mencionado no discurso, deveria ter sido incluída uma descrição sobre esse referente anteriormente. Para isso foram montados dois diferentes cenários para a pesquisa:

- Cenário 1: A pessoa é conhecida ou não pelo ouvinte/leitor.
- Cenário 2: A pessoa é o maior ou menor protagonista da notícia.

Para essa pesquisa foi utilizada a técnica de árvores de decisão para identificar as possibilidades dentro desses cenários.

Em (KASHANI; POPOWICH, 2006) foi desenvolvido por Kashani e Popowich um algoritmo para resolução de correferência no tratamento de pronomes. O algoritmo desenvolvido nesse trabalho, realiza a geração de pronomes em sumários automáticos, com o objetivo de evitar a repetição do nome. Para o desenvolvimento dos experimentos foi utilizada a ferramenta Lingpipe[2] que realiza a tokenização, detecção de entidades mencionadas e resolução de correferência. O funcionamento do algoritmo dá-se da seguinte forma: ele descobre o referente e troca pelo pronome correspondente. São considerados apenas pronomes da 3a. pessoa do singular, pois se considerassem outros tipos de pronomes haveria mudança na estrutura da sentença. Os autores destacam que esse trabalho pode, ainda, ser aplicado na sumarização de multi-documentos.

Um dos trabalhos de maior relevância para essa dissertação é apresentado em (STEINBERGER et al., 2007) por Steinberger et al. Esse trabalho propõe duas formas de uso para aplicação de correferência na área de sumarização. A primeira forma de uso teve como proposta a geração de sumários explorando a informação lexical e a resolução automática de correferência. Foram realizados experimentos de comparação entre sumários gerados usando somente informação lexical e sumários gerados utilizando informação lexical e resolução anafórica. A segunda forma de uso realiza uma verificação no sumário, com objetivo de realizar a correção dos referentes. Para a resolução de correferência, nos dois experimentos, foi utilizado o GUITAR (POESIO; KABADJOV, 2004) como ferramenta que possibilita resolução de pronomes e resolução de descrições definidas. No primeiro experimento, os autores utilizaram a informação das cadeias, extraídas com o GUITAR, para realizar a troca dos termos nominais anafóricos pelo primeiro elemento da sua cadeia de correferência. Vejamos um exemplo dessa transformação em um texto como demonstra a Figura 5.

No texto, como mostra a Figura 5, foi possível encontrar 8 cadeias anafóricas:

[2]http://www.alias-i.com/lingpipe/

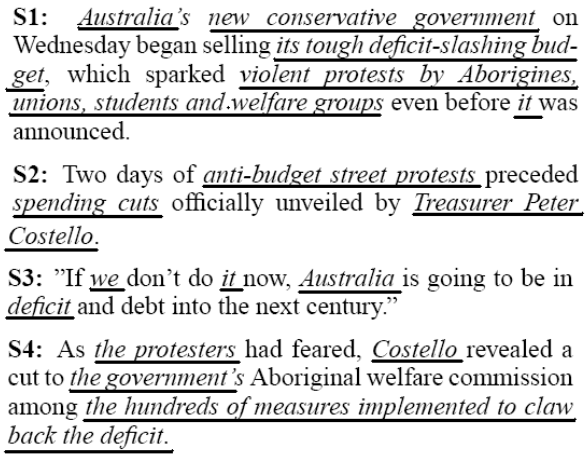

Figura 5: Texto retirado do artigo

- Cadeia 1: Australia - we - Australia
- Cadeia 2: its new conservative government (Australia's new conservative government) - the government
- Cadeia 3: its tough deficit-slashing budget (Australia's tough deficit-slashing budget) - it
- Cadeia 4: violent protests by Aborigines, unions, students and welfar e groups - anti-budget street protests
- Cadeia 5: Aborigines, unions, students and welfare groups - the protesters
- Cadeia 6: spending cuts - it - the hundreds of measures implemented to claw back the deficit
- Cadeia 7: Treasurer Peter Costello - Costello
- Cadeia 8: deficit - the deficit

Na Figura 6 temos a forma como ficaram as sentenças após o processamento do experimento. Observamos que o algoritmo proposto nesse trabalho efetuou a troca de todas as expressões pelo primeiro termo da cadeia. Todas as expressões que foram trocadas estão em destaque na Figura 6. Após esse processamento de troca de expressões, os autores

processam um algoritmo que irá selecionar os principais termos do texto. As sentenças que possuem os principais termos serão selecionadas para compor o sumário.

S1: *Australia's new conservative government* on Wednesday began selling *its tough deficit-slashing budget*, which sparked *violent protests by Aborigines, unions, students and welfare groups* even before *it* was announced.

S2: Two days of *anti-budget street protests* preceded *spending cuts* officially unveiled by *Treasurer Peter Costello*.

S3: "If *we* don't do *it* now, *Australia* is going to be in *deficit* and debt into the next century."

S4: As *the protesters* had feared, *Costello* revealed a cut to *the government's* Aboriginal welfare commission among *the hundreds of measures implemented to claw back the deficit*.

Figura 6: Texto retirado do artigo com a saída produzida pela ferramenta.

No segundo experimento foi realizada a verificação das cadeias nos sumários gerados, o método funciona da seguinte forma:

- Aplicar o algoritmo de resolução anafórica no texto e criar as cadeias de correferência.
- Identificar as sentenças que são extraídas para geração do sumário.
- Analisar que para cada cadeia de correferência é efetuada a troca da primeira ocorrência da cadeia no sumário pela primeira ocorrência no texto-fonte. Após esse passo, todas as cadeias que aparecem no sumário e no texto-fonte começarão com a mesma forma lexical.
- Rodar o algoritmo de resolução anafórica no sumário
- Analisar para toda expressão nominal no sumário: se a expressão é parte de uma cadeia no texto-fonte e ela não está resolvida no sumário (por não resolvida entende-se que não foi possível encontrar seu antecedente) ou se ela passou a fazer parte de uma diferente cadeia no sumário então se troca a expressão anafórica pelo núcleo da primeira expressão da cadeia do texto-fonte.

Esta dissertação segue a mesma linha de (STEINBERGER et al., 2007). Diferentemente do que é proposto pelos autores, optar por sempre substituir o termo anafórico pela primeira expressão da cadeia, nossa abordagem é desenvolver heurísticas para indicar qual é a "melhor" expressão que deverá ser escolhida para substituição.

Os trabalhos relacionados que utilizam correferência na sumarização, citados acima, mostram que há um considerável interesse em pesquisa nessa área. Cabe ressaltar que os trabalhos relacionados consideram, principalmente, a língua inglesa, sendo esta uma primeira proposta para desenvolvimento de um sistema com a utilização da correferência na edição de sumários para a língua portuguesa.

Capítulo 4

Materiais e Métodos

Este capítulo apresenta os materiais e métodos utilizados no desenvolvimento deste trabalho. Servem de base ao trabalho o corpus Summ-it e as ferramentas PALAVRAS, MMAX, ROUGE, GistSumm e SuPor-2. O analisador sintático PALAVRAS é usado na análise gramatical dos textos; a ferramenta MMAX é usada na anotação de correferência do corpus; a ferramenta ROUGE é usada para realizar a avaliação automática dos sumários; e os sumarizadores automáticos GistSumm e SuPor-2 são utilizados na geração dos sumários extrativos.

4.1 PALAVRAS

O analisador sintático PALAVRAS, descrito em (BICK, 2000), é uma ferramenta robusta utilizada para a análise sintática automática do português. Na análise morfossintática, o PALAVRAS traz informações como classe gramatical, gênero, número e flexão verbal. Apresenta, também, informações sobre a estrutura da sentença, sua análise estrutural (sintagmas nominais e verbais) e a função de seus constituintes.

O analisador possui três formatos de saída. O primeiro formato utiliza a forma gráfica de árvore, que representa a estrutura do texto, em que as folhas são compostas

pelas expressões lingüísticas e os ramos da árvore representam a análise sintática da sentença. Na Figura 7 temos a análise da frase “O fumo é extremamente prejudicial à saúde”, cuja análise sintática indica, por exemplo, as funções S (sujeito) e P (predicado). Mais abaixo, temos a análise morfossintática, por exemplo, art (artigo) e n (substantivo).

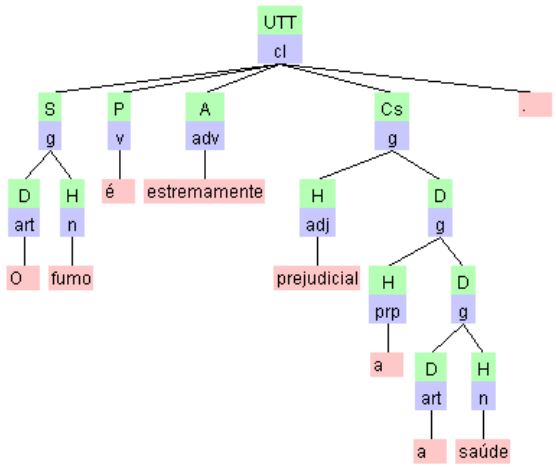

Figura 7: Árvore morfossintática gerada pelo PALAVRAS

O segundo formato gerado é uma saída no formato texto, Figura 8, que traz as mesmas informações geradas na árvore, em que, em cada linha, o primeiro símbolo representa a função sintática para cada elemento ou grupo; depois dos dois pontos, temos a categoria da palavra; entre parênteses, temos forma canônica e as informações de flexão, gênero e número e, após os parênteses, temos a palavra da sentença. O símbolo “=”, no início de cada linha, representa o nível da expressão na árvore sintática.

```
UTT:cl(fcl)
S:g(np)
=D:art('o' <artd> M S)      O
=H:n('fumo' M S)      fumo
prejudicial [prejudicial] ADJ M S @<SC
Cs:g(ap)
=D:adv('extremamente' <quant>)      extremamente
=H:adj('prejudicial' M S)      prejudicial
=D:g(pp)
==H:prp('a' <sam->)      a
==D:g(np)
===D:art('o' <artd> <-sam> F S)      a
===H:n('saúde' F S)      saúde
```

Figura 8: Formato texto gerado pelo PALAVRAS

O terceiro formato é um arquivo XML no padrão de anotação Tiger[1]. Nesse formato, o modelo de dados é baseado em grafos de sintaxe, isto é, grafos direcionados acíclicos com uma única raiz. Palavras, etiquetas de *part-of-speech*, etiquetas morfológicas e lemma são atributos do elemento "terminal". Elementos não-terminais são representados através de um elemento chamado "nonterminal" e apontam aos terminais correspondentes através de um identificador. Um exemplo de uso desse formato de codificação pode ser visto na Figura 9

```
- <body>
  - <corpus>
    - <s id="s1" ref="1" source="Running text" forest="1" text="O fumo é extremamente prejudicial a saúde.">
      - <graph root="s1_500">
        - <terminals>
            <t id="s1_1" word="O" lemma="o" pos="art" morph="M S" sem="--" extra="--" />
            <t id="s1_2" word="fumo" lemma="fumo" pos="n" morph="M S" sem="cm" extra="--" />
            <t id="s1_3" word="é" lemma="ser" pos="v-fin" morph="PR 3S IND VFIN" sem="--" extra="mv" />
            <t id="s1_4" word="extremamente" lemma="extremamente" pos="adv" morph="--" sem="--" extra="quant" />
            <t id="s1_5" word="prejudicial" lemma="prejudicial" pos="adj" morph="M S" sem="--" extra="--" />
            <t id="s1_6" word="a" lemma="o" pos="art" morph="F S" sem="--" extra="--" />
            <t id="s1_7" word="saúde" lemma="saúde" pos="n" morph="F S" sem="state-h" extra="--" />
            <t id="s1_8" word="." lemma="--" pos="pu" morph="--" sem="--" extra="--" />
          </terminals>
        - <nonterminals>
          - <nt id="s1_500" cat="s">
              <edge label="STA" idref="s1_501" />
            </nt>
          - <nt id="s1_501" cat="fcl">
              <edge label="S" idref="s1_502" />
              <edge label="P" idref="s1_3" />
              <edge label="Cs" idref="s1_503" />
              <edge label="fCs" idref="s1_504" />
            </nt>
          - <nt id="s1_502" cat="np">
              <edge label="DN" idref="s1_1" />
              <edge label="H" idref="s1_2" />
            </nt>
          - <nt id="s1_503" cat="adjp">
              <edge label="DA" idref="s1_4" />
              <edge label="H" idref="s1_5" />
            </nt>
          - <nt id="s1_504" cat="np">
              <edge label="DN" idref="s1_6" />
              <edge label="H" idref="s1_7" />
            </nt>
          </nonterminals>
        </graph>
      </s>
    </corpus>
  </body>
```

Figura 9: Formato TIGER

4.2 MMAX

O MMAX - Multi-Modal Annotation in XML - (MüLLER; STRUBE, 2001) é um software utilizado para anotações de discurso. O MMAX foi utilizado para anotação de correferência do corpus Summ-it. Essa anotação foi realizada de forma manual em

[1]http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/doc/html/TigerXML.html

duas etapas. Na primeira etapa, os anotadores utilizaram a ferramenta para delimitação dos sintagmas nominais. Na segunda etapa, foi realizada a identificação das cadeias e a classificação das expressões em novas no discurso, associativas, diretas e indiretas. Suas definições foram apresentadas na seção 2.2.2 deste trabalho.

Na Figura 10, mostrada abaixo, temos um exemplo da formação da cadeia através dos referentes "o mal de Alzheimer, doença degenerativa do cérebro que mais afeta os idosos pelo mundo", "a doença", "o mal de Alzheimer" e, novamente, "a doença" formando uma cadeia de correferência.

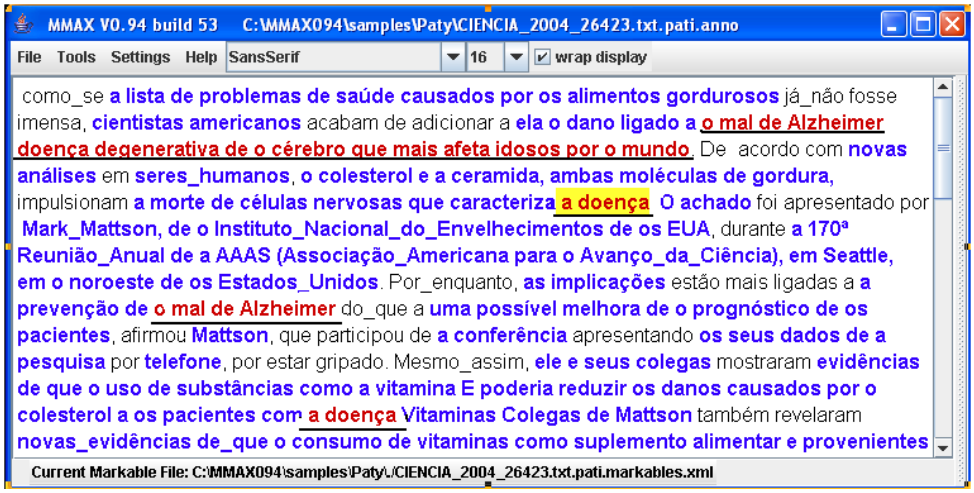

Figura 10: Interface gráfica do MMAX

Como entrada da ferramenta MMAX, é fornecido um arquivo XML com as palavras que fazem parte do texto que se quer anotar. Cada palavra possui um atributo identificador *(id)*. O arquivo XML de entrada do MMAX pode ser visualizado na Figura 11.

```
<words>
      <word id="word_1">como_se</word>
      <word id="word_2">a</word>
      <word id="word_3">lista</word>
      <word id="word_4">de</word>
      <word id="word_5">problemas</word>
...
</words>
```

Figura 11: Arquivo base para o MMAX

Como saída, a ferramenta MMAX produz um arquivo XML, representado na Fi-

gura 12. Cada anotação realizada usando o MMAX é identificada pelo elemento *markable*, como pode ser visto na Figura 12. As informações de delimitação e classificação do sintagma nominal são dadas através dos atributos *span* e *np_form*. Além dessas informações, o arquivo disponibiliza a informação de correferência através dos atributos *member, status* e *is_anaphoric*. O valor "old", para o atributo *status*, indica que a expressão é "velha" no discurso, isto é, uma entidade já mencionada. Se o valor de status for "new", indica que a expressão é "nova" e é a primeira vez que ela aparece no texto. O atributo *is_anaphoric* indica a classe anafórica, que pode ser: direta (*direct*) ou indireta (*indirect*). Esse atributo somente receberá um desses valores se o atributo *status* tiver o valor "old".

```
…
<markable id="markable_2" span="word_3..word_5" np_n="yes" np_form="bare-np"
          member="set_16"/>
<markable id="markable_3" span="word_47..word_51" is_anaphoric="direct" np_form="def-np"
          status="old" member="set_16" />
<markable id="markable_4" span="word_73..word_76" is_bridging="other-brigding"
          np_form="def-np" status="associative" pointer="markable_4" /
...
```

Figura 12: Arquivo XML de saída do MMAX

Com a saída gerada pelo MMAX, é possível identificar a cadeia de correferência através do atributo *member*, pois toda a cadeia terá o mesmo identificador para esse atributo. Veja, como exemplo, a Figura 13, que mostra os atributos *member* com o valor *"set_7"* indicando que todas as expressões fazem parte da mesma cadeia de correferência.

```
<markables>
<markable id="markables_1"  span="word_1..word_2" np_n="yes"
     np_form="def-np"       status="new"  member="set_7" />
<markable id="markables_11" span="word_35..word_36"
    is_anaphoric="indirect" np_form="def-np"  status="old" member="set_7" />
<markable id="markables_38" span="word_124..word_125"
    is_anaphoric="direct"   np_form="def-np"  status="old" member="set_7" />
<markable id="markables_85" span="word_300..word_302"
    is_anaphoric="direct"   np_form="def-np"  status="old" member="set_7" />
</markables>
```

Figura 13: Arquivo de saída do MMAX com as Cadeias de Correferência

Como mostra a Figura 13, as expressões estão apontando para o *markable_1*, e este *markable* está com o atributo *status* com o valor *new* indicando que o *markable_1* é a expressão nova no discurso. Todos os *markables* da Figura 13 fazem parte da mesma

cadeia de correferência.

4.3 GISTSUMM

O GistSumm foi desenvolvido por Thiago Pardo (PARDO, 2005a) no NILC - Núcleo Interinstitucional de Lingüística Computacional. O GistSumm é um sumarizador automático de texto que utiliza abordagem superficial para selecionar as sentenças que irão compor o sumário. O sumarizador utiliza o método extrativo e possui duas premissas:

- Todo texto possui uma idéia principal;
- É possível identificar em um texto uma sentença que melhor representa sua idéia principal (sentença-gist).

Com base nessas premissas o GistSumm tem como objetivo a identificação da sentença-gist e as sentenças que a complementam para composição do sumário extrativo. Na Figura 14 temos a arquitetura do sistema GistSumm.

O processo inicia com a entrada de um texto-fonte, o qual se deseja sumarizar (passo 1). No passo 2, inicia-se a segmentação sentencial, a qual delimita todas as sentenças do texto-fonte observando os sinais de pontuação tradicionais (ponto final, exclamação e interrogação). Logo em seguida, no passo 3, as palavras das sentenças são armazenadas em vetores. No passo 4, o GistSumm transforma todas as palavras em minúsculas buscando uma padronização. Com a ajuda de um léxico, o sistema transforma todas as palavras em sua forma lexical e, além disso, utilizando uma *stoplist*, faz a retirada de *stopwords* da frase (por exemplo, artigos e preposições). Após essa preparação, é aplicado o método de ranqueamento das sentenças. Esse método é selecionado pelo usuário no momento da sumarização. O GistSumm implementa dois métodos de ranqueamento: o método de *Keywords* e *Average Keywords*.

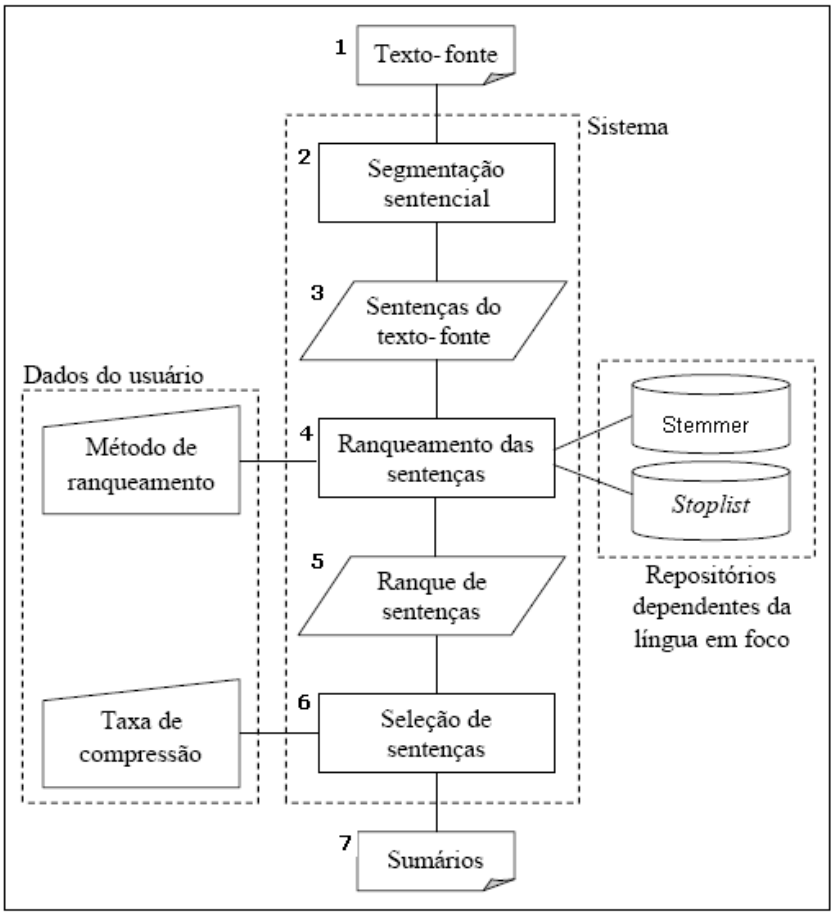

Figura 14: Arquitetura do sistema GistSumm

Por qualquer um dos métodos, a sentença com maior pontuação é considerada como sendo a sentença-gist do texto-fonte. Por isso, no GistSumm, os métodos de ranqueamento são utilizados para determinar a idéia principal do texto-fonte. As sentenças e suas pontuações são passadas para o passo 6, para realizar a seleção de sentenças. No processo de seleção de sentenças do texto-fonte para formar o sumário, o GistSumm executa os seguintes itens:

1. Calcula a média da pontuação das sentenças do texto-fonte e assume essa como sendo a *baseline* para corte das possíveis sentenças que formarão o sumário;

2. Seleciona, para formar o sumário, juntamente com a sentença-gist, todas as sentenças do texto-fonte que contenham pelo menos uma palavra que tenha uma das canônicas da sentença-gist e possuam uma pontuação maior que a *baseline* calculada

no item 1.

