Alinhamento automático de textos e sumários multidocumento

Verônica Agostini

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura:_____________________________

Alinhamento automático de textos e sumários multidocumento

Verônica Agostini

Orientador: **Prof. Dr. Thiago Alexandre Salgueiro Pardo**

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional. *EXEMPLAR DE DEFESA*

USP – São Carlos
Fevereiro de 2014

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

Agostini, Verônica
A275a Alinhamento automático de textos e sumários multidocumento / Verônica Agostini; orientador Thiago Alexandre Salgueiro Pardo. -- São Carlos, 2014.
119 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2014.

1. alinhamento. 2. sumarização automática. 3. sumarização multidocumento. 4. alinhamento sentencial. I. Pardo, Thiago Alexandre Salgueiro, orient. II. Título.

Agradecimentos

Agradeço ao meu orientador, pelo ensinamento e pelas palavras de apoio; aos meus pais, pelo apoio constante e indispensável; aos meus amigos, que muito ouviram e me ajudaram; à Renata, pela parceria em duas anotações de córpus; ao Roque, pela ajuda com o desenvolvimento das abordagens propostas neste trabalho; aos colegas de laboratório e a todos que me ajudaram de alguma forma a concluir este trabalho de mestrado.

"...before tomorrow comes, you could change everything…"

Resumo

Com o excesso de informação disponível *online*, a Sumarização Automática tornou-se uma área de bastante interesse na grande área da Inteligência Artificial. Alguns autores tentaram caracterizar o processo de sumarização para compreender melhor como sumarizadores o realizam. O alinhamento de um sumário e seus textos fonte pode ser encarado como uma caracterização desse processo. Com relação à sumarização automática, a técnica de alinhamento consiste em obter relações entre segmentos de um ou vários textos e seu sumário e, da forma que o conteúdo de um segmento esteja contido no outro. Uma vez obtidas essas relações, torna-se possível (i) aprender como sumarizadores profissionais realizam a sumarização, (ii) explicitar regras e modelos para a sumarização, e (iii) criar métodos automatizados utilizando as regras e modelos explicitados, o que traz uma contribuição à Sumarização Automática. Neste trabalho, foram propostas três abordagens dentro das abordagens superficiais e profundas do Processamento de Língua Natural para realizar os alinhamentos de forma automática. A primeira utiliza três métodos superficiais, sendo eles *Word overlap*, tamanho relativo e posição relativa. A segunda caracteriza-se em uma técnica de alinhamento com mais conhecimento linguístico, pois nela foi utilizada uma teoria discursiva, a CST (*Cross-Document Structure Theory*). A terceira utiliza Aprendizado de Máquina, caracterizando uma abordagem híbrida dada a característica de seus atributos superficiais e profundos, relativo à primeira e à segunda abordagem. Uma avaliação comparativa entre elas, e também entre um trabalho da literatura, foi realizada. Quando os dados do aprendizado de máquina eram balanceados, foi atingido o valor de 97,2% de medida-F, maior valor encontrado. O método superficial *Word overlap* também obteve um bom resultado, sendo ele 66,2% de medida-F.

Abstract

With the huge amount of online information, the Automatic Summarization has become an area of great interest in the Artificial Intelligence area. Some authors have tried to characterize the process of summarization to aim a better understanding of how the summarizers perform the summarization. The alignment of an abstract and its source texts can be seen as a characterization of this process. Regarding the automatic summarization, the alignment technique consists in obtaining relations between segments of one, or more text, and its abstract, in a way that the segment content is contained in the other segment. Once obtained these relationships, it becomes possible to (i) learn how professional summarizers perform the summarization, (ii) explicit rules and models for summarization, and (iii) create automated methods using the rules and the models, which brings a contribution to the Automatic Summarization area. In this dissertation, three approaches within the Natural Language Processing superficial and deep approaches have been proposed to carry the alignments automatically. The first uses three superficial methods, namely "Word overlap", "relative size" and "relative position". The second is an alignment technique with more linguistic knowledge, because was used a discursive theory, the CST (Cross-document Structure Theory). The third uses Machine Learning, featuring a hybrid approach given the characteristics of its deep and superficial attributes, relative to the first and second approaches. A comparative evaluation among them was performed. When the machine learning data were balanced, the value achieved was 97.2% of F-measure, the highest value found. The superficial method "Word overlap" also achieved a good result, which is 66.2% of F-measure.