No passo 7, os sumários são gerados e as sentenças são escolhidas baseadas numa taxa de compressão definida pelo usuário no início do processo de geração do sumário.

A escolha pela ferramenta GistSumm justifica-se pelo fato de ser uma ferramenta robusta de geração de sumários extrativos, que não requer treinamento específico (que é o caso de sumarizadores que utilizam técnicas de *machine learning*) nem anotação adicional (como no caso de sumarizadores que utilizam anotações de abordagens lingüísticas para sumarizar).

4.4 SuPor-2

O SuPor-2 (LEITE; RINO, 2006a) é uma versão modificada do SuPor (MóDOLO, 2003) (Ambiente para Sumarização Automática de Textos em Português), um sumarizador extrativo que depende de informações fornecidas por um engenheiro de conhecimento para treino e combinação de diversos métodos de extração de informações relevantes. O SuPor-2 utiliza um classificador para treino e seleção das sentenças mais relevantes do texto-fonte. O algoritmo de classificação utilizado foi o Naïve-Bayes. Para treinamento e classificação o sistema utiliza a ferramenta Weka (WITTEN; FRANK, 2000)

As features utilizadas pelo SuPor-2 são baseadas na freqüência das palavras, tamanho e posição da sentença, ocorrência de nomes próprios, análise de cadeias lexicais, e outras análises do discurso (detalhadas em (LEITE; RINO, 2006a)).

A Figura 15 mostra os passos para o treinamento do classificador. O sistema SuPor-2 usa um Léxico e uma *StopList* na fase de pré-processamento das informações, logo em seguida o sistema processa as *features* que serão usadas no classificador. O sistema monte um conjunto de tuplas que é processado pelo sistema Weka. A saída do módulo de treinamento é um arquivo com os parâmetros do classificador. Como foi usado o Naïve-Bayes, esse arquivo contém as probabilidades usadas pelo classificador na etapa

de extração.

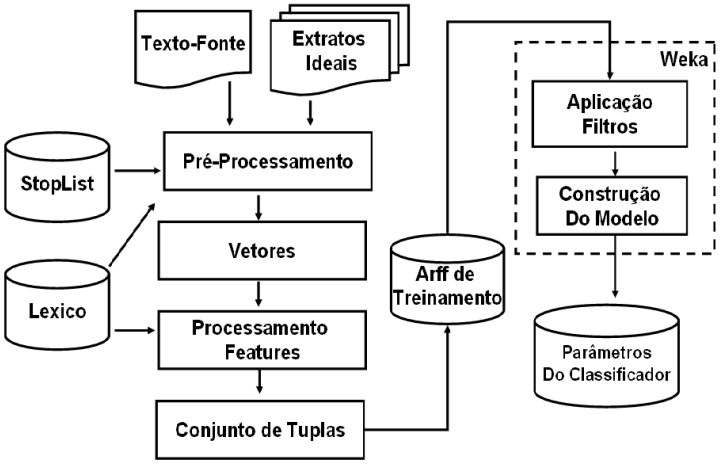

Figura 15: Módulo de treinamento do SuPor-2

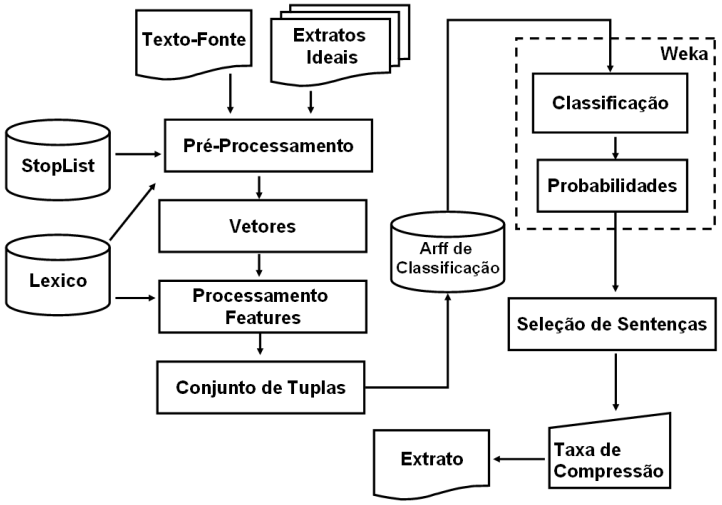

Figura 16: Módulo de seleção do SuPor-2

A Figura 16 mostra as etapas de seleção de sentenças. Para isso, o sistema usa novamente um léxico e a *StopList* na fase de pré-processamento. O Conjunto de tuplas é montado com suas respectivas *features* e processado pelo classificador que foi gerado no módulo de treinamento. O classificador, então, realiza a seleção das sentenças de maior relevância do texto fonte para geração no sumário. O número de sentenças que irão compor esse sumário estará relacionado a taxa de compressão desejada.

A justificativa pela escolha do sistema de sumarização SuPor-2 é pelo fato de ser um dos melhores sistemas de sumarização automática para o português, conforme descrito

em (LEITE; RINO, 2006b) e (LEITE et al., 2007).

4.5 Sistema de Resolução Automática de Correferência

O sistema desenvolvido por Souza em (SOUZA, 2007) tem como objetivo automatizar a resolução de correferência para a língua portuguesa usando uma abordagem baseada em aprendizado de máquina supervisionado.

O sistema seleciona subconjuntos de expressões (cadeias de correferência) de um texto, através da identificação dos pares de expressões anafóricas. Com o objetivo específico de aprender um classificador, uma base de dados deverá ser constituída através da extração de exemplos de um corpus anotado.

Esses subconjuntos são extraídos juntamente com 10 características que serão utilizadas pelo classificador. Essas características são:

1. Comparação de núcleo: compara o núcleo dos dois sintagmas;
2. Distância: determina a distância em frases entre os dois sintagmas.
3. Antecedente é pronome: verifica se o núcleo do sintagma eleito como antecedente é um pronome.
4. Anáfora é pronome: verifica se o núcleo do sintagma eleito como anáfora é um pronome.
5. São nomes próprios: verifica se ambos sintagmas são nomes próprios
6. Concordância de gênero: verifica caso o gênero (masculino/feminino) dos dois sintagmas coincidam.

7. Concordância de número: verifica se os dois sintagmas concordam em número (ou seja, ambos estão no singular ou ambos no plural);

8. Sujeito: verifica se ambos sintagmas são sujeitos

9. Concordância semântica: caso os dois nomes núcleos sejam diferentes e possuam etiquetas semânticas idênticas

10. Mesmo grupo semântico: caso os dois nomes núcleos sejam diferentes e possuam etiquetas semânticas que pertençam ao mesmo grupo, o valor deste atributo é verdadeiro.

O sistema seleciona, classifica e agrupa as expressões para a montagem das cadeias de correferência.

É importante salientar que esse trabalho é a primeira abordagem para a resolução de correferência de sintagmas nominais de qualquer tipo para a língua portuguesa. Outros trabalhos apresentam soluções restritas à resolução anafórica pronominal como os trabalhos apresentados em (COELHO, 2005) e (CHAVES, 2007).

Esse sistema foi utilizado em um dos experimentos realizados neste trabalho, seus resultados são discutidos na seção 6.3 do capítulo 6.

4.6 ROUGE

A ROUGE (LIN, 2004) é um sistema de pacote de medidas implementado para realizar avaliação de informatividade de sumários de forma automática. A vantagem da utilização de um sistema automático está no baixo custo, se comparado com uma avaliação manual, e de fácil reprodução. A ferramenta Rouge é independente de língua, sendo possível aplicá-la para o português.

O sistema utiliza como base um sumário de referência para ser comparado com

o sumário automático. Para realizar a análise, o sistema usa co-ocorrência de n-gramas, isto é verifica se as palavras do sumário de referência ocorrem no sumário automático.

O pacote de medidas da ROUGE oferece cinco medidas:

- Rouge-1: utiliza a contagem de co-ocorrência de unigramas para avaliação da informatividade.
- Rouge-2: verifica a freqüência de cada par de palavras (bigramas) na comparação entre os sumários.
- Rouge-3 e Rouge-4: semelhante às outras medidas, mas utilizam a comparação de 3-grama e 4-grama para verificação de co-ocorrência. Não são muito utilizadas, pois é incomum acontecer seqüências de 3 e 4 palavras entre os sumários automático e referência.
- Rouge-L: localiza as maiores seqüências entre os dois sumários, realizando uma avaliação similar à co-ocorrência de n-gramas.

Dentre essas medidas, a mais utilizada é a Rouge-1. O sistema fornece precisão, cobertura e F-measure para cada texto processado e a média em relação a um conjunto de textos processados.

4.7 Descrição do Corpus Summ-it

O corpus Summ-it (COLLOVINI et al., 2007), utilizado neste estudo, constitui-se de 50 textos jornalísticos da Folha de São Paulo, retirados do caderno de ciências do jornal, escritos em português do Brasil. O mesmo foi disponibilizado através do Projeto PLN-BR[2].

[2]http://www.nilc.icmc.usp.br:8180/portal/

O corpus foi processado pelo parser PALAVRAS (BICK, 2000), para extrair informações morfossintáticas e anotado, manualmente, com informações de correferência, utilizando-se a ferramenta MMAX que foi descrita na seção 4.2.

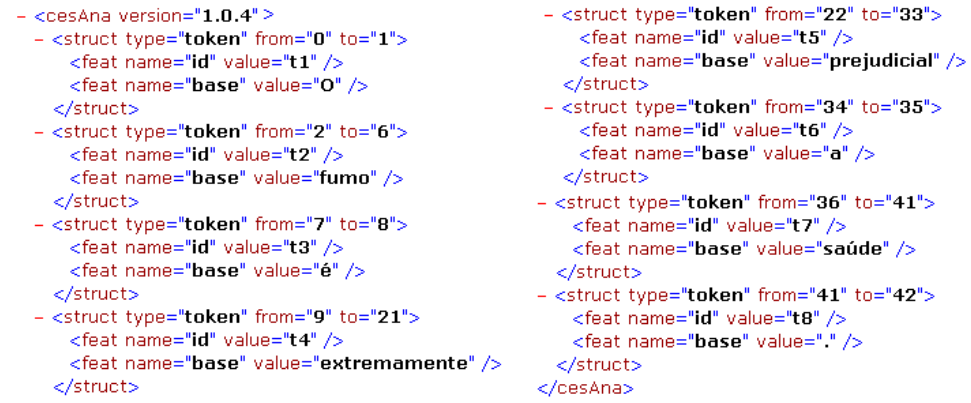

Figura 17: Arquivo de Tokens

Após o processamento dos textos no PALAVRAS, usou-se um conversor desenvolvido no Laboratório de Engenharia da Linguagem para a geração de três arquivos XML. O arquivo que aparace na Figura 17 mostra o arquivo XML com as palavras (tokens). O segundo arquivo possui as informações de *part-of-speech*, Figura 18, e o último arquivo apresenta as informações de sintaxe do texto, como mostra a Figura 19.

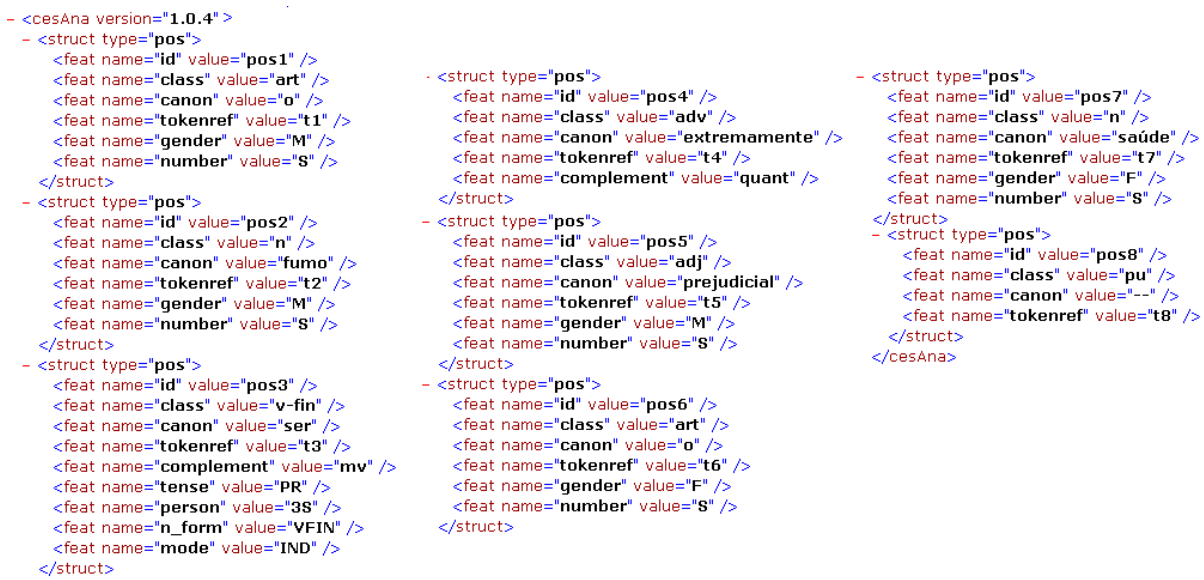

Figura 18: Arquivo de *part-of-speech*

Os arquivos XML, com as informações lingüísticas, estão em três arquivos separados, pois foi utilizado o princípio da separabilidade e foi mantida uma codificação básica para todos arquivos utilizando o princípio de uniformidade. O modelo adotado para armazenamento dessas informações é o XCES[3]. Esse padrão foi também adotado pelo Projeto PLN-BR. As únicas *tags* utilizadas são as *tags* <struct> e <feature> e elas são responsáveis por toda estrutura de armazenamento de informações lingüísticas.

```
- <cesAna version="1.0.4" xmlns="http://www.xces.org/schema/2003">
  - <struct type="phrase" from="t1" to="t7">
      <feat name="id" value="phr1" />
      <feat name="cat" value="s" />
    </struct>
  - <struct type="phrase" from="t1" to="t7">
      <feat name="id" value="phr2" />
      <feat name="cat" value="fcl" />
      <feat name="function" value="STA" />
    </struct>
  - <struct type="phrase" from="t1" to="t2">
      <feat name="id" value="phr3" />
      <feat name="cat" value="np" />
      <feat name="function" value="S" />
      <feat name="head" value="t2" />
    </struct>
  - <struct type="phrase" from="t4" to="t5">
      <feat name="id" value="phr4" />
      <feat name="cat" value="adjp" />
      <feat name="function" value="Cs" />
      <feat name="head" value="t5" />
    </struct>
  - <struct type="phrase" from="t6" to="t7">
      <feat name="id" value="phr5" />
      <feat name="cat" value="np" />
      <feat name="function" value="fCs" />
      <feat name="head" value="t7" />
    </struct>
  </cesAna>
```

Figura 19: Arquivo de informações de sintaxe

Cada texto do corpus possui um sumário manual feito por sumarizadores humanos (COELHO, 2007). O corpus conta, também, com relatórios HTML de cadeias de correferência, conforme ilustrado na Figura 20.

O corpus Summ-it possui um total de 5047 sintagmas nominais, compondo 560 cadeias de correferência, sendo que a cadeia mais extensa possui 16 elementos. A tabela 1 ilustra os resultados da anotação das descrições definidas, seguindo a classificação apresentada na seção 2.2.2.

[3]http://www.cs.vassar.edu/XCES/

Relatorio Correferencia: CIENCIA_2004_6480.txt

	Classificação	Sintagma
CADEIA : set_18		
word_1..word_4	---new	O ministro Roberto_Rodrigues (Agricultura)
word_35	---	Rodrigues
word_233	---	Rodrigues
CADEIA : set_12		
word_7..word_12	---new	o nascimento de a bezerra Vitoriosa
word_162..word_171	indirect---old	o resultado de um experimento realizado por a Embrapa (Empresa_Brasileira_de_Pesquisa_Agropecuária)
CADEIA : set_21		
word_10..word_12	---new	a bezerra Vitoriosa
word_14..word_15	indirect---old	O animal
word_17..word_25	---	um clone gerado a_partir_de um clone -a vaca Vitória
word_37..word_38	indirect---old	a cria
word_160	---	Vitoriosa
word_173	---	Ela
word_206..word_210	indirect---old	"O clone de o clone
word_254	---	Vitoriosa
word_262..word_273	---new	a terceira tentativa de o órgão de criar um clone a_partir_de outro

Figura 20: Arquivo HTML com informações com as cadeias de correferência

Tabela 1: Anotação de classificação do corpus Summ-it

Classificações	**Quantidades**
Novas no Discurso	1428
Anáforas Associativas	183
Anáforas Diretas	407
Anáforas Indiretas	291
Total de descrições definidas classificadas:	**2309**

Neste capítulo foram apresentadas as ferramentas PALAVRAS, MMAX, ROUGE, GISTSUMM e SUPOR-2, que são importantes no contexto deste trabalho para os processos de criação do corpus, geração e avaliação de sumários automáticos.

Capítulo 5

Sistema CorrefSum

Neste capítulo serão abordadas questões relativas à implementação do sistema, detalhando cada um dos módulos desenvolvidos.

5.1 Visão Geral

O sistema, desenvolvido neste trabalho, tem como objetivo realizar a correção da coesão referencial dos sumários, usando como fonte de informação as informações das cadeias de correferência presentes no texto-fonte.

A solução foi implementada utilizando a linguagem de programação Java[1], que permite tanto a portabilidade entre sistemas operacionais, quanto torna o sistema implementado livre para distribuição.

O sistema implementado possui 4 grandes módulos: módulo leitura de arquivos, módulo de processamento de informações, módulo de revisão dos sumários e módulo de interface.

Na Figura 21, temos uma visão geral do sistema. Os módulos de processamento de informações e revisão dos sumários são considerados módulos principais deste sistema.

[1]http://java.sun.com/

Nas próximas seções, os módulos componentes deste sistema serão detalhados.

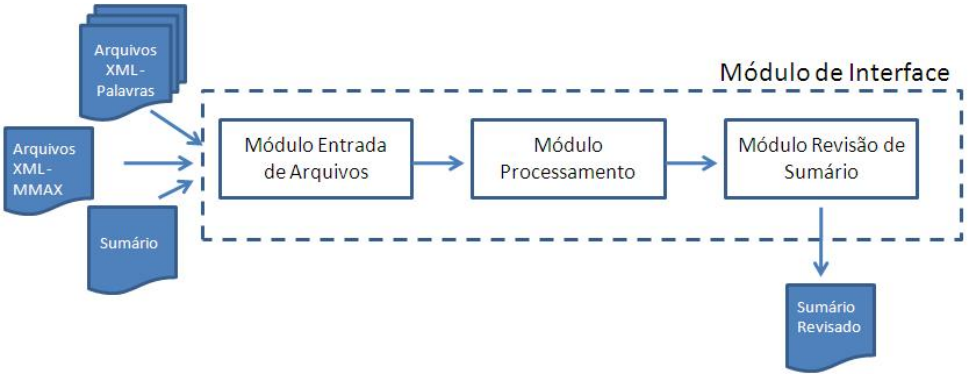

Figura 21: Visão geral do sistema

5.2 Módulo de Leitura do Arquivo

O objetivo deste módulo é realizar a leitura e o armazenamento das informações referentes aos sumários, os quais serão corrigidos em estruturas de dados que possibilitem uma rapidez de processamento.

O sistema recebe cinco arquivos de entrada para cada sumário que será processado. Três arquivos são fornecidos pelo parser PALAVRAS (BICK, 2000) (citado na seção 4.1), contendo a análise sintática do texto-fonte. O quarto arquivo é resultado da anotação manual das cadeias de correferência, usando o *software* MMAX (MüLLER; STRUBE, 2001), (citado na seção 4.2). O quinto arquivo é o sumário produzido por um sumarizador automático. É importante salientar que a geração desses arquivos é considerada, neste trabalho, etapa de pré-processamento.

Os arquivos fornecidos pelo PALAVRAS e pelo MMAX são arquivos em formatos XML. O padrão XML utilizado para armazenamento das informações lingüísticas é o XCES, tal como proposto pelo projeto PLN-BR[2]. O formato XCES foi escolhido por permitir o armazenamento de vários níveis lingüísticos com o mesmo formato de anotação, tornando o processamento ágil e rápido. É importante destacar que uma saída em formato

[2]http://www.nilc.icmc.usp.br:8180/portal/

XCES não é fornecida pelos sistemas PALAVRAS e MMAX. Para isso, foi implementado um conversor de formatos pela equipe do Laboratório da Engenharia da Linguagem da Unisinos. O arquivo de saída do sumarizador automático está em formato texto (.TXT). Ele é utilizado no formato produzido pelo sumarizador.

5.3 Módulo Processamento das Informações

Esse módulo é responsável por localizar as sentenças do texto-fonte que foram incluídas no sumário, selecionar todas as cadeias de correferência presentes (relativas aos sintagmas nominais presentes no sumário) e realizar a pontuação de cada elemento dessas cadeias.

Vejamos um exemplo de um texto na Figura 22 e um sumário na Figura 23.

[S1]Ao contrário do que muita gente pensa, a internet não está reduzindo os contatos entre as pessoas , nem substituindo-os por relações impessoais conduzidas por computador. [S2]A conclusão é de Barry Ellman, do Centro para Estudos Urbanos e Comunitários da Universidade de Toronto, Canadá. [S3]Segundo o pesquisador, os contatos via redes de computadores estão, na verdade, ampliando a socialização das pessoas. [S4]Um dos exemplos que ele apresenta é o de um estudo feito em um subúrbio de Toronto, segundo o qual as pessoas "plugadas" em uma rede local conheciam três vezes mais vizinhos do que os não-conectados. [S5]Além disso, vizinhos conectados se encontraram pessoalmente 60% mais do que os excluídos da rede. [S6]Os números gerais da internet apontam o mesmo fenômeno, diz Ellman. [S7]Segundo ele, pessoas ligadas por computadores tiveram mais contatos pessoais com seus amigos e parentes do que pessoas não-conectadas. [S8]O artigo do pesquisador está na edição de hoje da revista "Science".

Figura 22: Texto CIENCIA_2001_6410

Segundo ele, pessoas ligadas por computadores tiveram mais contatos pessoais com seus amigos e parentes do que pessoas não-conectadas.

Figura 23: Sumário gerado pelo Gistsumm do texto CIENCIA_2001_6410.

No exemplo, a sentença 7 (S7) do texto-fonte foi incluída no sumário. Após a localização da sentença, é iniciado o processo de seleção de todas as cadeias que fazem parte do sumário, o sistema procura todos os sintagmas nominais que estão dentro do intervalo da sentença. Seguimos no mesmo exemplo do texto CIENCIA_2001_6410, na Figura 24 temos a representação em XML, que define o intervalo de palavras pertencente a essa sentença.

```
<struct type=''phrase'' from=''word_136'' to=''word_155'' >
  <feat name=''id'' value=''phr7'' />
  <feat name=''cat'' value=''s'' />
</struct>
```

Figura 24: Trecho do arquivo XML-Phrases do texto CIENCIA_2001_6410

```
...
<struct type=''markable'' from=''word_137'' to=''word_137'' >
  <feat name=''id'' value=''markable_42'' />
  <feat name=''np_n'' value=''no'' />
  <feat name=''member'' value=''set_14'' />
  </struct>
<struct type=''markable'' from=''word_139'' to=''word_142'' >
  <feat name=''id'' value=''markable_43'' />
  <feat name=''np_n'' value=''yes'' />
  <feat name=''np_form'' value=''bare-np'' />
  </struct>
<struct type=''markable'' from=''word_142'' to=''word_142'' >
  <feat name=''id'' value=''markable_44'' />
  <feat name=''np_n'' value=''yes'' />
  <feat name=''np_form'' value=''bare-np'' />
  <feat name=''member'' value=''set_11'' />
  </struct>
 <struct type=''markable'' from=''word_144'' to=''word_151'' >
  <feat name=''id'' value=''markable_45'' />
  <feat name=''np_n'' value=''yes'' />
  <feat name=''np_form'' value=''quant-np'' />
  </struct>
...
```

Figura 25: Trecho do arquivo XML-Markables do texto CIENCIA_2001_6410

Com base nesse exemplo (Figura 24), observamos que a sentença 7, identificada pelo id="phr7", é composta pelo conjunto de palavras que estão no intervalo de id="word_136", até a palavra com id= "word_155". Para localizar todos os markables, que pertencem a essa sentença, é necessário selecionar todos os markables neste mesmo intervalo de palavras("words"), usando as informações do arquivo XML-Markables fornecido pelo MMAX. A Figura 25 mostra a representação em XML dessas informações.

Observamos, na Figura 25, que os markables com os seguintes valores de id´s: "markables_42", "markables_43", "markables_44" e "markables_45" foram selecionados, pois seu intervalo está contido dentro do intervalo de palavras da frase selecionada. No próximo passo, os markables "markables_43" e "markables_45" são desconsiderados, pois não fazem parte de nenhuma cadeia de correferência. Isso pode ser verificado por não possuírem o atributo "member" como informação. O sistema seleciona os markables "markables_42" e "markables_44" identificando então a presença de dois elementos

pertencentes a cadeias de correferência nesse sumário. Essas cadeias são definidas pelos atributos "member", que informam que as cadeias de correferência "set_14" e "set_11" deverão ser analisadas pelo sistema. Esse processo é repetido para cada uma das frases do sumário.

Após a identificação das cadeias de correferência, é realizada a localização de todos elementos pertencentes a cada uma das cadeias. Essa informação também é fornecida pelo mesmo arquivo XML de markables. A Figura 26 ilustra a localização de todos os membros das cadeias "set_14" e "set_11".

Identificados todos os termos de cada uma das cadeias pertencentes ao sumário, o sistema realiza a pontuação para cada membro da cadeia. Esse sistema de pontuação foi inspirado no trabalho de Mitkov em (MITKOV, 1998), no entanto, os critérios foram definidos pela autora do trabalho. A pontuação utilizada segue os seguintes critérios:

1. Nome Próprio: é atribuído 1 ponto se o sintagma nominal possuir algum nome próprio. Para selecionar essa informação, é usado o arquivo de *part-of-speech* (POS) fornecido pelo Palavras. As palavras do texto são etiquetadas com a etiqueta pos="prop" indicando que a palavra é um nome próprio.

2. Maior: é atribuído 1 ponto caso o sintagma seja o maior sintagma da sua cadeia, em número de caracteres.

3. Primeiro: é atribuído 1 ponto caso o sintagma seja o primeiro elemento de sua cadeia.

4. Aposto: é atribuído 1 ponto caso o sintagma possua vírgulas (geralmente usada como marca de aposto).

Todos os markables, selecionados na etapa anterior, são pontuados com base nessas características. Os pontos são cumulativos e servirão como critério de seleção do elemento de cadeia que deverá ser escolhido para substituir o termo no sumário.

```
....
<struct type=''markable'' from=''word_32'' to=''word_45'' >
  <feat name=''id'' value=''markable_54'' />
   <feat name=''np_form'' value=''pn'' />
  <feat name=''member'' value=''set_14'' />
  <feat name=''text'' value=''Barry Ellman, do Centro para Estudos Urbanos e
                          Comunitários da Universidade de Toronto, Canadá'' />
 </struct>
<struct type=''markable'' from=''word_159'' to=''word_160'' >
  <feat name=''id'' value=''markable_49'' />
  <feat name=''np_form'' value=''def-np'' />
  <feat name=''member'' value=''set_14'' />
  <feat name=''text'' value=''o pesquisador'' />
 </struct>
<struct type=''markable'' from=''word_73'' to=''word_73'' >
  <feat name=''id'' value=''markable_25'' />
  <feat name=''member'' value=''set_14'' />
  <feat name=''text'' value=''ele'' />
  </struct>
<struct type=''markable'' from=''word_134'' to=''word_134'' >
  <feat name=''id'' value=''markable_41'' />
  <feat name=''np_form'' value=''pn'' />
  <feat name=''member'' value=''set_14'' />
  <feat name=''text'' value=''Ellman'' />
  </struct>
<struct type=''markable'' from=''word_137'' to=''word_137'' >
  <feat name=''id'' value=''markable_42'' />
  <feat name=''member'' value=''set_14'' />
  <feat name=''text'' value=''ele'' />
  </struct>
<struct type=''markable'' from=''word_159'' to=''word_160'' >
  <feat name=''id'' value=''markable_49'' />
  <feat name=''np_form'' value=''def-np'' />
  <feat name=''member'' value=''set_14'' />
  <feat name=''text'' value=''o pesquisador'' />
 </struct>

<struct from=''word_56'' to=''word_56'' type=''markable''>
  <feat name=''id'' value=''markable_19'' />
  <feat name=''np_form'' value=''bare-np'' />
  <feat name=''member'' value=''set_11'' />
  <feat name=''text'' value=''computadores'' />
  </struct>

<struct from=''word_142'' to=''word_142'' type=''markable''>
  <feat name=''id'' value=''markable_44'' />
  <feat name=''np_form'' value=''bare-np'' />
  <feat name=''member'' value=''set_11'' />
  <feat name=''text'' value=''computadores''/>
  </struct>
...
```

Figura 26: Identificação de todos os termos das duas cadeias que aparecem no texto CIENCIA_2001_6410

Ainda usando o texto CIENCIA_2001_6410 como exemplo, mostramos, abaixo, o sistema de pontos para cada elemento das cadeias presentes no sumário.

1. Cadeia “set_14”

    - Barry Ellman, do Centro para Estudos Urbanos e Comunitários da Universidade de Toronto, Canadá - 4 pontos
    - o pesquisador - 0 pontos
    - ele - 0 pontos
    - Ellman -1 ponto
    - ele - 0 pontos
    - o pesquisador - 0 pontos

2. Cadeia “set_11”

    - Computadores - 1 ponto
    - Computadores - 0 pontos

Podemos observar que o sintagma “Barry Ellman, do Centro para Estudos Urbanos e Comunitários da Universidade de Toronto, Canadá” foi pontuado com 4 pontos, pois ele possui as 4 características pontuadas (nome próprio, maior, primeiro e aposto); o sintagma “Ellman” foi pontuado com 1 ponto, por ser nome-próprio. Em relação à outra cadeia, o primeiro sintagma “computadores” recebeu 1 ponto apenas, por ser o primeiro; o outro sintagma “computadores” não foi pontuado, pois não possui nenhuma das características.

Essas informações serão utilizadas pelo módulo de revisão dos sumários, que utilizará a informação dos pontos para selecionar o melhor elemento da cadeia para fazer a substituição.

Caso ocorra uma próxima frase no sumário, em que apareça um sintagma de uma cadeia de correferência que foi tratada pelo algoritmo descrito acima, este sintagma não

é substituído. Tomemos a sentença 8, do texto CIENCIA_2001_6410 ([S8] da Figura 22), como exemplo: *"O artigo do pesquisador está na edição de hoje da revista Science."* Se essa sentença estivesse incluída no sumário, o sistema desconsideraria o tratamento da cadeia do elemento "o pesquisador", pois essa cadeia já teria sido tratada anteriormente, quando foi analisada a cadeia do elemento "ele" na sentença 7.

5.4 Módulo de Revisão dos Sumários

O módulo de revisão dos sumários tem como objetivo substituir sintagmas nominais pela melhor expressão, ou a mais informativa de sua cadeia, gerando um novo sumário (sumário revisado), idealmente, sem quebras das cadeias de correferência.

Para escolher a melhor expressão, o sistema utiliza a pontuação gerada pelo módulo de processamento de informações, levando em conta a taxa de compressão do sumário. Os sumários originais, gerados pelo sumarizador automático, foram selecionados com uma taxa de compressão de 70%, gerando sumários de 30% em relação ao texto-fonte. Em favor da coesão referencial, neste trabalho estamos considerando uma taxa de compressão superior, podendo, o sumário revisado, chegar a 40% (com aplicação deste sistema) com as trocas de expressões.

Ainda tomando o texto CIENCIA_2001_6410 como exemplo, o sistema deve verificar qual expressão dentro da cadeia de correferência deverá ser substituída. Nesse momento, o critério de seleção da melhor expressão será decidido através da pontuação. O elemento da cadeia que tiver o maior número de pontos será selecionado. Com base nos pontos das cadeias do sumário, o elemento "ele" no sumário original deverá ser substituído pelo sintagma nominal "Barry Ellman, do Centro para Estudos Urbanos e Comunitários da Universidade de Toronto, Canadá", gerando um sumário revisado. Essa geração do sumário revisado também é feita por esse módulo, que gera um arquivo no formato texto. Na Figura 27, temos o sumário revisado já com a substituição dos termos.

Segundo Barry Ellman, do Centro para Estudos Urbanos e Comunitários da Universidade de Toronto, Canadá, pessoas ligadas por computadores tiveram mais contatos pessoais com seus amigos e parentes do que pessoas não-conectadas.

Figura 27: Sumário revisado do texto CIENCIA_2001_6410

Com relação à taxa de compressão, podemos observar que o sumário original do texto CIENCIA_2001_6410, gerado pelo sistema GistSumm, possui uma taxa de compressão de 12%[3]). Esse sumário, após a troca da expressão, ficou com uma taxa de compressão de 22%.

O módulo de revisão dos sumários é responsável por resguardar a taxa de compressão configurada pelo sistema. Caso o sumário revisado ultrapasse a taxa, foi desenvolvido um algoritmo que seleciona apenas a primeira parte do sintagma (quando o sintagma tiver vírgula). Por exemplo, se o sumário do texto CIENCIA_2001_6410 estivesse com a taxa de compressão no limite, o sintagma "Barry Ellman, do Centro para Estudos Urbanos e Comunitários da Universidade de Toronto, Canadá" seria selecionado somente até a parte que antecede a vírgula: "Barry Ellman". O sistema elimina, também, os parênteses, caso eles apareçam no sintagma escolhido para substituição.

5.5 Módulo de Interface

Foi desenvolvida uma interface que permite ao usuário operar o sistema de forma fácil e interativa. A interface pode ser divida em duas partes. A primeira parte possibilita a seleção dos arquivos para processamento. Nela, o sistema permite a parametrização da taxa máxima de compressão, que será respeitada no momento da troca da melhor expressão da cadeia. Há, também, a escolha da substituição automática pelo algoritmo ou a seleção de manipulação manual das substituições. A Figura 28 mostra a interface do sistema exibindo a primeira parte da seleção dos arquivos.

[3]Os sumários foram gerados pelo Gistsumm com taxa de compressão de 70%, mas o sistema tem um dispositivo de segurança que impede o acréscimo de mais sentenças no sumário, caso seja gerado um sumário maior que a taxa de compressão desejada.

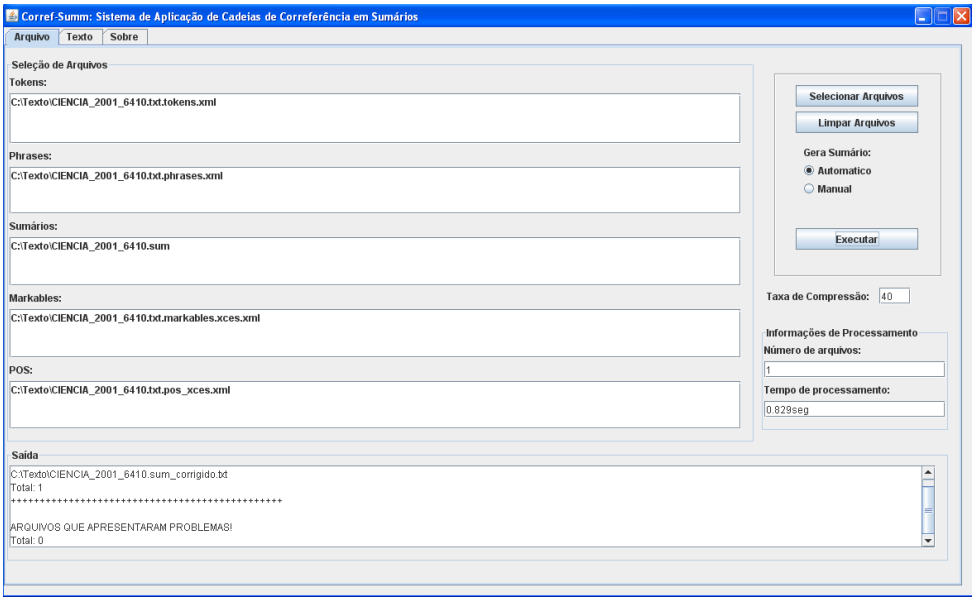

Figura 28: Interface do sistema - seleção dos arquivos.

A segunda parte da interface permite que o usuário do sistema efetue as trocas das cadeias de correferência de forma manual. O usuário clica no botão (ao lado do sintagma nominal) e na metade direita da interface aparecem as opções (todos os elementos daquela cadeia) que poderão ser usadas para efetuar a troca das expressões. O usuário tem a total liberdade de escolher a melhor expressão, conforme sua análise pessoal. A Figura 29 traz a ilustração da interface com a possibilidade de troca manual de expressões.

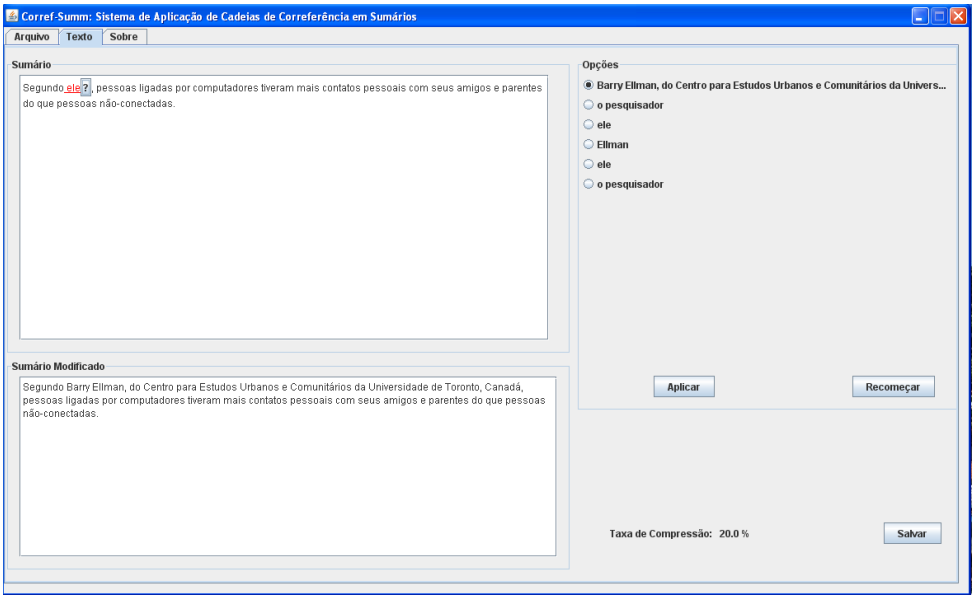

Figura 29: Interface do sistema - troca de expressões e análise das cadeias manualmente.

Quando o usuário escolhe a opção automática (Figura 28), no campo “sumário revisado” da tela (Figura 29), aparece o sumário revisado automaticamente. O sistema permite reiniciar as trocas das expressões quantas vezes o usuário desejar. Caso a opção manual seja selecionada, esse campo da tela será preenchido com o sumário original e ficará aguardando as substituições propostas pelo usuário.

Este capítulo apresentou o sistema desenvolvido para realizar a pesquisa sobre cadeias de correferência, sumarização automática e coesão referencial proposta nesta dissertação. Os experimentos realizados com o sistema são descritos na próxima seção.

Capítulo 6

Experimentos e Avaliação

Neste capítulo apresentamos os experimentos realizados utilizando o sistema CorrefSum descrito no capítulo 5. Para os experimentos utilizamos dois sumarizadores extrativos: GistSumm e o Supor-2.

Os experimentos foram realizados com base no corpus Summ-it (descrito na seção 4.7). Para o experimento inicial, o corpus foi dividido em duas partes, uma parte foi utilizada no desenvolvimento das heurísticas e a outra parte foi usada para teste dessas heurísticas. Durante a fase de desenvolvimento, os textos e os sumários gerados pela ferramenta GistSumm foram observados e serviram como base de estudo para o desenvolvimento de heurísticas de troca de expressões de elementos da mesma cadeia. Foram considerados 30 textos para desenvolvimento das heurísticas e na fase de teste foram utilizados 20 textos. Os sumários gerados pelo sistema SuPor-2 foram utilizados de forma integral, pois as heurísticas já estavam desenvolvidas e testadas.

Os seguintes itens foram analisados para cada texto:

- Quantidade de cadeias no texto
- Quantidade de cadeias no sumário
- Quantidade de trocas efetuadas

- Taxa de compressão antes da troca
- Taxa de compressão depois da troca

Para avaliar os sumários revisados gerados pelo CorrefSum, utilizamos duas formas de avaliação: avaliação automática e avaliação subjetiva. Para avaliação automática, foi utilizada a medida de avaliação Rouge (LIN, 2000). A Rouge utiliza como parâmetro de comparação um sumário de referência. Para isso, usamos os sumários gerados manualmente por sumarizadores profissionais, conforme descrito em (COELHO, 2007). A Rouge fornece as seguintes medidas: Precisão, Cobertura e F-Measure. Essas medidas são disponibilizadas para cada texto individualmente e também pelo conjunto de textos processados. Neste trabalho, estamos usando a Rouge-1 que usa a comparação por unigramas para avaliação. A Rouge-1 é a medida que tem sido utilizada e aceita por vários trabalhos nessa área, como por exemolo, os trabalhos de (CARBONEL, 2007), (LEITE et al., 2007) e (FILHO; PARDO; NUNES, 2007).