Índice

Lista de Quadros

Lista de Figuras

Lista de Tabelas

Capítulo 1. Introdução

1.1. Contextualização

Com a grande quantidade de informação disponível *online*, a Sumarização Automática tornou-se uma área de bastante interesse no Processamento de Língua Natural, subárea da Inteligência Artificial.

A sumarização automática, de acordo com Mani (2001), tem o objetivo de extrair conteúdo de uma fonte de informação e apresentá-lo ao usuário de uma forma condensada e de uma maneira suscetível aos interesses do usuário ou de uma aplicação.

A atividade de produzir sumários (resumos) é uma atividade comum que as pessoas realizam diariamente, por exemplo, quando alguém deseja narrar uma história para outra pessoa. A sumarização é útil também quando alguém procura saber do que se trata um livro, um filme ou mesmo um artigo ou uma dissertação, lendo, dessa forma, um sumário sobre o conteúdo dos mesmos.

Além disso, sumários são bastante úteis, pois é sabido que nem toda parcela de informação é relevante a quem procura, e muito do que os meios de comunicação provêm são informações repetidas e até contraditórias. Nesse contexto, torna-se útil a produção automática de sumários provenientes de mais de um texto. A sumarização automática multidocumento é, portanto, a produção de um único texto condensado que contenha as informações mais relevantes dos textos fonte que versem sobre um assunto em comum, ao mesmo tempo em que são removidas redundâncias e que são levadas em conta as similaridades e diferenças dos textos (Mani, 2001).

A sumarização automática surgiu depois da sumarização humana. É sabido que sumarizadores profissionais realizam a sumarização em alguns passos (Pinto Molina, 1995; Cremmins, 1996) e também que utilizam técnicas, como a de corta e cola (Jing e McKeown, 1999) para produzir sumários monodocumento, porém não há um campo profissional envolvendo a sumarização multidocumento, pois a ideia de se ter um único sumário a partir de vários textos é um tanto externa ao mundo da sumarização profissional (Mani, 2001).

Exemplos de sumários multidocumento, retirados do córpus CSTNews (Cardoso *et al.*, 2011b), podem ser vistos nos Quadro 1 e 2. O sumário do Quadro 1 foi produzido por um humano, enquanto que o sumário do Quadro 2 foi produzido por um sumarizador automático, o CSTSumm (Castro Jorge e Pardo, 2010).

A brasileira Fabiana Murer conquistou a medalha de ouro no salto com vara ao saltar 4m60, um novo recorde pan-americano, 20 cm a mais que sua antiga marca. A medalha de prata ficou com a americana April Steiner com 4m40 e a de bronze com a cubana Yarisley Silva com 4m30.
Fabiana conseguiu o ouro em três tentativas. Tentou ainda bater o próprio recorde sul-americano de 4m66, mas não conseguiu. A outra brasileira, Joana Costa, ficou na quinta posição, com 4m20, mostrando que o nervosismo pode atrapalhar as competições em casa.

Quadro 1: Exemplo de sumário multidocumento humano

O PRIMEIRO - Murer salta para quebrar recorde pan-americano; primeiro ouro do atletismo.
RIO - Depois da queda de April Steiner, a brasileira Fabiana Murer leva a medalha de ouro no salto com vara, com 4m50 - novo recorde pan-americano.
A brasileira Fabiana Murer conquistou o primeiro ouro do atletismo para o Brasil, nesta segunda-feira, na prova de salto com vara.
A medalha de prata ficou com a americana April Steiner, com a marca de 4m40 e o bronze foi para a cubana Yarisley Silva, com 4m30.