Para a avaliação subjetiva, foram escolhidos 10 textos originais e revisados pelo CorrefSum que apresentaram uma maior diferença de desempenho para a medida ROUGE. Para essa avaliação, usamos 5 juízes humanos, falantes nativos da língua e especialistas na lígua portuguesa. A avaliação foi separada em duas partes: na seção 6.1.3, temos os resultados da avaliação dos sumários do GistSumm e na seção 6.2.3, a avaliação dos sumários do Supor-2. Os juízes responderam a um questionário sobre os sumários extrativos com e sem revisão, com objetivo de avaliar informatividade e legibilidade. Os sumários presentes no questionário não foram identificados como original e corrigido com a intenção de não influenciar na resposta do avaliador. Os questionários encontram-se nos Anexos A e B.

Este capítulo traz, ainda, uma análise qualitativa das substituições realizadas pelo GistSumm no corpus Summ-it e, na última seção, temos algumas discussões sobre questões relacionadas à implementação do sistema.

6.1 Experimentos e Avaliação - GistSumm

Nesta seção, são descritos os experimentos, avaliação automática e subjetiva, realizados com o sumarizador GistSumm.

6.1.1 Experimento

Os sumários gerados pelo GistSumm a partir do corpus Summ-it foram divididos em duas partes, uma para estudo e análise, utilizada no desenvolvimento das heurísticas de troca de expressões, ao qual chamamos conjunto de treino, e uma segunda parte para testes das heurísticas. Na Tabela 2, temos os resultados para o conjunto de treino onde foram selecionados 30 sumários do corpus Summ-it.

Tabela 2: Resultados do conjunto de treino do Summ-it

NOME DO TEXTO	QTDE CADEIAS NO TEXTO	QTDE CADEIAS NO SUMÁRIO	QTDE. TROCAS	TX.COMPRESSÃO ANTES (%)	TX.COMPRESSÃO DEPOIS (%)
CIENCIA_2000_17082	10	6	1	30	30
CIENCIA_2000_17088	11	7	2	27	29
CIENCIA_2000_17101	17	8	1	29	29
CIENCIA_2000_17108	9	6	1	30	36
CIENCIA_2000_17109	12	8	4	19	26
CIENCIA_2000_17112	9	6	3	22	28
CIENCIA_2000_17113	16	9	3	23	26
CIENCIA_2001_19858	11	6	2	28	31
CIENCIA_2002_22005	12	6	2	30	35
CIENCIA_2002_22010	10	6	1	28	32
CIENCIA_2002_22015	16	7	4	27	28
CIENCIA_2002_22023	12	8	4	21	27
CIENCIA_2002_22027	22	10	3	26	28
CIENCIA_2002_22029	19	15	3	30	36
CIENCIA_2003_24212	19	10	2	27	27
CIENCIA_2003_24219	13	8	1	27	28
CIENCIA_2003_24226	15	8	3	28	30
CIENCIA_2004_26415	6	4	1	16	17
CIENCIA_2004_26417	14	6	2	30	33
CIENCIA_2004_26423	24	13	3	28	29
CIENCIA_2004_26425	21	12	3	26	30
CIENCIA_2005_28743	9	4	1	27	28
CIENCIA_2005_28747	11	7	4	16	21
CIENCIA_2005_28752	14	7	0	27	27
CIENCIA_2005_28754	15	10	1	27	29
CIENCIA_2005_28755	14	12	3	27	28
CIENCIA_2005_28756	13	7	3	24	34
CIENCIA_2005_28764	12	8	1	33	34
CIENCIA_2005_28766	23	17	3	27	30
CIENCIA_2005_28774	21	11	2	25	27
SOMA	430	252	67	-	-
MÉDIA	14,33	8,4	2,23	26,16	29,1

Podemos observar que, em média, os textos-fonte possuem 14,33 cadeias de correferência. Por sua vez, 8,4 dessas cadeias aparecem nos sumários. Como resultado do processamento, obtivemos um total de 67 trocas, com 2,23 trocas por texto, em média. Das 252 cadeias analisadas, 185 não necessitaram de troca, pois os elementos da cadeia, avaliados como melhores pelas heurísticas, já estavam contidos no sumário. O número máximo de trocas efetuadas em um único sumário foi 4, houve também 1 caso em que nenhuma troca foi necessária.

A taxa de compressão média dos sumários originais gerados pelo GistSumm foi de 26,16% e os sumários revisados após a aplicação do sistema CorrefSum obtiveram uma

média de taxa de compressão de 29,10%. Observamos que os sumários aumentaram em média 3% em relação ao seu tamanho original.

Na Tabela 3 temos os resultados para os 20 sumários restantes do corpus Summ-it usados como base de teste para o sistema. Esses textos, em média possuem 7,8 cadeias de correferência e os sumários possuem 3,9 diferentes cadeias, em média. Em relação a trocas de expressões observamos um total de 22 trocas tendo em média 1,1 trocas por texto, 55 expressões analisadas não necessitaram de troca. Em relação à taxa de compressão observamos uma média de 24,1% e após o processamento pelo sistema CorrefSum os sumários alcançaram uma média de 27,25% de taxa de compressão em relação ao seu texto-fonte.

Tabela 3: Resultados do conjunto de teste do Summ-it

NOME DO TEXTO	QTDE CADEIAS NO TEXTO	QTDE CADEIAS NO SUMÁRIO	QTDE. TROCAS	TAXA COMPRESSÃO ANTES (%)	TAXA COMPRESSÃO DEPOIS (%)
CIENCIA_2000_6380	10	4	0	26	26
CIENCIA_2000_6381	11	5	1	21	25
CIENCIA_2000_6389	9	2	1	11	16
CIENCIA_2000_6391	6	3	1	24	30
CIENCIA_2001_6406	6	1	1	19	21
CIENCIA_2001_6410	8	2	1	12	20
CIENCIA_2001_6414	8	4	2	18	18
CIENCIA_2001_6416	8	3	2	26	35
CIENCIA_2001_6423	3	2	0	40	40
CIENCIA_2002_6441	5	3	2	25	31
CIENCIA_2003_6457	9	8	1	28	30
CIENCIA_2003_6465	11	5	0	25	25
CIENCIA_2003_6472	4	3	1	36	39
CIENCIA_2004_6480	10	4	2	27	33
CIENCIA_2004_6488	5	4	0	27	27
CIENCIA_2004_6494	8	3	1	17	19
CIENCIA_2005_6507	6	6	1	27	29
CIENCIA_2005_6514	8	4	1	23	23
CIENCIA_2005_6515	11	7	1	29	29
CIENCIA_2005_6518	10	5	3	21	29
SOMA	156	78	22	-	-
MÉDIA	7,8	3,9	1,1	24,1	27,25

Com base nesses resultados observamos que houve trocas em grande parte dos sumários, indicando a possibilidade de problemas na coesão referencial ou a existência de uma expressão mais completa que pudesse tornar o texto mais informativo.

6.1.2 Avaliação Rouge dos Sumários do GistSumm Revisados

A Tabela 4 mostra os resultados da medida ROUGE para os sumários originais gerados pelo GistSumm e os sumários corrigidos. A avaliação automática usa como base os sumários de referência (COELHO, 2007), construídos manualmente por sumarizadores humanos.

Tabela 4: Dados Rouge - Sumários Originais GistSumm e Sumários Corrigidos- Dados de treino

	GISTSUMM-ORIGINAL			GISTSUMM-CORRIGIDO		
NOME TEXTO	COBERTURA	PRECISÃO	F-MEASURE	COBERTURA	PRECISÃO	F-MEASURE
CIENCIA_2000_17082	61,36	56,84	59,02	61,36	57,45	59,34
CIENCIA_2000_17088	38,71	38,30	38,50	41,94	39,80	40,84
CIENCIA_2000_17101	45,55	45,10	45,32	48,52	47,57	48,04
CIENCIA_2000_17108	56,52	57,14	56,83	58,70	49,09	53,47
CIENCIA_2000_17109	29,21	46,43	35,86	40,45	48,00	43,90
CIENCIA_2000_17112	22,00	29,33	25,14	33,00	34,02	33,50
CIENCIA_2000_17113	45,04	48,36	46,64	31,30	38,68	34,60
CIENCIA_2001_19858	55,37	64,90	59,76	59,32	63,64	61,40
CIENCIA_2002_22005	57,31	64,05	60,49	73,68	68,85	71,19
CIENCIA_2002_22010	67,90	78,18	72,68	67,90	70,49	69,17
CIENCIA_2002_22015	47,89	57,63	52,31	53,05	57,95	55,39
CIENCIA_2002_22023	25,20	32,98	28,57	34,96	35,83	35,39
CIENCIA_2002_22027	62,75	70,59	66,44	65,36	69,93	67,57
CIENCIA_2002_22029	67,14	71,50	69,25	77,93	68,60	72,97
CIENCIA_2003_24212	66,46	69,43	67,91	64,02	66,88	65,42
CIENCIA_2003_24219	54,13	53,15	53,64	55,05	51,72	53,33
CIENCIA_2003_24226	45,18	52,66	48,63	50,76	53,76	52,22
CIENCIA_2004_26415	29,13	60,00	39,22	33,98	63,64	44,30
CIENCIA_2004_26417	34,57	37,58	36,01	40,74	40,49	40,62
CIENCIA_2004_26423	44,49	55,56	49,41	48,31	55,61	51,70
CIENCIA_2004_26425	57,14	71,80	63,64	63,27	68,13	65,61
CIENCIA_2005_28743	54,61	61,94	58,04	53,95	60,29	56,94
CIENCIA_2005_28747	20,88	38,78	27,14	32,97	45,46	38,22
CIENCIA_2005_28752	46,63	54,29	50,17	46,01	53,19	49,34
CIENCIA_2005_28754	55,83	89,15	68,66	64,56	88,67	74,72
CIENCIA_2005_28755	59,55	58,89	59,22	61,80	58,20	59,95
CIENCIA_2005_28756	41,29	50,79	45,55	54,84	47,22	50,75
CIENCIA_2005_28764	47,76	55,81	51,47	52,24	51,22	51,72
CIENCIA_2005_28766	51,50	63,98	57,06	60,50	65,05	62,69
CIENCIA_2005_28774	49,28	54,84	51,91	55,56	55,83	55,69
MÉDIA	48,41	56,60	51,83	53,15	56,03	54,25

Observamos que os sumários do GistSumm obtiveram 48,41% de cobertura em relação ao sumário de referência, e com aplicação do CorrefSum, a cobertura passou para 53,15%. Em relação à precisão, o valor passou de 56,60% para 56,03%. A medida F-measure que mostra uma média harmônica entre precisão e cobertura que passou de

51,83% para 54,25%.

A Tabela 5 mostra os dados obtidos com os sumários originais e sumários revisados em relação aos outros 20 textos do corpus Summ-it.

Tabela 5: Dados Rouge - Sumários Originais GistSumm e Sumários Corrigidos- Dados de teste

	GISTSUMM-ORIGINAL			GISTSUMM-CORRIGIDO		
NOME TEXTO	COBERTURA	PRECISÃO	F-MEASURE	COBERTURA	PRECISÃO	F-MEASURE
CIENCIA_2000_6380	34,21	37,14	35,62	34,21	37,14	35,62
CIENCIA_2000_6381	27,78	35,09	31,01	36,11	39,39	37,68
CIENCIA_2000_6389	17,14	38,71	23,76	25,71	42,86	32,14
CIENCIA_2000_6391	25,81	37,21	30,48	46,77	50,00	48,33
CIENCIA_2001_6406	60,71	89,47	72,34	66,07	88,10	75,51
CIENCIA_2001_6410	24,00	57,14	33,80	50,00	71,43	58,82
CIENCIA_2001_6414	16,28	28,00	20,59	16,28	27,45	20,44
CIENCIA_2001_6416	27,63	29,17	28,38	52,63	42,11	46,78
CIENCIA_2001_6423	48,84	33,33	39,62	48,84	33,33	39,62
CIENCIA_2002_6441	75,00	91,30	82,35	73,21	68,33	70,69
CIENCIA_2003_6457	62,79	65,06	63,91	62,79	61,36	62,07
CIENCIA_2003_6465	51,14	62,50	56,25	51,14	60,00	55,22
CIENCIA_2003_6472	81,13	75,44	78,18	81,13	69,36	74,78
CIENCIA_2004_6480	41,94	49,37	45,35	47,31	46,32	46,81
CIENCIA_2004_6488	56,00	82,35	66,67	56,00	82,35	66,67
CIENCIA_2004_6494	19,05	28,57	22,86	26,98	36,17	30,91
CIENCIA_2005_6507	56,36	81,58	66,67	58,18	74,42	65,31
CIENCIA_2005_6514	26,47	34,62	30,00	26,47	34,62	30,00
CIENCIA_2005_6515	60,94	52,70	56,52	60,94	53,43	56,94
CIENCIA_2005_6518	36,91	51,67	43,06	45,24	46,91	46,06
MÉDIA	43,12	53,50	46,98	48,57	53,12	50,13

Com a avaliação dos dados de teste foram alcançados os seguintes resultados, uma cobertura de 43,12% para os sumários do GistSumm e com aplicação do CorrefSum a cobertura passou para 48,57%. A precisão passou de 53,50% para 53,12%. A medida F-measure passou de 46,98% (sumários originais) para 50,13% (sumários revisados).

Em relação a esses resultados podemos observar que as trocas realizadas com objetivo de recuperar a coesão referencial nos sumários podem melhorar a informatividade. Nos dois conjuntos de dados, treino e teste, houve um acréscimo das medidas de Cobertura e F-measure. Temos o aumento de cobertura sem perda de precisão. Entretanto, este ganho está relacionado ao aumento do sumário.

A partir da avaliação com a Rouge, observamos que um maior número de subs-

tituições nem sempre indica um maior ganho de informatividade. A Tabela 6 traz os resultados da Rouge para os sumários com pelo menos uma troca e para aqueles que apresentaram pelo menos duas trocas. O conjunto de sumários com mais trocas (duas ou mais) apresenta uma diferença em relação aos sumários originais similar àquela apresentada pelos sumários com uma ou mais trocas. A diferença em F-measure manteve-se em torno de 3% em ambos os casos.

Tabela 6: Resultados Rouge: Comparação com textos com 1 ou mais trocas e 2 ou mais trocas

	GISTSUMM-ORIGINAL			GISTSUMM-CORRIGIDO		
	COBERTURA	PRECISÃO	F-MEASURE	COBERTURA	PRECISÃO	F-MEASURE
1 OU MAIS TROCAS	45,40	55,05	49,23	51,23	54,95	52,60
2 OU MAIS TROCAS	44,59	52,83	48,16	50,94	52,30	51,41

Em uma análise texto a texto, podemos observar que o sumário do texto CIENCIA_2001_6410, por exemplo, obteve um maior ganho de F-measure passando de 33,80% para 58,82%, após a revisão do sumário, com apenas 1 substituição. Por outro lado, observamos que o texto CIENCIA_2005_6518 obteve um ganho menor, passando de 43,06% para 46,06%, com 3 substituições.

6.1.3 Avaliação Subjetiva dos Sumários do GistSumm Revisados

Esta seção traz os resultados da avaliação subjetiva entre os sumários originais do GistSumm e os sumários revisados pelo CorrefSum. Foram escolhidos 10 textos, correspondendo a 20% do corpus. Os textos escolhidos foram aqueles cujo sumário corrigido apresentou uma F-measure (Rouge) com um maior aumento em relação ao sumário original.

Para a avaliação subjetiva foram analisados os quesitos de legibilidade e informatividade, tais como interpretados pelos juizes, a partir do questionário fornecido (Anexo A). Uma informação relevante sobre os questionários é que o texto original e o corrigido não

foram identificados, para que isso não pudesse influenciar na opinião pessoal do avaliador. Na Tabela 7, temos os resultados da avaliação de legibilidade e na Tabela 8 os resultados da informatividade.

Tabela 7: Avaliação Subjetiva da Legibilidade

	JUIZ 1			JUIZ 2			JUIZ 3			JUIZ 4			JUIZ 5		
	O	C	A	O	C	A	O	C	A	O	C	A	O	C	A
CIENCIA_2000_6389		X			X			X				X			X
CIENCIA_2000_6391		X			X			X		X			X		
CIENCIA_2000_17109		X			X			X				X	X		
CIENCIA_2000_17112		X			X			X		X			X		
CIENCIA_2001_6410		X			X			X				X			X
CIENCIA_2001_6416		X				X			X			X			X
CIENCIA_2002_22005		X			X			X				X	X		
CIENCIA_2004_6494	X			X			X					X	X		
CIENCIA_2005_28747		X			X			X				X	X		
CIENCIA_2005_28766		X			X			X				X	X		
SOMA	1	9	0	1	8	1	1	8	1	2	0	8	7	0	3

Tabela 8: Avaliação Subjetiva da Informatividade

	JUIZ 1			JUIZ 2			JUIZ 3			JUIZ 4			JUIZ 5		
	O	C	A	O	C	A	O	C	A	O	C	A	O	C	A
CIENCIA_2000_6389		X			X			X			X			X	
CIENCIA_2000_6391		X			X				X		X			X	
CIENCIA_2000_17109		X			X			X		X				X	
CIENCIA_2000_17112		X			X			X				X		X	
CIENCIA_2001_6410		X			X			X			X			X	
CIENCIA_2001_6416		X			X			X			X			X	
CIENCIA_2002_22005		X			X			X			X			X	
CIENCIA_2004_6494	X			X				X			X		X		
CIENCIA_2005_28747		X			X		X				X			X	
CIENCIA_2005_28766		X			X		X				X			X	
SOMA	1	9	0	1	9	0	2	7	1	1	8	1	1	9	0

LEGENDA
O=ORIGINAL
C=CORRIGIDO
A=AMBOS

Com base nesses resultados, podemos observar que 3 avaliadores concordaram em relação à legibilidade, nos informando que os sumários corrigidos são mais legíveis. O juiz 4 não identificou diferença na legibilidade e o juiz 5, discordando dos demais, acredita que 7 dos sumários originais são mais legíveis que os corrigidos

Na avaliação da informatividade, observamos uma concordância entre os 5 juizes, na maioria dos casos, eles concordam que os sumários corrigidos são mais informativos.

Acredita-se, com essa avaliação, que o objetivo de manter a legibilidade dos sumários e aumentar a informatividade foi atingido, e confirmam os resultados fornecidos pela Rouge, que mediu a informatividade de forma automática. Na seção 6.2.3 são discutidos os resultados da avaliação subjetiva dos sumários gerados pelo SuPor-2.

6.2 Experimentos e Avaliação - Supor-2

6.2.1 Experimento

Nesta seção serão analisados e discutidos os experimentos usando o sistema Supor-2. O corpus Summ-it foi utilizado de forma integral (50 textos) para realização desse experimento. A utilização do corpus de forma integral é dado pelo fato de não haver mais necessidade de observar as heurísticas, pois elas já foram desenvolvidas e testadas usando os sumários do GistSumm. A Tabela 9 mostra os resultados obtidos com os sumários do SuPor-2.

Tabela 9: Resultados dos 50 textos do Summ-it

NOME DO TEXTO	QTDE CADEIAS NO TEXTO	QTDE CADEIAS NO SUMÁRIO	QTDE. TROCAS TROCAS	TX.COMPRESSÃO ANTES (%)	TX.COMPRESSÃO DEPOIS (%)
CIENCIA_2000_6380	10	4	1	31	33
CIENCIA_2000_6381	11	9	3	25	38
CIENCIA_2000_6389	9	7	2	33	38
CIENCIA_2000_6391	6	5	1	40	40
CIENCIA_2000_17082	10	9	2	31	32
CIENCIA_2000_17088	11	10	3	31	33
CIENCIA_2000_17101	17	10	0	31	31
CIENCIA_2000_17108	9	6	0	36	36
CIENCIA_2000_17109	12	11	0	42	42
CIENCIA_2000_17112	9	7	3	37	39
CIENCIA_2000_17113	16	11	1	36	38
CIENCIA_2001_6406	6	3	0	25	25
CIENCIA_2001_6410	8	7	2	39	40
CIENCIA_2001_6414	8	5	1	37	38
CIENCIA_2001_6416	8	4	2	26	38
CIENCIA_2001_6423	3	2	1	30	29
CIENCIA_2001_19858	11	9	1	35	36
CIENCIA_2002_6441	5	5	0	44	44
CIENCIA_2002_22005	12	8	1	34	35
CIENCIA_2002_22010	10	5	0	30	30
CIENCIA_2002_22015	16	7	2	31	33
CIENCIA_2002_22023	12	7	3	37	38
CIENCIA_2002_22027	22	11	4	31	34
CIENCIA_2002_22029	19	14	2	32	35
CIENCIA_2003_6457	9	8	3	34	38
CIENCIA_2003_6465	11	8	1	33	34
CIENCIA_2003_6472	4	3	0	56	56
CIENCIA_2003_24212	19	10	1	33	39
CIENCIA_2003_24219	13	11	2	32	33
CIENCIA_2003_24226	15	9	1	33	34
CIENCIA_2004_6480	10	4	2	27	33
CIENCIA_2004_6488	5	4	0	27	27
CIENCIA_2004_6494	8	8	1	38	38
CIENCIA_2004_26415	6	5	1	39	40
CIENCIA_2004_26417	14	5	1	30	30
CIENCIA_2004_26423	24	16	3	31	32
CIENCIA_2004_26425	21	14	0	30	30
CIENCIA_2005_6507	6	6	1	27	29
CIENCIA_2005_6514	8	7	2	33	34
CIENCIA_2005_6515	11	7	1	29	29
CIENCIA_2005_6518	10	9	1	33	36
CIENCIA_2005_28743	9	4	0	35	35
CIENCIA_2005_28747	11	8	1	31	32
CIENCIA_2005_28752	14	9	0	37	37
CIENCIA_2005_28754	15	10	2	33	33
CIENCIA_2005_28755	14	11	2	30	31
CIENCIA_2005_28756	13	11	0	32	32
CIENCIA_2005_28764	12	9	1	30	30
CIENCIA_2005_28766	23	17	5	33	39
CIENCIA_2005_28774	21	13	0	32	32
SOMA	586	402	67	-	-
MÉDIA	11,72	8,04	1,34	33,24	34,96

Com base nesses resultados, observamos que a quantidade de cadeias que apareceram no sumário foram em média 8,04 e a quantidade média de trocas efetuadas foram de 1,34 por texto. Em 13 textos nenhuma substituição foi efetuada. Nos demais textos as trocas variam de 1 a 5 trocas por texto.

Podemos observar também que a taxa de compressão dos sumários originais gerados pelo Supor-2 foram de 33,24% em média e após a revisão, os sumários ficaram com uma taxa média de 34,96%, correspondendo ao aumento de 1,72% em relação ao tamanho do sumário original.

Acredita-se que a diferença de número de substituições encontradas entre o GistSumm (89) e o Supor-2 (67) deve-se ao fato dos sumários gerados pelo Supor-2 são maiores.

Na próxima seção, é discutida a avaliação com a ferramenta Rouge dos sumários do Supor-2 corrigidos pelo CorrefSum.

6.2.2 Avaliação Rouge dos Sumários do Supor-2 Revisados

Usamos novamente a ferramenta ROUGE que permite avaliar a informatividade dos sumários (originais e revisados) em relação aos sumários de referência gerados por humanos. Os dados em relação aos sumários originais gerados pelo Supor-2 e os sumários revisados pelo CorrefSum foram analisados e podem ser analisados na Tabela 10.

Tabela 10: Avaliação Rouge com sumários originais e corrigidos gerados pelo Supor-2

	SUPOR-2-ORIGINAL			SUPOR-2-CORRIGIDO		
	COBERTURA	PRECISÃO	F-MEASURE	COBERTURA	PRECISÃO	F-MEASURE
MÉDIA	63,60	59,34	60,94	64,70	57,15	60,26

Podemos observar que diferente dos resultados obtidos pelo GistSumm os resultados do Supor-2 sofreram poucas alterações. Dentre as medidas fornecidas, observamos que a cobertura obteve um pequeno acréscimo passando de 63,60% para 64,70%. Entretanto, a taxa de precisão obteve uma queda de 2,19% passando de 59,34% para 57,15%. A medida F-measure permaneceu, praticamente, inalterada em torno de 60%.

Como esses resultados podem estar relacionados com a taxa de compressão dos sumários do SuPor-2 (33,24% em média), foi realizado outro experimento observando a taxa de compressão máxima de 30%. A Tabela 11 mostra os novos dados.

Tabela 11: SuPor-2 - Limite de taxa de compressão máxima de 30%

	QTDE. TROCAS	TAXA COMPRESSÃO ANTES (%)	TAXA COMPRESSÃO DEPOIS (%)
SUPOR-2*	67	33,24	34,96
SUPOR-2**	75	23,14	25,52

* Experimento com sumários gerados pelo SuPor-2

** Experimento com taxa de 30% de limite máximo de compressão dos sumários

Observamos na Tabela 11 que o número de substituições aumentou de 67 para 75. A taxa de compressão ficou em média 25,52%, menor do que os 34,96% do experimento anterior. Os valores da Rouge para os sumários gerados pelo Supor-2, com a limitação da taxa de compressão, podem ser observados na Tabela 12.

Tabela 12: Avaliação Rouge com sumários originais e corrigidos gerados pelo Supor-2 com limite de taxa de compressão

	SUPOR-2-ORIGINAL			SUPOR-2-CORRIGIDO		
	COBERTURA	PRECISÃO	F-MEASURE	COBERTURA	PRECISÃO	F-MEASURE
MÉDIA	48,37	63,07	54,33	53,15	64,08	57,36

Comparando os resultados das Tabelas 10 e 12, observamos que limitando o valor máximo da taxa de compressão em 30% obteve-se um resultado semelhante aos resultados do GistSumm (seção 6.1.2). A cobertura obteve um acréscimo, passando de 48,37% para 53,15%. As medidas de precisão e F-measure, também tiveram um acréscimo passando de 63,07% para 64,08% e 54,33% para 57,36%, respectivamente.

Uma avaliação subjetiva com o objetivo de avaliar a informatividade e a legibilidade dos sumários corrigidos pelo CorrefSum e seus resultados serão discutidos na seção seguinte.

6.2.3 Avaliação Subjetiva dos Sumários do Supor Revisados

Nesta seção, é discutida a avaliação subjetiva entre os sumários gerados pelo sistema Supor-2 e os sumários revisados pelo CorrefSum. Do conjunto de 50 sumários foram escolhidos 10 para realização da avaliação subjetiva. Para escolher os 10 sumários que fizeram parte dessa avaliação, foi usado o mesmo critério da seção 6.1.3, onde os sumários que apresentaram uma maior diferença na F-measure gerada pela ROUGE foram escolhidos. Essa avaliação segue os moldes da avaliação subjetiva do GistSumm com a utilização do mesmo questionário (no Anexo B). A Tabela 13 apresenta os dados em relação à legibilidade e na Tabela 14, dados em relação a informatividade.

Tabela 13: Avaliação Subjetiva da Legibilidade - SuPor-2

	JUIZ 1			JUIZ 2			JUIZ 3			JUIZ 4			JUIZ 5		
	O	C	A	O	C	A	O	C	A	O	C	A	O	C	A
CIENCIA_2000_17088	X			X			X			X			X		
CIENCIA_2000_17112			X			X		X		X			X		
CIENCIA_2000_17113		X			X			X		X					X
CIENCIA_2001_6410		X			X			X		X					X
CIENCIA_2002_22005		X			X			X			X				X
CIENCIA_2003_24212		X			X			X		X					X
CIENCIA_2003_24219			X			X			X	X					X
CIENCIA_2004_6480	X			X			X					X			X
CIENCIA_2004_26415			X			X			X	X					X
CIENCIA_2004_26423		X		X			X				X			X	
SOMA	2	5	3	3	4	3	3	5	2	7	2	1	2	1	7

Tabela 14: Avaliação Subjetiva da Informatividade - SuPor-2

	JUIZ 1			JUIZ 2			JUIZ 3			JUIZ 4			JUIZ 5		
	O	C	A	O	C	A	O	C	A	O	C	A	O	C	A
CIENCIA_2000_17088	X			X			X					X		X	
CIENCIA_2000_17112		X			X			X				X		X	
CIENCIA_2000_17113		X			X			X				X		X	
CIENCIA_2001_6410		X			X			X				X		X	
CIENCIA_2002_22005		X			X			X				X		X	
CIENCIA_2003_24212		X			X			X				X		X	
CIENCIA_2003_24219			X			X			X			X			X
CIENCIA_2004_6480		X			X			X			X			X	
CIENCIA_2004_26415		X			X			X				X		X	
CIENCIA_2004_26423		X		X			X				X			X	
SOMA	1	8	1	2	7	1	2	7	1	0	2	8	0	9	1

LEGENDA
O=ORIGINAL
C=CORRIGIDO
A=AMBOS

Com relação aos dados apresentados pelas Tabelas 13 e 14, observamos que, conforme os juízes, a legibilidade nos sumários corrigidos não foi afetada. As maiorias dos juizes concordaram que os sumários corrigidos estão mais legíveis que os sumários originais. Entretanto, o juiz 4 discorda e informa que a maioria dos sumários originais são mais legíveis. Para o juiz 5, não foi identificado diferença na legibilidade.

Em relação a informatividade, grande parte dos juízes indicaram que os sumários corrigidos, na maioria dos casos, são mais informativos que os sumários originais. Esse resultado é muito parecido com o que foi observado pela avaliação do GistSumm (seção 6.1.3) e extremamente importante, pois, esses resultados vão ao encontro das medidas fornecidas pela Rouge (seção 6.2.2) na Tabela 10. Apesar dos valores da Rouge não demonstrarem aumento significativo na informatividade para os textos escolhidos, vimos que isso não se confirmou na avaliação subjetiva.

Para uma melhor avaliação do sistema se faz necessária uma análise mais específica em relação à coesão referencial e coerência textual de todos os sumários corrigidos.

6.3 Experimentos com Sistema de Resolução de Correferência Automático

Um sistema de resolução automática de cadeias de correferência foi implementado por Souza (SOUZA, 2007) (descrito na seção 4.5). O sistema fornece as cadeias de correferência de cada texto processado. Foi realizada uma integração entre os sistemas de resolução de correferência e o CorrefSum. Foram processados os 50 textos do Summ-it, sumarizados com o GistSumm. Os resultados podem ser vistos na Tabela 15.

Tabela 15: Resultados dos experimentos com sistema de correferência automática

	ANOTAÇÃO MANUAL			ANOTAÇÃO AUTOMÁTICA		
	Nº CADEIAS NO TEXTO	Nº CADEIAS NO SUMÁRIO	Nº TROCAS	Nº CADEIAS NO TEXTO	Nº CADEIAS NO SUMÁRIO	Nº TROCAS
SOMA	586	330	89	393	194	36
MÉDIA	11,72	6,60	1,78	7,86	3,88	0,72

Com base nesses resultados, observamos que foram encontradas 330 cadeias de correferência em todo conjunto de textos, que representa 67,06% em relação à anotação manual. Deste total de 330 de cadeias encontradas, 194 cadeias estavam contidas nos sumários. Com a anotação automática, o sistema CorrefSum realizou 36 substituições, correspondendo a 40,45% das substituições realizadas com a anotação manual. Essas substituições ocorreram num total de 28 sumários, foram realizadas de 1 a 2 substituições em cada sumário. As taxas de compressão antes e depois do processamento obtiveram pequenas alterações passando de 25,33% para 26,18% em média.

Na Tabela 16 temos os dados da avaliação de informatividade. Os sumários corrigidos usando a anotação automática das cadeias de correferência são comparados com a anotação manual das cadeias e sumários originais. Para essa avaliação foi usada a ferramenta ROUGE.

Tabela 16: Resultados Rouge - comparação entre anotação manual e anotação automática

	Precisão	Cobertura	F-Measure
Sumários Originais	45,59	54,94	49,26
Sumários Corrigidos Cadeias Manuais	50,85	54,74	52,28
Sumários Corrigidos Cadeias Automat.	54,60	47,03	49,96

Com base nos resultados da Tabela 16, podemos observar que os sumários corrigidos com as cadeias geradas automaticamente apresentam algumas melhoras. A precisão obteve um acréscimo de 45,59% para 54,60%. Entretanto, a cobertura apresenteou um descréscimo de 54,94% para 47,03%. Com relação a F-measure observamos uma melhora em relação aos sumários originais passando de 49,26% para 49,96%.

É importante ressaltar que a tarefa de resolução de correferência ainda é desafio na área de PLN. A implementação utilizada foi resultado do primeiro trabalho considerando todos os tipos de sintagmas nominais e a língua portuguesa. Esse trabalho reportou uma F-measure de 59,60% quando avaliado no corpus Summ-it.

6.4 Avaliação Qualitativa das Substituições

Nesta seção é discutida a avaliação qualitativa das substituições dos elementos textuais das cadeias de correferência, a partir de um análise feita pelo autor. Para uma melhor análise, separamos os textos em grupos. O grupo A tem a análise dos sumários que sofreram apenas 1 substituição. O grupo B são sumários que sofreram 2 substituições. O grupo C são sumários que sofreram 3 e o grupo D são sumários com 4 ou mais substituições. Foi atribuída uma pontuação para a importância da troca. Essa análise foi realizada de forma subjetiva. A pontuação foi considerada da seguinte forma:

- 1 ponto: Para a troca que não expresse melhoria no resultado, isto é, sem a troca, o texto era coerênte. Por exemplo, a troca de “a pele humana” por “a pele humana normal”.

- 2 pontos: Para a troca que, de alguma forma, colaborou para o entendimento do sumário, trazendo informação adicional, aumentando a informatividade do sumário. Por exemplo, a troca de “‘a Ciência e Tecnologia” por “o MCT (Ministério da Ciência e Tecnologia)”.

- 3 pontos: Para a troca que contribui significantemente para o entendimento do sumário, trazendo informatividade e que resolveu um problema de coesão referencial. Por exemplo, a troca de “Guerra” por “o agrônomo Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina)”.

A Tabela 17 mostra as trocas efetuadas nos sumários do grupo A, usando o sistema CorrefSum e a análise subjetiva, conforme sistema de pontuação descrito acima.

Tabela 17: Substituições do Grupo A

NOME TEXO	PONTOS	EXPRESSÃO ORIGINAL	EXPRESSÃO SUBSTITUTA
CIENCIA_2000_6381	3	a ministra	A ministra da Justiça do país, Elisabeth Guigou
CIENCIA_2000_6389	3	Guerra	o agrônomo Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina)
CIENCIA_2000_6391	3	aquela carga	compensados de madeira exportados pela Selvaplac, subsidiária brasileira de um grupo da Malásia
CIENCIA_2000_17082	1	O desmatamento da Amazônia	o desmatamento com queimadas
CIENCIA_2000_17101	1	as mulheres que não receberam AZT	outro grupo de grávidas com HIV
CIENCIA_2000_17108	2	esse inseto	Um ser que invade corpos e domina a mente alheia, forçando suas vítimas a fazer o que ele ordena
CIENCIA_2001_6406	3	o gene	o gene da HBsAg
CIENCIA_2001_6410	3	ele	Barry Ellman, do Centro para Estudos Urbanos e Comunitários da Universidade de Toronto, Canadá
CIENCIA_2002_22010	3	os cientistas	Jennifer Leonard, da Universidade da Califórnia em Los Angeles, e colegas de instituições do Peru e do México
CIENCIA_2003_6457	1	o comitê central do Partido Comunista Chinês	o comitê -cuja reunião foi liderada pelo presidente chinês, Hu Jintao
CIENCIA_2003_6472	1	ele	Um estudo defendendo a idéia
CIENCIA_2003_24219	2	a Ciência e Tecnologia	o MCT (Ministério da Ciência e Tecnologia)
CIENCIA_2004_6494	1	o aquecimento global	o aquecimento global causado por esse gás
CIENCIA_2004_26415	3	o pesquisador	Bioantropólogo da Universidade MacMaster, no Canadá
CIENCIA_2005_6507	1	a radiação	a radiação cósmica de fundo
CIENCIA_2005_6514	1	O leito do rio	o fundo do Taquari
CIENCIA_2005_6515	3	a região	o Pantanal
CIENCIA_2005_28743	1	a chamada gripe espanhola	Essa epidemia de gripe, ou influenza
CIENCIA_2005_28754	2	As protuberâncias encontradas	arestas do material usado para preencher o vão entre as telhas térmicas da barriga do ônibus
CIENCIA_2005_28764	1	a borda da floresta	a borda da floresta e das savanas
SOMA	39		
MÉDIA	1,95		

A Tabela 17 mostra a análise de 20 sumários do corpus que tiveram apenas 1 substituição. Nesse grupo, temos 8 substituições que receberam 3 pontos indicando importantes substituições, 3 trocas com 2 pontos e 9 trocas receberam 1 ponto. A Tabela 18 mostra essa análise no grupo de textos que tiveram 2 substituições.

Tabela 18: Substituições do Grupo B

NOME TEXO	PONTOS	EXPRESSÃO ORIGINAL	EXPRESSÃO SUBSTITUTA
CIENCIA_2000_17088	3	o país	o Brasil
	3	a equipe carioca	Pesquisadores do Museu Nacional do Rio de Janeiro
CIENCIA_2001_6414	3	a região	a Antártida
	3	a península	a península Antártica
CIENCIA_2001_6416	1	o influenza	o H5 N1, o vírus influenza que, em 1997, matou 6 das 18 pessoas infectadas, em Hong Kong
	3	Os cientistas	pesquisadores da Universidade de Wisconsin-Madison (EUA)
CIENCIA_2001_19858	3	os cientistas	Cientistas do Centro de Estudos Saclay, na França
	3	Mirabel	Felix Mirabel, pesquisador que liderou o grupo
CIENCIA_2002_6441	1	um modelo alternativo	Um modelo testado em simulações muito detalhadas
	1	os gigantes gasosos	planetas gigantes extra-solares que foram encontra dos nos últimos anos
CIENCIA_2002_22005	3	brasileiros	Os brasileiros -Gilson Rambelli, Paulo Bava de Camargo e Flávio Rizzi, pesquisadores do Museu de Arqueologia e Etnologia (MAE) da USP-
	1	artefatos de várias origens ali depositados por a sucessivas levas de ocupantes do território português	inúmeros fragmentos, principalmente cerâmicos, que vão desde ânforas romanas até utensílios inteiros do século 16, passando pela cerâmica árabe
CIENCIA_2003_24212	1	Iowa	Iowa, EUA
	1	Os dois nascimentos	Os filhotes
CIENCIA_2004_6480	1	O animal	a terceira tentativa do órgão de criar um clone a partir de outro
	3	Rodrigues	O ministro Roberto Rodrigues (Agricultura)
CIENCIA_2004_26417	2	uma amostra natural	a amostra de efluentes ácidos analisada
	3	os cientistas	os cientistas liderados por Jillian Banfield, da Universidade da Califórnia em Berkeley
CIENCIA_2005_28774	2	uma pele biônica	uma pele artificial para robôs que imita uma parcela significativa das qualidades e capacidades da pele humana normal
	1	a pele humana	a pele humana normal
SOMA	42		
MÉDIA	2,10		

No grupo B, composto por 10 textos que obtiveram 2 trocas, somando um total de 20 substituições, podemos observar que 10 trocas receberam 3 pontos indicando que as substituições fez diferença na informatividade dos sumários e 2 trocas receberam 2 pontos. Dentre esse conjunto de 20 substituições, 8 delas não contribuíram para aumentar a informatividade do sumário. A Tabela 19 traz os dados referente aos textos do grupo C.

Tabela 19: Substituições do Grupo C

NOME TEXO	PONTOS	EXPRESSÃO ORIGINAL	EXPRESSÃO SUBSTITUTA
CIENCIA_2000_17112	1	o planeta	O mundo
	2	As metrópoles	zonas de intensa urbanização recente, como o sul dos EUA e o norte do México
	2	recursos hídricos	as reservas do líquido disponíveis em uma região
CIENCIA_2000_17113	1	os camundongos transgênicos	Os camundongos com essa alteração genética
	2	seu corpo	o corpo dos bichos
	2	a massa muscular	o volume total do corpo dos bichos
CIENCIA_2002_22027	2	um animal doméstico	gatos ou cachorros
	3	esse estudo	o trabalho de Allen e colegas
	2	as pessoas	as pessoas com animais domésticos
CIENCIA_2002_22029	3	O Maldi	espectrômetro de massa de ionização por dessorção a laser com auxílio de matriz (Maldi, na sigla em inglês)
	2	a Fapesp	A Fapesp (Fundação de Amparo à Pesquisa do Estado de São Paulo)
	3	Vasconcelos	a física nuclear Suzana Salem Vasconcelos, do LIP
CIENCIA_2003_24226	3	a Sars	e Sars (sigla em inglês para síndrome respiratória aguda grave)
	1	o vírus da gripe espanhola	o da gripe de 1918, a chamada gripe espanhola
	1	humanos	seres humanos
CIENCIA_2004_26423	3	ele	Mark Mattson, do Instituto Nacional do Envelhecimentos dos EUA
	1	EUA	os Estados Unidos
	2	a beta-amilóide	a proteína beta-amilóide
CIENCIA_2004_26425	3	ele	Domingos Matos, 36, médico da Universidade Federal do Pará
	3	Esse método	uma estratégia de interrupção estruturada
	3	o nosso trabalho	um estudo feito por Domingos Matos, 36, médico da Universidade Federal do Pará
CIENCIA_2005_6518	3	Os cientistas	o grupo do Butantan e do Goeldi
	1	a aranha	Seis das nove aranhas descobertas pelo grupo do Butantan e do Goeldi
	1	a cabeça da formiga	a cabeça daquelas formigas graúdas [as saúvas]
CIENCIA_2005_28755	1	pesquisadores da USP	Stevani e seus colaboradores
	1	fungo	uma descrita, a Gerronema viridilucens
	1	sensor de poluição	sensores vivos de poluição
CIENCIA_2005_28756	3	o animal	o mais antigo mamífero sul-americano do Paleoceno, o período geológico que marca o começo do reinado de seu grupo no planeta, logo depois da extinção dos dinossauros, há 65 milhões de anos
	1	os pesquisadores	pesquisadores argentinos
	3	o Cretáceo	o período anterior, o Cretáceo (quando os dinos ainda eram a forma dominante de vertebrado terrestre)
CIENCIA_2005_28766	3	duas moléculas	glicoproteínas (grosso modo, proteínas unidas a uma forma de açúcar)
	3	o pesquisador	Jonas Perales, do Laboratório de Toxinologia
	2	essas aplicações	a ação das substâncias contra doenças como o câncer
SOMA	68		
MÉDIA	2,06		

Com a Tabela 19 podemos observar que os textos do grupo C, textos com 3 substituições cada, tiveram os seguintes resultados, 13 trocas tiveram maior impacto na infor-

matividade, colaborando com a preservação da coesão referencial do sumário, 9 trocas impactaram no nível da informatividade do sumário, recuperando no texto-fonte expressões com mais informação. Dentre essas substituições, 11 delas não apresentaram impacto na informatividade. A Tabela 20 mostra os resultados o último grupo de análise.