Quadro 2: Exemplo de sumário multidocumento automático

Como pode ser evidenciado pelos exemplos de sumário dos Quadro 1 e 2, a sumarização automática multidocumento traz alguns desafios, chamados de fenômenos multidocumento, que não estão presentes na sumarização monodocumento, como o supracitado problema da redundância. Além disso, é necessário levar em conta os diferentes estilos de escrita dos autores dos textos fonte, a ordenação dos eventos no tempo, as informações contraditórias, etc.. Tais questões podem ser encaradas como problemas para a sumarização, mas também podem ser úteis. No caso dos exemplos, a primeira sentença do Quadro 2 contém a informação que a atleta Fabiana Murer saltou e ganhou uma medalha de ouro. Porém essa informação também está presente na terceira sentença do mesmo sumário, caracterizando um exemplo de informação redundante. Já o sumário humano, apresentado no Quadro 1, não possui esse tipo de problema.

Os sumarizadores automáticos são criados dentro de duas principais abordagens de sumarização: a superficial e a profunda. Na abordagem superficial, é utilizado pouco conhecimento linguístico para realizar a sumarização, como informações estatísticas/empíricas sobre os textos; e, na abordagem profunda, é utilizado mais conhecimento linguístico, como informações semânticas e de discurso. O CSTSumm é um exemplo de sumarizador automático da abordagem profunda, pois utiliza métodos de seleção de conteúdo baseado em uma teoria discursiva para realizar a sumarização.

Ainda, juntamente com a sumarização automática, existem diversas outras aplicações de processamento de língua natural, como a tradução automática e a simplificação textual, e a técnica do alinhamento pode fazer parte de várias delas.

O alinhamento é uma tarefa que consiste em relacionar segmentos textuais, sejam eles palavras, sentenças, parágrafos, ou até documentos inteiros. Existem várias formas de alinhamento nas diversas aplicações do processamento de língua natural, como o alinhamento utilizado para auxiliar a tradução automática (por exemplo, Gale e Church, 1993; Yamada e Knight, 2001.), em que palavras ou sentenças de textos acompanhados de suas traduções são alinhadas. É também um exemplo o alinhamento de segmentos textuais de perguntas e respostas (por exemplo, Soricut e Brill, 2004); o alinhamento entre sentenças originais com versões simplificadas das mesmas, na aplicação de simplificação textual (por exemplo, Specia, 2010); e o alinhamento na sumarização automática (por exemplo, Marcu, 1999; Hirao *et al.*, 2004), que é o foco deste trabalho.

Vários propósitos são buscados quando se realiza o alinhamento, e o propósito buscado neste trabalho de mestrado é auxiliar a sumarização automática citada anteriormente. Uma das formas que a sumarização automática pode ser auxiliada com o alinhamento é com a descoberta de regras de sumarização. Além disso, o alinhamento pode ajudar no entendimento dos fenômenos multidocumento, como o fenômeno da redundância. O alinhamento pode, também, auxiliar na caracterização de sumários multidocumento, de forma a descobrir quais transformações são realizadas por humanos quando eles sumarizam textos. Com uma caracterização mais detalhada dos sumários multidocumento, seria possível refinar as técnicas de sumarização automática, melhorando o desempenho dos sumarizadores automáticos e por sua vez a qualidade dos sumários resultantes dos mesmos.

No caso de alinhamento sentencial na tradução automática, que se trata de alinhamento entre um texto e sua tradução, todas, ou praticamente todas, as sentenças dos textos são alinhadas. Já na sumarização automática, muito provavelmente restarão sentenças do(s) texto(s) fonte(s) sem alinhamento, pois se trata de um alinhamento de texto(s) fonte(s) com seu sumário, e o sumário é menor que os textos que o originaram, contendo apenas a informação mais relevante dos textos. Outra diferença presente nesse tipo de alinhamento é o fato de, muitas vezes, uma sentença do sumário ter sido gerada de diversas outras sentenças no(s) texto(s) fonte(s). Um exemplo de alinhamento sentencial na sumarização pode ser visto a seguir (Quadro 3 a 5). Esse alinhamento foi realizado manualmente entre dois textos fonte e seu sumário humano, e as sentenças grifadas dos textos fonte representam sentenças que foram alinhadas. O alinhamento efetivo é apresentado no Quadro 6. Nos Quadro 3 a 4, D indica o número do documento, e S indica o número da sentença. Na expressão S_Sn, contida no Quadro 5, o primeiro S indica “sumário” e o segundo, “sentença”.

[D1_S1] SÃO LUÍS - Após quase 24 horas de tensão, terminou no fim da manhã desta quarta-feira a rebelião na Central de Custódia de Presos de Justiça (CCJP) no Maranhão.