Tabela 20: Substituições do Grupo D

NOME TEXO	PONTOS	EXPRESSÃO ORIGINAL	EXPRESSÃO SUBSTITUTA
CIENCIA_2000_17109	2	células-tronco da medula óssea	células não-especializadas, capazes de dar origem a qualquer tipo de tecido
	1	células hepáticas	outro tipo de célula -células hepáticas-
	3	as sanguíneas	e células sanguíneas
	3	os pesquisadores	pesquisadores do Imperial College, em Londres
CIENCIA_2002_22015	3	o astro	o buraco negro GRO J 1655-40
	2	buracos negros	buracos negros distribuídos ao redor da Via Láctea
	2	a estrela	estrela recém-nascida
	1	a força da gravidade	a gravidade
CIENCIA_2002_22023	2	bomba atômica	armamento nuclear ao redor do globo
	3	o sistema	uma rede de satélites do Departamento de Defesa dos EUA
	1	pequenos asteróides	os asteróides com mais de um quilômetro de diâmetro, capazes de destruir civilizações inteiras
	1	a atmosfera da Terra	a atmosfera terrestre
CIENCIA_2005_28747	1	As diferenças	diferenças como essas
	1	o objeto central	o objeto central de um quadro
	1	os chineses	Pessoas nascidas na China
	3	a universidade	a Universidade de Michigan em Ann Arbor, nos Estados Unidos
SOMA	30		
MÉDIA	1,87		

O grupo D consiste em 4 textos que tiveram 4 substituições cada, somando 16 substituições analisadas. Nesta análise vemos 5 trocas que receberam 3 pontos, 4 que receberam 2 pontos e 7 que receberam 1 ponto. Na Tabela 21, temos um resumo das Tabelas 17, 18, 19 e 20. Logo abaixo os dados são comentados:

Tabela 21: Tabela Resumida dos Grupos A, B, C e D

GRUPO	MÉDIA	1 PONTO	2 PONTOS	3 PONTOS
GRUPO A	1,95	9	3	8
GRUPO B	2,10	8	2	10
GRUPO C	2,06	11	9	13
GRUPO D	1,87	7	4	5
TOTAL	2,01	35 (39,33%)	18 (20,22%)	36 (40,45%)

Estamos considerando que as substituições com 2 e 3 pontos significam, respectivamente, bom e excelente. Com base nesses resultados observamos que 54 trocas (60,67%) tiveram um desempenho de bom à excelente, e que 35 (39,33%) não tiveram impacto na informatividade dos sumários. Em relação à média alcançada por cada troca observamos que em todos os grupos esse valor médio ficou em torno de 2. Com isso podemos concluir que as trocas, de maneira geral, ajudaram a contribuir na recuperação da coesão do sumário.

Realizando uma análise mais detalhada nesses resultados (Tabelas 17, 18, 19 e 20), observamos casos em que a troca não se fazia necessária, escolhemos alguns exemplos onde isso foi percebido:

- o desmatamento da Amazônia → o desmatamento com queimadas
- as mulheres que não receberam AZT → Outro grupo de grávidas com HIV
- um modelo alternativo → Um modelo testado em simulações muito detalhadas
- a pele humana → a pele humana normal
- o planeta → O mundo
- a atmosfera da Terra → a atmosfera terrestre
- os chineses → Pessoas nascidas na China
- EUA → os Estados Unidos

De forma geral, essas expressões não tem característica de anaforicidade: são expressões completas cujo significado é auto-contido, isto é, as interpretações são de certa forma independentes de contexto. Por isso, acreditamos que o sistema pode, ainda, ser melhorado com o desenvolvimento de um módulo de detecção de termos anafóricos, onde a substituição só será efetuada se o sistema identificar a necessidade.

Outro caso interessante, refere-se às substituições que foram realizadas com o objetivo de melhorar a coesão referencial, mas o antecedente escolhido para ser o substituto apresentou um problema de coesão referencial em uma expressão interna do sintagma. Vejamos abaixo alguns exemplos:

- A ministra → A ministra da Justiça do país, Elisabeth Guigou
- esse inseto → Um ser que invade corpos e domina a mente alheia, forçando suas vítimas a fazer o que ele ordena
- esse estudo → O trabalho de Allen e colegas
- pesquisadores da USP → Stevani e seus colaboradores

Por exemplo, a substituição da expressão "A ministra da Justiça do país, Elisabeth Guigou" ocasionou um problema de coesão referencial, onde não se consegue identificar a qual país a expressão se refere. A expressão "Um ser que invade corpos e domina a mente alheia, forçando suas vítimas a fazer o que ele ordena", acarreta um problema, pois não se consegue identificar que a expressão "ser" se refere a um "inseto". Outro problema é nas expressões "O trabalho de Allen e colegas" e "Stevani e seus colaboradores", onde não se consegue identificar os referentes para os nomes "Allen" e "Stevani". O tratamento de expressões internas do sintagma é outro ponto onde esse trabalho pode ser aperfeiçoado.

6.5 Discussões

Esta seção discute questões relacionadas à implementação do sistema, e apresenta uma análise dos resultados das trocas realizadas em expressões dos sumários.

Alguns problemas foram encontrados na anotação dos textos pelo parser Palavras, como por exemplo:

- Subtítulos: Por falta de pontuação no texto fonte, o parser não delimita e separa os subtítulos dos textos, ocasionando problemas na delimitação das sentenças. Esse foi um problema comum encontrado em diversos textos do corpus. Na busca de uma solução para esse problema, foi implementada uma rotina que identifica os subtítulos e ignorá-os na montagem das sentenças. Isso foi importante, pois esse problema ocasionava um erro no momento de comparar as sentenças do sumário com o texto-fonte.

- Aspas: O parser elimina as aspas do texto, problema esse que ocasionou um prejuízo quando o sumário revisado foi gerado, pois todos os sumários revisados (gerados através das informações dos arquivos XML do parser) estão sem aspas. Inclusive, foi um quesito que chamou a atenção dos juizes humanos, pois a falta de aspas no sumário ocasionou uma dificuldade na leitura e compreensão dos sumários.

- Continuidade da frase: O parser, por vezes, considera os dois pontos ":" como final de frase, incorretamente. Esse problema ocasionou, no início deste trabalho, problemas na geração do sumário corrigido, pois os sumários corrigidos não eram formados com as sentenças completas.

6.5.1 Anotação de correferência

Realizando uma análise nos sumários, observamos um caso interessante em relação à anotação das cadeias de correferência. Observamos o texto CIENCIA_2002_22023 completo na Figura 30 e o sumário gerado pelo GistSumm do texto na Figura 31.

Na Figura 32 temos os sumário corrigido pelo CorrefSum e observamos que o sistema realizou quatro substituições(em negrito na Figura 32):

- bomba atômica → armamento nuclear ao redor do globo

- o sistema → uma rede de satélites do Departamento de Defesa dos EUA

A maioria dos cientistas concorda que **os asteróides com mais de um quilômetro de diâmetro, capazes de destruir civilizações inteiras**, são uma preocupação que só se justifica a cada punhado de dezenas de milhões de anos. Mas novos cálculos mostram que bólidos mais modestos, com 50 metros de diâmetro e a capacidade de destruir uma cidade, despencam do céu uma vez por milênio. Na verdade, trata-se de boa notícia. Estimativas anteriores sugeriam que um evento desses ocorresse em média a cada 200 ou 300 anos. Os novos cálculos, aprimorados com o uso de informação antes mantida secreta pelo governo americano, oferecem uma estimativa mais precisa sobre a periodicidade desses episódios. Durante os últimos oito anos, **uma rede de satélites do Departamento de Defesa dos EUA** tem monitorado a atmosfera terrestre com o objetivo de detectar explosões _obviamente na tentativa de monitorar o uso de **armamento nuclear ao redor do globo**. Registros de **bomba atômica** nunca apareceram, mas, em compensação, **o sistema** foi capaz de apontar diversos eventos de explosões _todas causadas pela entrada de **pequenos asteróides na atmosfera da Terra** e sua subsequente quebra pelo atrito com o ar. Para os militares a coisa acabou não sendo lá muito útil, mas os dados se tornaram um prato cheio para os astrônomos. "Em oito anos, detectamos mais de 300 eventos, graças ao nosso sistema de calibragem dos dados de satélite", conta Douglas Revelle, do Laboratório Nacional de Los Alamos, um dos autores do estudo, que está publicado na edição de hoje da revista britânica "Nature" (www.nature.com). Incidências de rochas espaciais de poucos metros de diâmetro na atmosfera acontecem com razoável frequência _anualmente, segundo os pesquisadores. "Esses corpos medidos em metros são interessantes cientificamente, mas não oferecem absolutamente nenhum perigo aos humanos", diz Robert Jedicke, da Universidade do Arizona, escolhido pela "Nature" para comentar o estudo. A ameaça só existe quando os bólidos têm 50 metros ou mais. Foi um meteoro desse tipo (ou um disco voador, segundo fãs de ufologia) que explodiu sobre Tunguska, na Sibéria, em 1908, destruindo centenas de quilômetros quadrados de floresta. Se um desses explodisse sobre uma região habitada, poderia matar milhões. Felizmente, com base na nova estimativa, parece haver ainda nove séculos para catalogar os pedregulhos espaciais e se preparar para futuras colisões.

Figura 30: Texto CIENCIA_2002_22023

Registros de **bomba atômica** nunca apareceram, mas, em compensação, **o sistema** foi capaz de apontar diversos eventos de explosões _todas causadas pela entrada de **pequenos asteróides** n**a atmosfera da Terra** e sua subseqüente quebra pelo atrito com o ar. "Em oito anos, detectamos mais de 300 eventos, graças ao nosso sistema de calibragem dos dados de satélite", conta Douglas Revelle, do Laboratório Nacional de Los Alamos, um dos autores do estudo, que está publicado na edição de hoje da revista britânica "Nature" (www.nature.com).

Figura 31: Sumário GistSumm do texto CIENCIA_2002_22023

- pequenos asteróides → os asteróides com mais de um quilômetro de diâmetro, capazes de destruir civilizações inteiras
- a atmosfera da Terra → a atmosfera terrestre

Observamos que o sistema realizou as substituições de forma correta (levando em consideração o sistema de pontuação implementado). Mas a troca da expressão "pequenos asteróides" por "os asteróides com mais de um quilômetro de diâmetro, capazes de destruir civilizações inteiras" acabou alterando totalmente o sentido do sumário. Isso foi interessante, pois se percebeu que esses dois termos não deveriam estar na mesma cadeia de correferência, pois não podem ser substituídos por não terem o mesmo referente.

Registros de **armamento nuclear ao redor do globo** nunca apareceram, mas em compensação, **uma rede de satélites do Departamento de Defesa dos EUA** foi capaz de apontar diversos eventos de explosões -todas causadas pela entrada de **os asteróides com mais de um quilômetro de diâmetro, capazes de destruir civilizações inteiras** n**a atmosfera terrestre** e sua subsequente quebra pelo atrito com o ar. Em oito anos, detectamos mais de 300 eventos, graças ao nosso sistema de calibragem dos dados de satélite, conta Douglas Revelle, do Laboratório Nacional de Los Alamos, um dos autores do estudo, que está publicado na edição de hoje da revista britânica Nature (www.nature.com)

Figura 32: Sumário gerado pelo GistSumm e corrigido pelo CorrefSum do texto CIENCIA_2002_22023

6.5.2 Análise de Substituições

Neste item, veremos algumas considerações em relação a algumas substituições realizadas pelo sistema.

- Pronomes: pronomes oblíquos (o, a, lhe) e pronomes pessoais de 3ª pessoa do plural (nós) não podem ser simplesmente substituídos, pois os seus complementos nominais e verbais não concordam com as substituições, tornando o texto lido, após a substituição, incoerente.

- Substituições simples: observamos no texto CIENCIA_2000_17109 que o sistema procurou substituir o termo "as sangüíneas", em "das sanguíneas", por "células sangüíneas", mas ocasionou um erro no sumário, resultando "dcélulas sangüíneas". Foi implementada uma rotina que verifica se antes do elemento que será substituído aparece a preposição "de". Dessa forma, o sistema realiza a edição necessária no momento da substituição. Nesse caso, o sistema trocou "das sangüíneas" por "de células sangüíneas"

- Substituições diretas simples: o sistema não realiza substituições de sintagmas em que a relação entre eles for direta (simples), por exemplo:

  – a pressão - pressão (não é substituído)

  Mas quando o sintagma trouxer alguma informação adicional o sistema considera uma troca relevante e realiza as substituições, como por exemplo:

  – a Sars - Sars (sigla em inglês para síndrome respiratória aguda grave)

- Cadeias com mais de um elemento com a mesma pontuação: caso haja uma situação onde o elemento a ser substituído tenha a mesma pontuação que o elemento a ser trocado o sistema não efetua a troca, por exemplo, texto CIENCIA_2005_28754 cadeia "set_51":

  – o trabalho - 1 ponto (regra do primeiro elemento cadeia)

  – o procedimento de reparo - 1 ponto (regra do maior elemento da cadeia)

  – a missão - 0 pontos

  – o procedimento - 0 pontos

  Neste exemplo, o sumário em questão possui a expressão "o trabalho" onde o melhor candidato para troca é o sintagma "o procedimento de reparo". Como a pontuação é a mesma o sistema não realiza a substituição.

- Durante esse trabalho se observou que nem sempre o primeiro elemento da cadeia é o mais pontuado elemento para a substituição, vejamos alguns exemplos:
  Caso 1: Texto CIENCIA_2003_24212 - cadeia "set_46"

  – os filhotes - 1 ponto (regra do primeiro elemento cadeia)

  – clones de outra espécie, o banteng, um tipo de gado ameaçado de extinção - 1 ponto (regra aposto)

  – os dois batengs produzidos em Iowa - 1 ponto (regra nome-próprio)

  – os dois filhotes - 0 pontos

  – eles - 0 pontos

  – cópias genéticas idênticas de um banteng macho que morreu no Parque Selvagem Animal de San Diego em 1980 - 2 pontos (regra maior elemento da cadeia e nome-próprio)

  Caso 2: Texto CIENCIA_2005_28766 - cadeia "set_57"

  – duas moléculas - 1 ponto (regra do primeiro elemento cadeia)

- as substâncias antiofídicas - 0 pontos
- glicoproteínas (grosso modo, proteínas unidas a uma forma de açúcar) - 2 pontos (regra do maior elemento da cadeia e aposto)
- a DM43 e a DM64 - 1 ponto (regra nome-próprio)
- as substâncias - 0 pontos
- elas - 0 pontos

Observamos, por exemplo, no caso 1, que o elemento mais pontuado dessa cadeia é o último elemento e, em relação ao caso 2, o elemento com maior pontuação estava no meio da cadeia. Apesar da maioria das substituições realizadas pelo sistema optarem pelo primeiro elemento da cadeia, há casos em que outros elementos acabam sendo selecionados.

Este capítulo apresentou os resultados dos experimentos usando sumários gerados por dois diferentes sistemas, o GistSumm e o SuPor-2. Foram discutidas duas abordagens de avaliação: a avaliação automática usando a ferramenta Rouge e a avaliação subjetiva onde os sumários revisados foram avaliados por 5 juízes humanos. Os ítens avaliados subjetivamente foram legibilidade e informatividade. Por fim, foi realizada uma avaliação qualitativa das substituições feitas pelo sistema CorrefSum. As considerações finais desta dissertação são apresentadas no próximo capítulo.

Capítulo 7

Considerações Finais

Problemas na geração de sumários automáticos são levantados em alguns trabalhos[(COELHO, 2007), (CARBONEL, 2007)], como por exemplo, problema de coesão referencial, que acaba dificultando a interpretação (coerência) do sumário automático. Esse problema é agravado quando o sistema de sumarização utiliza o método extrativo para compor os sumários, pois esse método acaba selecionando sentenças inteiras na composição do sumário sem levar em conta os elos referenciais do texto-fonte.

Um dos problemas mais comuns é a ocorrência de expressões referenciais pouco significativas nos sumários. A carência informacional, por vezes, pode causar incompreensão do sumário, acarretando problemas de interpretação.

É nesse contexto que esse trabalho se insere. O trabalho têm como foco a recuperação da coesão referencial nos sumários extrativos através da verificação e análise das cadeias de correferência do texto-fonte.

Este trabalho propõe a pós-edição de sumários automáticos, buscando reescrevê-lo de forma mais coerente, sem problemas nos elos referenciais. Para isso, as expressões nominais dos sumários são analisadas com base na anotação de correferência, com o objetivo de buscar dentro da cadeia, expressões representativas da entidade evocada.

Os experimentos realizados neste trabalho tiveram como base dois sumarizadores:

o GistSumm e o SuPor-2. Os resultados obtidos através dos sumários do GistSumm foram considerados satisfatórios, tanto na avaliação automática (onde se conseguiu aumentar a F-measure de 46,98% para 50,13% - Tabela 5), quanto na avaliação subjetiva, onde os sumários corrigidos tiveram um grande impacto na avaliação da informatividade.

Com os sumários do SuPor-2, num primeiro momento, a F-measure ficou em torno dos 60% (Tabela 10), tanto nos sumários originais, quanto nos revisados. Quando o experimento foi repetido observando a taxa de compressão máxima de 30%(Tabela 12) observamos que os resultados foram próximos aos encontrados no GistSumm e que a F-measure obteve um acréscimo passando de 54,33% para 57,36%. Na avaliação subjetiva foi demonstrado um acréscimo na informatividade dos sumários corrigidos.

Foi ainda realizada uma avaliação qualitativa das substituições feitas pelo sistema. Esta avaliação aponta também para um ganho na informatividade, mas é necessária realizar uma avaliação mais detalhada para avaliar questões especificamente ligadas a coerência textual, como, perda de sentido por falta de contexto suficiente ou de antecedente textual.

Um intercâmbio com a Universidade Federal de São Carlos (UFSCAR) foi realizado durante o desenvolvimento deste trabalho, onde surgiu a oportunidade de interagir com a Profª Drª. Lucia Rino e sua equipe, pesquisadores na área de sumarização automática.

A seguir temos as contribuições, limitações e trabalhos futuros deste trabalho.

7.1 Contribuições

Destacam-se, nesta seção, as principais contribuições deste trabalho:

- Este trabalho é a primeira abordagem sobre revisão da coesão referencial em sumários para a língua portuguesa;
- Um sistema para revisão automática e semi-automática de expressões referenciais em sumários foi desenvolvido;

- Uma interface para manipulação das cadeias de correferência de forma manual foi desenvolvida;
- O sistema foi integrado com um sistema de resolução automática de correferência, com o objetivo de futuramente, gerar sumários automáticos utilizando abordagem superficial com manutenção dos elos referenciais;
- Foram avaliados dois sumarizadores automáticos, GistSumm e Supor-2, para geração e revisão dos sumários;
- A pesquisa considerou dois métodos de avaliação: avaliação automática e avaliação subjetiva;
- Um artigo foi aceito para publicação no PROPOR[1] - International Conference on Computational Processing of Portuguese Language.
  - Título: CorrefSum: Referencial Cohesion Recovery in Extractive Summaries
  - Autores: Patrícia Nunes Gonçalves, Lucia Rino e Renata Vieira

7.2 Limitações

São limitações deste trabalho:

- Uso de somente textos do gênero jornalístico de divulgação científica;
- Dependência da anotação manual das cadeias de correferência;
- Avaliação superficial de legibilidade e informatividade, sem uma análise qualitativa específica dos problemas de coesão referencial.

[1] http://www.propor2008.org/

7.3 Trabalhos Futuros

Como continuidade da pesquisa realizada neste projeto de mestrado, apontamos alguns itens que poderão ser utilizados para futura pesquisa deste trabalho:

- Realizar experimentos e avaliar o sistema CorrefSum com sumarizadores que utilizam abordagem profunda, como por exemplo, o VeinSumm (CARBONEL, 2007).
- Integrar o sistema desenvolvido com um sistema de classificação de expressões anafóricas para verificar a necessidade de efetuar a substituição.
- Implementar um módulo que resolva os problemas de coesão referencial dos sintagmas internos.
- Implementar um módulo que gere expressões referenciais com base na cadeia de correferência e que essa expressão seja utilizada para substiuição no sumário automático.
- Usar as informações das cadeias associativas, pois, acredita-se que de alguma forma, podem enriquecer os sumários extrativos.
- Construir e avaliar sumarizadores automáticos que levem em consideração as cadeias de correferência na escolha das sentenças relevantes.

Referências

AMO, P. et al. Orthografic co-reference resolution between proper nouns throught the calculation of the relation of replicancia. In: *Workshop Coreference and Its Applications*. Maryland, USA: [s.n.], 1999.

AZZAM, S.; HUMPHREYS, K.; GAIZAUSKAS, R. Using coreference chains for text summarization. In: *Proceedings of The Relation of Discourse/Dialogue Structure and Reference*. [S.l.: s.n.], 1999.

BAXENDALE, P. Machine- made index for technical literature an experiment. *IBM Journal of Research and Development*, v. 2, p. 354–365, 1958.

BICK, E. *The Parsing System "PALAVRAS- Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Tese (Doutorado) — Department of Linguistics, University of Århus, DK., 2000.

CARBONEL, T. I. *Estudo e validação de teorias do domínio lingüístico com vistas à melhoria do tratamento de cadeias de co-referência em Sumarização Automática*. Dissertação (Mestrado) — Universidade Federal de São Carlos (UFSCAR). São Carlos, Agosto 2007.

CHAVES, A. *A resolução de anáforas pronominais da língua portuguesa com base no algoritmo de Mitkov*. Dissertação (Mestrado) — Universidade Federal de São Carlos (UFSCAR). São Carlos, Julho 2007.

COELHO, J. C. B. *Uso de Informação de Correferência e Anáfora para Verificação da Coesão e Coerência Textual na Sumarização Automática*. Junho 2007. Trabalho de Conclusão de Curso de Letras. Unisinos - São Leopoldo.

COELHO, J. C. B. et al. Resolving portuguese nominal anaphora. In: VIEIRA, R. et al. (Ed.). *7th Workshop on Computational Processing of Written and Spoken Language (PROPOR'2006)*. Itatiaia, RJ: Springer, 2006.

COELHO, T. T. *Resolução de anáfora pronominal em português utilizando o algoritmo de Lappin e Leass*. Dissertação (Mestrado) — Departamento de Computação, Universidade Estadual de Campinas - Unicamp, 2005.

COLLOVINI, S. *Análise de Expressões Referenciais em Corpus Anotado da Língua Portuguesa*. Dissertação (Mestrado) — Departamento de Computação, Universidade do Vale do Rio dos Sinos - Unisinos, 2005.

COLLOVINI, S. et al. Summit: Um corpus anotado com informações discursivas visando à sumarização automática. In: *5º Workshop em Tecnologia da Informação e da Linguagem Humana (TIL'2007)*. Rio de Janeiro, RJ: Proceedings of the SBC, 2007.

COLLOVINI, S.; VIEIRA, R. Anáforas nominais definidas: balanceamento de corpus e classificação. In: *IV Workshop de Tecnologia da Informação e Linguagem Humana TIL*. Ribeirão Preto, SP: Proceeding of the Brazilian Symposium on Artificial Intelligence, 2006.

COLLOVINI, S.; VIEIRA, R. Análise de expressões referenciais em corpus anotado da língua portuguesa. In: *V Best MSc dissertation/PhD thesis contest (CTDIA'2006)*. Ribeirão Preto, SP: Proceedings of the SBIA-IBERAMIA, 2006.

CRISTEA, D.; IDE, N.; ROMARY, L. Veins theory: A model of global discourse cohesion and coherence. In: *COLING-ACL*. [S.l.: s.n.], 1998. p. 281–285.

FILHO, P. P. B.; PARDO, T.; NUNES, M. d. G. V. Summarizing scientific texts: Experiments with extractive summarizers. In: . Rio de Janeiro, Brasil: Proceedings of the Seventh International Conference on Intelligent Systems Design and Applications ISDA,, 2007.

GASPERIN, C.; GOULART, R.; VIEIRA, R. Uma ferramenta para resolução automática de co-referência. In: *Encontro Nacional de Inteligência Artificial (ENIA 2003)*. Campinas, SP: [s.n.], 2003.

JONES, J. S. *Automatic Summarizing: factors and directions. In I. Mani and M. Maybury (eds.), Advances in automatic text summarization*. [S.l.]: The MIT Press, 1999.

JORDAN, M. Short texts to explain problem-solution structures and vice versa. *Instructional Science*, v. 9, p. 221–252, 1980.

JURAFSKY, D.; MARTIN, J. Speech and language processing. In: _____. [S.l.]: Alan Apt, 2000. cap. Discourse, p. 670–718.

KASHANI, M. M.; POPOWICH, F. Pronoun generation for text summarization and question answering. In: *Proceedings of 5th Slovenian and 1st international Language Technologies Conference 2006*. [S.l.: s.n.], 2006.

KOCH, I. G. V. *O texto e a construção dos sentidos*. [S.l.]: São Paulo: Contexto, 2000.

KOCH, I. G. V. *Desvendando os Segredos do texto*. [S.l.]: São Paulo: Cortez, 2003.

KOCH, I. G. V.; SILVA, M. C. d. S. *Lingüística Aplicada ao Português: Sintaxe*. [S.l.]: São Paulo: Cortez, 2002.

KOCH, I. G. V.; TRAVAGLIA, L. C. *A coerência textual*. [S.l.]: São Paulo: Contexto, 1990.

KOCH, I. G. V.; TRAVAGLIA, L. C. *A coesão textual*. [S.l.]: São Paulo: Contexto, 1996.

LAPPIN, S.; LEASS, H. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, v. 20(4), p. 535–561, 1994.

LEITE, D.; RINO, L. *SuPor: extensões e acoplamento a um ambiente para mineração de dados*. [S.l.], 2006.

LEITE, D.; RINO, L. Uma comparação entre sistemas de sumarização automática extrativa. In: *IV Workshop de Tecnologia da Informação e Linguagem Humana TIL*. Ribeirão Preto, SP: Proceeding of the Brazilian Symposium on Artificial Intelligence, 2006.

LEITE, D. et al. Extractive automatic summarization: Does more linguistic knowledge make a difference? In: *C. Biemann, I. Matveeva, R. Mihalcea, and D. Radev (eds.)*. Rochester, NY, USA: Proceedings of the HLT/NAACL Workshop on TextGraphs-2: Graph-Based Algorithms for Natural Language Processing, 2007.

LIN, C.-Y. Rouge: A package for automatic evaluation of summaries. In: *Workshop on Automatic Summarization*. Philadelphia,USA: Proceedings of ACL-02, 2000.

LIN, C.-Y. Looking for a few good metrics: Automatic summarization evaluation - how many samples are enough? In: *Proceedings of 4th Workshop NTCIR*. [S.l.: s.n.], 2004.

LUHN, H. P. The automatic creation of literature abstracts. *j-IBM-JRD*, v. 2, p. 159–165, 1958. ISSN 0018-8646.

LUO, X. et al. A mention-synchronous coreference resolution algorithm based on the bell tree. In: *Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 2004.

MANI, I. *Automatic Summarization*. [S.l.]: John Benjamins Publishing Co., 2001.

MANN, W. C.; THOMPSON, S. A. Rhetorical structure theory: Description and construction of text structures. In: KEMPEN, G. (Ed.). *Natural Language Generation: New Results in Artificial Intelligence, Psychology, and Linguistics*. Dordrecht: Nijhoff, 1987. p. 85–95.

MóDOLO, M. *SuPor: an Environment for Exploration of Extractive Methods for Automatic Text Summarization for Portuguese (in Portuguese)*. Dissertação (Mestrado) — Universidade Federal de São Carlos (UFSCAR). São Carlos, Dezembro 2003.

MILLER, G. A. *WordNet: a lexical database for English*. [S.l.]: Communications of the ACM. Volume 38, Issue 11, 1995.

MITKOV, R. Robust pronoun resolution with limited knowledge. In: *Conference on Computational Linguistics*. [S.l.: s.n.], 1998.

MITKOV, R. *Anaphora Resolution*. [S.l.]: Longman, 2002.

MüLLER, C.; STRUBE, M. Mmax: A tool for the annotation of multi-modal corpora. In: *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. Seattle, Washington: [s.n.], 2001. p. 45–50.

MUC-6. A mention-synchronous coreference resolution algorithm based on the bell tree. In: *Sixth Message Understanding Conference(MUC-6)*. [S.l.: s.n.], 1995.

NENKOVA, A.; SIDDHARTHAN, A.; MCKEOWN, K. Automatically learning cognitive status for multi-document summarization of newswire. In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2005.

NETO, M. C. M.; B, A. N.; GOMES, A. Satsumm - uma ferramenta para sumarização automática de textos jornalísticos. In: *Sétima Escola Regional de Computação Bahia-Sergipe*. Vitória da Conquista: [s.n.], 2007.

NG, V. *Machine Learning for Coreference Resolution: Recent Successes and Future Directions*. [S.l.], 2003.

NG, V. Machine learning for coreference resolution: From local classification to global ranking. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*. Michigan, US: [s.n.], 2005.

NG, V. Supervised ranking for pronoun resolution: Some recent improvements. In: *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05)*. Pittsburgh, Pennsylvania: [s.n.], 2005.

NG, V. Shallow semantics for coreference resolution. In: *International Joint Conferences on Artificial Intelligence (IJCAI'2007)*. Hyderabad, India: [s.n.], 2007.

NICOLAE, C. *Identification of Entity Mentions in Text and Their Coreference Resolution*. Dissertação (Mestrado) — University of Texasm at Dallas, December 2006.

PAICE, C. D. *The automatic generation of literature abstracts: an approach based on the identification of self- indicating phrases*. [S.l.]: Butterworth Co., 1981.

PARDO, T. *DMSumm: Um Gerador Automático de Sumários*. Dissertação (Mestrado) — Departamento de Computação, Universidade Federal de São Carlos. São Carlos-SP, Abril 2002.

PARDO, T. *GistSumm - GIST SUMMarizer: Extensões e Novas Funcionalidades*. [S.l.], 2005.

PARDO, T. *Métodos para Análise Discursiva Automática*. Tese (Doutorado) — Departamento de Computação, Universidade Federal de São Carlos. São Carlos-SP, 2005.

PARDO, T.; NUNES, M. Dizer an automatic discourse analyzer for brazilian portuguese. In: *V Best MSc dissertation/PhD thesis contest (CTDIA'2006)*. Ribeirão Preto, SP: Proceedings of the SBIA-IBERAMIA, 2006.

PARDO, T.; RINO, L. Dmsumm: Review and assessment. In: *3º International Conference Advances in Natural Language Processing*. Portugual: [s.n.], 2002.

PARDO, T.; RINO, L.; NUNES, M. Neuralsumm: Uma abordagem conexionista para a sumarização automática de textos. In: *Anais do IV Encontro Nacional de Inteligência Artificial*. Campinas, São Paulo: [s.n.], 2003.

PERINI, M. *Gramática descritiva do português*. [S.l.]: São Paulo: Editora Ática, 2003.

POESIO, M.; KABADJOV, M. A. A general-purpose,off the shelf anaphoric resolver. In: *Proceedings of International Conference on Language Resources and Evaluation*. Lisboa, Portugual: [s.n.], 2004.

PONZETTO, S.; STRUBE, M. Semantic role labeling for coreference resolution. In: *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'2006)*. Trento, Italy: [s.n.], 2006.

RADEV, D. *Text summarization*. Julho 2004.
Http://www.summarization.com/sigirtutorial2001.ppt.
Acessado em 30/05/2007. Tutorial ACM/SIGIR CLAIR: Computational Linguistics And Information Retrieval group.

RADEV, D.; JING, H.; BUDZIKOWSKA, M. Centroid-based summarization of multiple documents. In: *Workshop on Automatic Summarization*. Seatle,USA: Proceedings of ANLP/NAACL, 2000.

RIBEIRO-JUNIOR, L. C. et al. Uso de informações semânticas na identificação de anáforas indiretas e associativas. In: *5º Workshop em Tecnologia da Informação e da Linguagem Humana (TIL'2007)*. Rio de Janeiro, RJ: Proceedings of the SBC, 2007.

RINO, L.; CARBONEL, T. *Rhesuma-2: Análise dos sumários e estudos dos casos de quebra de cadeias de co-referência*. [S.l.], 2006.

RINO, L.; PARDO, T. *A Coleção TeMário e a Avaliação de Sumarização Automática*. [S.l.], 2006.

SENO, E. *RHeSumaRST: Um sumarizador automático de estruturas RST*. Dissertação (Mestrado) — Departamento de Computação, Universidade Federal de São Carlos. São Carlos-SP, 2005.

SOUZA, J. G. C. de. *Resolução automática de correferência aplicada à língua portuguesa*. Novembro 2007. Trabalho de conclusão. Unisinos - São Leopoldo.

STEINBERGER, J. et al. Two uses of anaphora resolution in summarization. In: *Information Processing and Management*. [S.l.: s.n.], 2007.

TEIXEIRA, M. *Coesão Referencial*. Junho 2007. Acessado em 10/06/2007
http://www.comunica.unisinos.br/professores/marlene/arquivos/referenciacao_2004_1.pdf.

VIEIRA, R. *Definite description processing in unrestricted text*. Tese (Doutorado) — University of Edinburgh, Edinburgh, 1998.

WITTEN, I. H.; FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation*. San Francisco: Morgan Kaufmann, 2000.

Anexo A - Questionários Sumários GistSumm

Sumários GistSumm Originais e Revisados

10 textos corpus Summit

CIENCIA_2000_6389

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

Guerra citou a micropropagação de vegetais (produção de mudas em laboratório, feita para evitar doenças e selecionar vegetais saudáveis) como exemplo de biotecnologia de baixo custo.

SUMARIO 2:

O agrônomo Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina) citou a micropropagação de vegetais (produção de mudas em laboratório, feita para evitar doenças e selecionar vegetais saudáveis) como exemplo de biotecnologia de baixo custo.

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

CIENCIA_2000_6391

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

Não temos certeza de que compensados de madeira exportados pela Selvaplac, subsidiária brasileira de um grupo da Malásia era ilegal, mas sabemos que 80% da atividade madeireira no Brasil é irregular e que a Selvaplac tem uma tradição de envolvimento com madeira ilegalmente extraída, disse a a Folha Rebeca Lerer, ativista brasileira do Greenpeace.

SUMARIO 2:

"Não temos certeza de que aquela carga era ilegal, mas sabemos que 80% da atividade madeireira no Brasil é irregular e que a Selvaplac tem uma tradição de envolvimento com madeira ilegalmente extraída", disse à Folha Rebeca Lerer, ativista brasileira do Greenpeace.

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

CIENCIA_2000_17109

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

Pesquisas em camundongos haviam mostrado que células não-especializadas, capazes de dar origem a qualquer tipo de tecido poderiam originar outro tipo de célula -células hepáticas-, além de células

sanguíneas. Para descobrir se o mesmo acontecia em seres humanos, pesquisadores do Imperial College, em Londres analisaram células do fígado de mulheres que haviam sofrido um transplante de medula óssea, cujo doador havia sido um homem.

SUMARIO 2:

Pesquisas em camundongos haviam mostrado que células-tronco da medula óssea poderiam originar células hepáticas, além das sanguíneas. Para descobrir se o mesmo acontecia em seres humanos, os pesquisadores analisaram células do fígado de mulheres que haviam sofrido um transplante de medula óssea, cujo doador havia sido um homem.

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

CIENCIA_2000_17112

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

Um estudo publicado na edição de hoje da revista "Science" afirma que 1,75 bilhão de pessoas já enfrentam severa escassez de água no planeta. "A demanda aumenta de forma drástica no mundo todo", afirmou o especialista em recursos hídricos José Galizia Tundisi, do Instituto Internacional de Ecologia, em São Carlos (SP). "As metrópoles não têm recursos hídricos suficientes para suportar o crescimento populacional", disse Tundisi.

SUMARIO 2:

Um estudo publicado na edição de hoje da revista Science afirma que 1,75 bilhão de pessoas já enfrentam severa escassez de água no mundo. A demanda aumenta de forma drástica no mundo todo, afirmou o especialista em as reservas do líquido disponíveis em uma região José Galizia Tundisi, do Instituto Internacional de Ecologia, em São Carlos (SP). Zonas de intensa urbanização recente, como o sul dos EUA e o norte do México não têm recursos hídricos suficientes para suportar o crescimento populacional, disse Tundisi

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

CIENCIA_2001_6410

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

Segundo ele, pessoas ligadas por computadores tiveram mais contatos pessoais com seus amigos e parentes do que pessoas não-conectadas.

SUMARIO 2:

Segundo Barry Ellman, do Centro para Estudos Urbanos e Comunitários da Universidade de Toronto, Canadá, pessoas ligadas por computadores tiveram mais contatos pessoais com seus amigos e parentes do que pessoas não-conectadas.

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

CIENCIA_2001_6416

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

Uma mudança de apenas uma base [letra] no gene PB2 [que resultou na modificação de um aminoácido na proteína por ele codificada] parece ser a causa da virulência de o H5 N1, o vírus influenza que, em 1997, matou 6 das 18 pessoas infectadas, em Hong Kong, explica. Pesquisadores da Universidade de Wisconsin-Madison (EUA) ainda não sabem exatamente qual o papel do PB2, mas ele parece codificar uma enzima responsável pela indução de um número maior de partículas virais nas células infectadas.

SUMARIO 2:

"Uma mudança de apenas uma base [letra] no gene PB2 [que resultou na modificação de um aminoácido na proteína por ele codificada] parece ser a causa da virulência do influenza", explica. Os cientistas ainda não sabem exatamente qual o papel do PB2, mas ele parece codificar uma enzima responsável pela indução de um número maior de partículas virais nas células infectadas.

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

CIENCIA_2002_22005

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

Mas a equipe internacional de arqueólogos subaquáticos, incluindo brasileiros, que trabalhou na foz do rio Arade, no sul de Portugal, já pôde pelo menos sepultar um mito _o de que ali haveria um navio viking naufragado. Mesmo sem confirmar o relato histórico feito por um cronista árabe, no ano de 996, de que navios vikings teriam afundado a caminho de atacar a cidade de Silves, os arqueólogos puderam encontrar um tesouro de outro tipo: a riqueza de artefatos de várias origens ali depositados pelas sucessivas levas de ocupantes do território português. A coordenação foi do arqueólogo português Francisco Alves, um dos pioneiros na área em Portugal e responsável pelos trabalhos no galeão Nossa Senhora dos Mártires, cujos achados foram a atração central do pavilhão de Portugal na exposição internacional de Lisboa em 1998, a Expo-98.

SUMARIO 2:

Mas a equipe internacional de arqueólogos subaquáticos, incluindo Os brasileiros -Gilson Rambelli, Paulo Bava de Camargo e Flávio Rizzi, pesquisadores do Museu de Arqueologia e Etnologia (MAE) da USP-, que trabalhou na foz do rio Arade, no sul de Portugal, já pôde pelo menos sepultar um mito -o de que ali haveria um navio viking naufragado. Mesmo sem confirmar o relato histórico feito por um cronista árabe, no ano de 996, de que navios vikings teriam afundado a caminho de atacar a cidade de Silves, os arqueólogos puderam encontrar um tesouro de outro tipo :a riqueza de inúmeros fragmentos, principalmente cerâmicos, que vão desde ânforas romanas até utensílios inteiros do século 16, passando pela cerâmica árabe. A coordenação foi do arqueólogo português Francisco Alves, um dos pioneiros na área em Portugal e responsável por os trabalhos no galeão Nossa Senhora dos Mártires, cujos achados foram a atração central do pavilhão de Portugal na exposição internacional de Lisboa em 1998, a Expo-98.

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

CIENCIA_2004_6494

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

Capitaneados por sir David King, o principal assessor científico do governo britânico, os pesquisadores não pouparam esforços para demonstrar que o aquecimento global já está pondo em risco as vidas e a economia humanas em diversas regiões.

SUMARIO 2:

Capitaneados por sir David King, o principal assessor científico do governo britânico, os pesquisadores não pouparam esforços para demonstrar que o aquecimento global causado por esse gás já está pondo em risco as vidas e a economia humanas em diversas regiões.

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

CIENCIA_2005_28747

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

As diferenças como essas não são minúsculas. Depois do primeiro segundo, os americanos olharam mais para o objeto central de um quadro do que para o fundo durante 600 milissegundos, enquanto isso só aconteceu por 40 milissegundos com Pessoas nascidas na China, disse à Folha Richard Nisbett, do Departamento de Psicologia de a Universidade de Michigan em Ann Arbor, nos Estados Unidos.

SUMARIO 2:

"As diferenças não são minúsculas. Depois do primeiro segundo, os americanos olharam mais para o objeto central do que para o fundo durante 600 milissegundos, enquanto isso só aconteceu por 40 milissegundos com os chineses", disse à Folha Richard Nisbett, do Departamento de Psicologia da universidade.

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

CIENCIA_2005_28766

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

Pesquisadores da Fiocruz(Fundação Oswaldo Cruz) identificaram glicoproteínas (grosso modo, proteínas unidas a uma forma de açúcar) no sangue dos gambás que têm essa função antiofídica e esperam utilizar- las não apenas para auxiliar quem sofre acidentes com cobras, mas também para tratar doenças humanas, como câncer e osteoartrite. Conforme as pesquisas progrediram, a equipe descobriu que a resistência não se estende só ao gambá propriamente dito, mas também às cuícas e outros parentes do animal, todos caçadores de cobras, que teriam tido vantagens em desenvolver tais defesas bioquímicas. Seja como for, a DM43 e a DM64 parecem especificamente talhadas para neutralizar os principais efeitos do veneno das serpentes da família das viperídeas, entre as quais se incluem as jararacas. Segundo Jonas Perales, do Laboratório de Toxinologia, a equipe inclusive pediu patentes sobre algumas da ação das substâncias contra doenças como o câncer, mas enquanto o pedido não for aprovado, Perales prefere não revelar exatamente do que se trata.

SUMARIO 2:

Pesquisadores da Fiocruz (Fundação Oswaldo Cruz) identificaram duas moléculas no sangue dos gambás que têm essa função antiofídica e esperam utilizá-las não apenas para auxiliar quem sofre acidentes com cobras, mas também para tratar doenças humanas, como câncer e osteoartrite. Conforme as pesquisas progrediram, a equipe descobriu que a resistência não se estende só ao gambá propriamente dito, mas também às cuícas e outros parentes do animal, todos caçadores de cobras, que teriam tido vantagens em desenvolver tais defesas bioquímicas. Seja como for, a DM43 e a DM64 parecem especificamente talhadas para neutralizar os principais efeitos do veneno das serpentes da família das viperídeas, entre as quais se incluem as jararacas. Segundo o pesquisador, a equipe inclusive pediu patentes sobre algumas dessas aplicações, mas, enquanto o pedido não for aprovado, Perales prefere não revelar exatamente do que se trata.

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

Anexo B - Questionários Sumários SuPor-2

Sumários SuPor-2 Originais e Revisados

10 textos corpus Summit

CIENCIA_2000_17088

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

Pesquisadores do Museu Nacional do Rio de Janeiro anunciaram ontem a descoberta de uma nova espécie de dinossauro no Brasil. Batizado de Santanaraptor placidus, o fóssil é o único a ser encontrado no país com restos de tecido mole, como fibras musculares, vasos sanguíneos e pele. Outra importante descoberta é que, na cadeia evolutiva dos dinossauros, o Santanaraptor ocuparia uma posição no grupo Tyrannoraptora, o mesmo do Tyrannossaurus rex, que habitou os EUA no final da era dos dinos. "O Santanaraptor pode ser a espécie que deu origem ao tiranossauro 68 milhões de anos mais tarde", explicou o geólogo.

SUMARIO 2:

Pesquisadores do Museu Nacional do Rio de Janeiro anunciaram ontem a descoberta de O exemplar de Santanaraptor encontrado pela equipe carioca no Brasil. Batizado de Santanaraptor placidus, o fóssil é o único a ser encontrado no país com restos de tecido mole, como fibras musculares, vasos sanguíneos e pele. Outra importante descoberta é que, na cadeia evolutiva dos dinossauros, o Santanaraptor ocuparia uma posição no grupo Tyrannoraptora, o mesmo do Tyrannossaurus rex, que habitou os EUA no período

Cretáceo (o último da era dos grandes répteis).. O Santanaraptor pode ser a espécie que deu origem ao tiranossauro 68 milhões de anos mais tarde, explicou Alexander Kellner.

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

CIENCIA_2000_17112

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

Um estudo publicado na edição de hoje da revista "Science" afirma que 1,75 bilhão de pessoas já enfrentam severa escassez de água no planeta. Por severa escassez de água potável, entende-se, segundo a ONU, o uso de mais de 40% das reservas do líquido disponíveis em uma região para consumo industrial, doméstico e agrícola. A projeção dos cientistas para o ano 2025 é que 3,3 bilhões de pessoas não tenham mais água para irrigação _a atividade humana que mais consome o líquido. "A demanda aumenta de forma drástica no mundo todo", afirmou o especialista em recursos hídricos José Galizia Tundisi, do Instituto Internacional de Ecologia, em São Carlos (SP).

SUMARIO 2:

Um estudo publicado na edição de hoje da revista Science afirma que 1,75 bilhão de pessoas já enfrentam severa escassez de água nO mundo. Por severa escassez de água potável, entende- se, segundo a ONU (Organização das Nações Unidas), o uso de mais de 40 das reservas do líquido disponíveis em uma região para consumo industrial, doméstico e agrícola. A projeção de a equipe de Vörösmarty para o ano 2025 é que 3,3 bilhões de pessoas não tenham mais água para irrigação -a atividade humana que mais consome o líquido. A demanda aumenta de forma drástica no mundo todo, afirmou o especialista em recursos hídricos José Galizia Tundisi, do Instituto Internacional de Ecologia, em São Carlos (SP).

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

CIENCIA_2000_17113

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

Um tratamento para a obesidade que faz você comer mais, perder peso e reduzir a taxa de gordura do corpo é o que sugere um estudo britânico publicado hoje na revista científica Nature. Por enquanto é só sugestão: o tratamento foi testado em camundongos. No centro do método está um gene humano descoberto recentemente, o UCP-3, cujos mecanismos de ação ainda não são totalmente conhecidos. O gene UCP-3 foi inserido em camundongos e manipulado para produzir, em excesso, a proteína determinada por ele. A porcentagem de tecido adiposo (gordura) sobre o volume total do corpo de Os camundongos com essa alteração genética também diminuiu -em os machos, em 44; nas fêmeas, em 57. Esse é um alvo viável para remédios contra a obesidade, disse um dos autores, John Clapham, da empresa farmacêutica SmithKline Beecham, que fez o estudo em colaboração com a Universidade de Cambridge, Reino Unido.

SUMARIO 2:

Um tratamento para a obesidade que faz você comer mais, perder peso e reduzir a taxa de gordura do corpo é o que sugere um estudo britânico publicado hoje na revista científica “Nature”. Por enquanto é só sugestão: o tratamento foi testado em camundongos. No centro do método está um gene humano descoberto recentemente, o UCP-3, cujos mecanismos de ação ainda não são totalmente conhecidos. O gene UCP-3 foi inserido em camundongos e manipulado para produzir, em excesso, a proteína determinada por ele. A porcentagem de tecido adiposo (gordura) sobre o volume total do corpo dos bichos também diminuiu _nos machos, em 44 %; nas fêmeas, em 57 %. “Esse é um alvo viável para remédios contra a obesidade”, disse um dos autores, John Clapham, da empresa farmacêutica SmithKline Beecham, que fez o estudo em colaboração com a Universidade de Cambridge, Reino Unido.

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

CIENCIA_2001_6410

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

Ao contrário do que muita gente pensa, a internet não está reduzindo os contatos entre as pessoas nem substituindo-os por relações impessoais conduzidas por computador. Um dos exemplos que ele apresenta é o de um estudo feito em um subúrbio de Toronto, segundo o qual as pessoas "plugadas" em uma rede local conheciam três vezes mais vizinhos do que os não-conectados.

SUMARIO 2:

Ao contrário do que muita gente pensa, redes de computadores não está reduzindo os contatos entre as pessoas nem substituindo- os por relações impessoais conduzidas por computador. Um dos exemplos que Barry Ellman apresenta é o de um estudo feito num subúrbio de Toronto, segundo o qual as pessoas plugadas em uma rede local conheciam três vezes mais vizinhos do que os não-conectados.

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

CIENCIA_2002_22005

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

Mas a equipe internacional de arqueólogos subaquáticos, incluindo brasileiros, que trabalhou na foz do rio Arade, no sul de Portugal, já pôde pelo menos sepultar um mito -o de que ali haveria um navio viking naufragado. Os brasileiros -Gilson Rambelli, Paulo Bava de Camargo e Flávio Rizzi, pesquisadores do Museu de Arqueologia e Etnologia (MAE) da USP- acharam mesmo restos de um ou dois navios antigos, possivelmente portugueses dos século 15 ou 16, da chamada tradição ibero-atlântica. Mesmo sem confirmar o relato histórico feito por um cronista árabe, no ano de 996, de que navios vikings teriam afundado a caminho de atacar a cidade de Silves, os arqueólogos puderam encontrar um tesouro de outro tipo: a riqueza de inúmeros fragmentos, principalmente cerâmicos, que vão desde ânforas romanas até utensílios inteiros do século 16, passando pela cerâmica árabe. Rambelli, que coordenou a participação brasileira, está lançando nesta semana o livro Arqueologia Até Debaixo DÁgua (Editora Maranta, São Paulo, 2002).

SUMARIO 2:

Mas a equipe internacional de arqueólogos subaquáticos, incluindo brasileiros, que trabalhou na foz do rio Arade, no sul de Portugal, já pôde pelo menos sepultar um mito _o de que ali haveria um navio viking naufragado. Os brasileiros _Gilson Rambelli, Paulo Bava de Camargo e Flávio Rizzi, pesquisadores do Museu de Arqueologia e Etnologia (MAE) da USP _ acharam mesmo restos de um ou dois navios antigos, possivelmente portugueses dos século 15 ou 16, da chamada tradição ibero-atlântica. Mesmo sem confirmar o relato histórico feito por um cronista árabe, no ano de 996, de que navios vikings teriam afundado a caminho de atacar a cidade de Silves, os arqueólogos puderam encontrar um tesouro de outro tipo: a riqueza de artefatos de várias origens ali depositados pelas sucessivas levas de ocupantes do território português. Rambelli, que coordenou a participação brasileira, está lançando nesta semana o livro "Arqueologia Até Debaixo D'Água" (Editora Maranta, São Paulo, 2002).

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

Biotecnólogos e ambientalistas travaram uma rara aliança na semana passada para comemorar um episódio singular: duas vacas deram cria em Iowa, EUA. Os filhotes não são delas, mas clones de outra espécie, o banteng, um tipo de gado ameaçado de extinção. Os dois nascimentos, ocorridos em 1º e 3 de abril, marcam o início de uma nova fase para um projeto que já atraía interesse _o Frozen Zoo (ou Zoológico Congelado, na tradução para o português). A idéia, iniciada em 1976, era coletar e preservar criogenicamente (em baixas temperaturas) amostras celulares de animais ameaçados de extinção, com a esperança de estudá-los e, quem sabe, ressuscitá-los quando a tecnologia assim o permitisse. O material vai desde os notórios pandas e condores até os menos conhecidos bantengs _parentes asiáticos raros do gado comum que estão à beira do esquecimento. A primeira tentativa de trazer um membro do Frozen Zoo de volta do mundo dos animais perdidos foi com um gauro, outra espécie rara de gado.

SUMARIO 2:

Biotecnólogos e ambientalistas travaram uma parceria entre a companhia de biotecnologia Advanced Cell Technology, de Massachusetts, o Centro Sioux, de Iowa, e o Centro para Reprodução de Espécies Ameaçadas da Sociedade Zoológica de San Diego, na Califórnia na semana passada para comemorar um episódio singular: duas vacas deram cria em Iowa, EUA. Os filhotes não são de elas, mas clones de outra espécie, o banteng, um tipo de gado ameaçado de extinção. Os dois nascimentos, ocorridos em 1º e 3 de abril, marcam o início de uma nova fase para um projeto que já atraía interesse -o Frozen Zoo (ou Zoológico Congelado, na tradução para o português) A idéia, iniciada em 1976, era coletar e preservar criogenicamente (em baixas temperaturas) amostras celulares de animais ameaçados de extinção, com a esperança de estudar- los e, quem sabe, ressuscitar- los quando a tecnologia assim o permitisse. O material vai desde os notórios pandas e condores até os menos conhecidos bantengs -parentes asiáticos raros do gado comum que estão à beira do esquecimento. A primeira tentativa de trazer um membro do Frozen Zoo de volta do mundo dos animais perdidos foi com um gauro, outra espécie rara de gado.

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

CIENCIA_2003_24219

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

Os ministérios da Agricultura e da Ciência e Tecnologia defenderam ontem o uso da soja transgênica na produção do biodiesel para abastecer parte da frota nacional de veículos. A idéia foi lançada pelo ministro Roberto Amaral (Ciência e Tecnologia) e detalhada ontem durante a abertura do 1º Congresso Internacional de Biodiesel, realizado em Ribeirão Preto e promovido pela USP (Universidade de São Paulo) da cidade. A intenção do governo é usar parte da soja transgênica já plantada no país, e que está com seu consumo proibido, na produção do combustível. O projeto, desenvolvido pela USP de Ribeirão, consegue produzir o biodiesel a partir da mistura de óleo vegetal _incluindo o de soja _ e etanol, álcool derivado da cana-de-açúcar.

SUMARIO 2:

Os ministérios da Agricultura e de o MCT (Ministério da Ciência e Tecnologia) defenderam ontem o uso da soja transgênica na produção do biodiesel para abastecer parte da frota nacional de veículos. A idéia foi lançada pelo ministro Roberto Amaral (Ciência e Tecnologia) e detalhada ontem durante a abertura do 1º Congresso Internacional de Biodiesel, realizado em Ribeirão Preto e promovido pela USP (Universidade de São Paulo) da cidade. A intenção do governo é usar parte da soja transgênica já plantada no Brasil, e que está com seu consumo proibido, na produção do combustível. O projeto, desenvolvido pela USP de Ribeirão, consegue produzir o biodiesel a partir da mistura de óleo vegetal -incluindo o de soja- e etanol, álcool derivado da cana-de-açúcar.

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

CIENCIA_2004_6480

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

O animal é um clone gerado a partir de um clone _a vaca Vitória, que havia sido clonada em 2001. Ela surgiu a partir de células isoladas de um pedaço de pele retirado da orelha da vaca Vitória, que foi o primeiro clone bovino da América Latina, nascida em 2001. “O clone do clone coloca o Brasil na vanguarda científica desse assunto, como já está no seqüenciamento [soletração] de genoma”, afirmou Rodrigues.

SUMARIO 2:

A terceira tentativa do órgão de criar um clone a partir de outro é um clone gerado a partir de um clone -a vaca Vitória, que havia sido clonada em 2001. Ela surgiu a partir de células isoladas de um pedaço de pele retirado da orelha da vaca Vitória, que foi o primeiro clone bovino da América Latina, nascida em 2001. O clone do clone coloca o Brasil na vanguarda científica desse assunto, como já está no seqüenciamento [soletração] de genoma, afirmou o ministro Roberto Rodrigues (Agricultura).

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

CIENCIA_2004_26415

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

Para um desavisado parece até obsessão freudiana, mas Bioantropólogo da Universidade MacMaster, no Canadá está pedindo a todos os seus conhecidos a maior quantidade de fezes possível -quanto mais velhas, melhores. Estamos recolhendo amostras de coprólitos [fezes fossilizadas] de duas cavernas em Israel com 40 mil anos, onde provavelmente Cro-Magnons [os primeiros humanos modernos] e neandertais viveram lado a lado, contou o pesquisador durante a reunião da AAAS (Associação Americana para

o Avanço da Ciência). Dadas as características muito especiais de preservação que as fezes podem alcançar, há grandes chances de elas terem preservado mais DNA do que o que se pode extrair de ossos, proteínas e outras moléculas.

SUMARIO 2:

Para um desavisado parece até obsessão freudiana, mas Hendrik Poynar está pedindo a todos os seus conhecidos a maior quantidade de fezes possível _quanto mais velhas, melhores. “stamos recolhendo amostras de coprólitos [fezes fossilizadas] de duas cavernas em Israel com 40 mil anos, onde provavelmente Cro-Magnons [os primeiros humanos modernos] e neandertais viveram lado a lado”, contou o pesquisador durante a reunião da AAAS (Associação Americana para o Avanço da Ciência). Dadas as características muito especiais de preservação que as fezes podem alcançar, há grandes chances de elas terem preservado mais DNA do que o que se pode extrair de ossos, bem como proteínas e outras moléculas.

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

CIENCIA_2004_26423

Avaliar em relação à legibilidade e informatividade.

A legibilidade é uma qualidade que determina a facilidade de leitura.

A informatividade é o que determina o nível de informação de cada sumário.

SUMARIO 1:

Como se a lista de problemas de saúde causados por os alimentos gordurosos já não fosse imensa, cientistas americanos acabam de adicionar a ela o dano ligado ao mal de Alzheimer, doença degenerativa do cérebro que mais afeta idosos pelo mundo. De acordo com novas análises em seres humanos, o colesterol e a ceramida, ambas moléculas de gordura, impulsionam a morte de células nervosas que caracteriza a doença. Colegas de Mark Mattson, do Instituto Nacional do Envelhecimentos dos EUA também revelaram novas evidências de que o consumo de vitaminas como suplemento alimentar e provenientes de vegetais seria capaz de prevenir o advento da doença, e mesmo de que as estatinas, hoje usadas para o combate ao colesterol na corrente sangüínea, também poderiam evitar o aparecimento do mal de Alzheimer. Ele e seus colegas descobriram uma espécie de combinação letal entre o aparecimento

de a proteína beta-amilóide e a presença de colesterol e de ceramida nos neurônios de pessoas mortas com o mal. Experimentos coordenados por Carl Cotman, da Universidade da Califórnia em Irvine, mostraram resultados animadores em cães na fase de envelhecimento (a partir dos 9 anos de vida), na qual os animais podem desenvolver uma doença que lembra Alzheimer.

SUMARIO 2:

Como se a lista de problemas de saúde causados pelos alimentos gordurosos já não fosse imensa, cientistas americanos acabam de adicionar a ela o dano ligado ao mal de Alzheimer, doença degenerativa do cérebro que mais afeta idosos pelo mundo. De acordo com novas análises em seres humanos, o colesterol e a ceramida, ambas moléculas de gordura, impulsionam a morte de células nervosas que caracteriza a doença. Colegas de Mattson também revelaram novas evidências de que o consumo de vitaminas como suplemento alimentar e provenientes de vegetais seria capaz de prevenir o advento da doença, e mesmo de que as estatinas, hoje usadas para o combate ao colesterol na corrente sangüínea, também poderiam evitar o aparecimento do mal de Alzheimer. Mattson e colegas descobriram uma espécie de combinação letal entre o aparecimento da beta-amilóide e a presença de colesterol e de ceramida nos neurônios de pessoas mortas com o mal. Experimentos coordenados por Carl Cotman, da Universidade da Califórnia em Irvine, mostraram resultados animadores em cães na fase de envelhecimento (a partir dos 9 anos de vida), na qual os animais podem desenvolver uma doença que lembra Alzheimer.

Você percebe alguma diferença de legibilidade entre os dois sumários?

( ) Sim. Qual dos dois é mais legível: 1( ) ou 2 ( )

( ) Não.

Você percebe alguma diferença de informatividade?

( ) Sim. Qual dos dois é mais informativo: 1( ) ou 2 ( )

( ) Não.

Anexo C - Tabela Dados Rouge SuPor-2

Tabela 22: Resultados do Rouge - Sumários SuPor-2

	SUPOR-2-ORIGINAL			SUPOR-2-CORRIGIDO		
NOME TEXTO	COBERTURA	PRECISÃO	F-MEASURE	COBERTURA	PRECISÃO	F-MEASURE
CIENCIA_2000_6380	81,58	72,94	77,02	81,58	68,13	74,25
CIENCIA_2000_6381	59,72	66,15	62,77	56,94	42,27	48,52
CIENCIA_2000_6389	68,57	60,76	64,43	71,43	53,76	61,35
CIENCIA_2000_6391	67,74	59,16	63,16	67,74	57,53	62,22
CIENCIA_2000_17082	53,41	47,96	50,54	53,41	46,54	49,74
CIENCIA_2000_17088	76,34	67,62	71,72	80,65	66,37	72,82
CIENCIA_2000_17101	70,30	64,55	67,30	70,30	64,55	67,30
CIENCIA_2000_17108	63,04	52,25	57,14	63,04	52,25	57,14
CIENCIA_2000_17109	68,54	51,26	58,65	68,54	51,26	58,65
CIENCIA_2000_17112	31,00	24,41	27,31	37,00	26,81	31,09
CIENCIA_2000_17113	51,15	43,51	47,02	54,20	45,22	49,31
CIENCIA_2001_6406	51,79	60,42	55,77	51,79	60,42	55,77
CIENCIA_2001_6410	58,00	42,65	49,15	64,00	46,38	53,78
CIENCIA_2001_6414	68,61	57,84	62,77	68,61	57,28	62,43
CIENCIA_2001_6416	40,79	49,21	44,60	53,95	43,62	48,24
CIENCIA_2001_6423	76,74	71,74	74,16	60,47	57,78	59,09
CIENCIA_2001_19858	75,71	69,07	72,24	76,27	67,50	71,62
CIENCIA_2002_6441	80,36	53,57	64,29	78,57	51,77	62,41
CIENCIA_2002_22005	66,67	67,06	66,86	69,59	66,85	68,20
CIENCIA_2002_22010	64,74	69,10	66,85	64,21	68,93	66,49
CIENCIA_2002_22015	54,93	56,80	55,85	56,34	52,86	54,55
CIENCIA_2002_22023	61,79	47,50	53,71	65,04	48,49	55,56
CIENCIA_2002_22027	75,82	69,05	72,27	67,97	63,03	65,41
CIENCIA_2002_22029	75,12	74,77	74,94	78,40	71,06	74,55
CIENCIA_2003_6457	68,61	64,13	66,29	72,09	59,62	65,26
CIENCIA_2003_6465	86,36	76,77	81,28	88,64	74,29	80,83
CIENCIA_2003_6472	83,02	50,00	62,41	83,02	50,00	62,41
CIENCIA_2003_24212	65,24	57,84	61,32	79,88	59,55	68,23
CIENCIA_2003_24219	63,30	50,74	56,33	65,14	50,71	57,03
CIENCIA_2003_24226	55,84	53,92	54,86	55,33	52,66	53,96
CIENCIA_2004_6480	41,94	49,37	45,35	47,31	46,32	46,81
CIENCIA_2004_6488	56,00	82,35	66,67	56,00	82,35	66,67
CIENCIA_2004_6494	76,19	55,81	64,43	76,19	53,33	62,75
CIENCIA_2004_26415	73,79	63,33	68,16	75,73	63,42	69,03
CIENCIA_2004_26417	40,12	43,92	41,94	40,12	43,62	41,80
CIENCIA_2004_26423	56,78	63,81	60,09	59,32	62,78	61,00
CIENCIA_2004_26425	55,10	59,34	57,14	55,10	59,02	56,99
CIENCIA_2005_6507	56,36	81,58	66,67	58,18	74,42	65,31
CIENCIA_2005_6514	100,00	87,18	93,15	94,12	81,01	87,08
CIENCIA_2005_6515	60,94	52,70	56,52	60,94	53,43	56,94
CIENCIA_2005_6518	58,33	52,13	55,06	57,14	46,60	51,34

Tabela 23: Resultados do Rouge - Sumários SuPor-2 (continuação)

	SUPOR-2-ORIGINAL			SUPOR-2-CORRIGIDO		
NOME DO TEXTO	COBERTURA	PRECISÃO	F-MEASURE	COBERTURA	PRECISÃO	F-MEASURE
CIENCIA_2005_28743	56,58	49,14	52,60	56,58	49,14	52,60
CIENCIA_2005_28747	57,14	56,52	56,83	57,14	54,17	55,62
CIENCIA_2005_28752	60,12	51,31	55,37	59,51	50,52	54,65
CIENCIA_2005_28754	68,45	87,58	76,84	69,90	86,23	77,21
CIENCIA_2005_28755	55,62	49,25	52,24	54,49	46,41	50,13
CIENCIA_2005_28756	62,58	57,74	60,06	62,58	57,06	59,69
CIENCIA_2005_28764	63,18	67,2	65,13	61,19	65,78	63,40
CIENCIA_2005_28766	64,00	64,98	64,48	74,00	61,93	67,43
CIENCIA_2005_28774	60,87	52,72	56,50	60,87	52,72	56,50
MÉDIA	63,60	59,34	60,94	64,70	57,15	60,26