[D1_S2] Depois que os presos entregaram o revólver usado para dar início ao motim, a Tropa de Choque da Polícia Militar entrou no presídio e liberou os 30 reféns - sendo 16 crianças.

[D1_S3] Alguns menores saíram desmaiados e foram conduzidos para o atendimento médico.

[D1_S4] Quatro pessoas teriam ficado feridas.

[D1_S5] A Tropa de Choque entrou no presídio depois que Charlene Ribeiro da Silva, esposa do detento Bruno Monteiro da Silva - suspeito de chefiar a rebelião - conversou com o diretor da unidade. [D1_S6] Segundo ela, o diretor assinou um termo de compromisso onde dá garantias de que os presos não serão torturados depois do motim e que Bruno será transferido para outra unidade. [D1_S7] - O preso Bruno agiu de forma traiçoeira e covarde, sem dar oportunidade para os agentes que distribuíam lanches às crianças. Ele saiu atirando, acertando a cabeça de um agente e o braço de outro - contou o secretário-adjunto de Administração Penitenciária, Sidones Cruz. [D1_S8] A Secretaria de Segurança deve abrir sindicância para averiguar como a arma entrou no presídio. [D1_S9] <u>O motim começou durante a festa do Dia das Crianças.</u> [D1_S10] As negociações foram feitas por um pequeno buraco no muro do presídio. [D1_S11] A água e a luz da unidade chegaram a ser cortadas. [D1_S12] A cadeia abriga 203 detentos, mas só tem capacidade para 80. [D1_S13] Neste mesmo pavilhão, no início deste mês, os presos quebraram objetos e fizeram um túnel para tentar uma fuga em massa. [D1_S14] Até agora as celas não foram reparadas. [D1_S15] Esta é a terceira rebelião em São Luís só neste mês.

Quadro 3: Documento 1

[D2_S1] <u>Terminou a rebelião de presos no Centro de Custódia de Presos de Justiça (CCPJ), em São Luís, no começo da tarde desta quarta-feira (17).</u> [D2_S2] <u>Os presos entregaram as armas e a polícia faz uma revista dentro da unidade.</u> [D2_S3] <u>O motim começou durante a festa do Dia das Crianças, realizada na terça-feira (16).</u> [D2_S4] <u>As 16 crianças e 14 adultos foram libertados.</u> [D2_S5] Segundo informações da polícia, o líder da rebelião foi transferido para o Presídio de Pedrinhas, na capital maranhense. [D2_S6] Os presos receberam garantias, por parte do diretor da unidade, de que não

haveria represálias e novas transferências. [D2_S7] Os presos tentaram fugir durante a festa, mas o plano foi descoberto. [D2_S8] <u>No começo da rebelião quatro pessoas ficaram feridas, entre elas uma auxiliar de enfermagem e um agente de polícia que trabalham no presídio.</u> [D2_S9] A unidade ficou sem luz e água e as negociações para a libertação dos reféns foi retomada na manhã desta quarta-feira. [D2_S10] Segundo informações da polícia, os presos temiam uma transferência em massa depois de terem iniciado uma outra rebelião durante a greve de policiais no estado, na semana passada. [D2_S11] A CCPJ tem capacidade para cerca de 80 presos, mas abriga 203 homens.

Quadro 4: Documento 2

[S_S1] Terminou a rebelião de presos no Centro de Custódia de Presos de Justiça (CCPJ), em São Luís, no começo da tarde desta quarta-feira (17). [S_S2] O motim começou durante a festa do Dia das Crianças. [S_S3] Depois que os presos entregaram o revólver usado para dar início ao motim, a Tropa de Choque da Polícia Militar entrou no presídio e liberou os 30 reféns - sendo 16 crianças. [S_S4] Alguns menores saíram desmaiados e foram conduzidos para o atendimento médico. [S_S5] Quatro pessoas teriam ficado feridas.

Quadro 5: Sumário

S_S1 ↔ D1_S1, D2_S1 S_S2 ↔ D1_S9, D2_S3 S_S3 ↔ D1_S2, D2_S2, D2_S4 S_S4 ↔ D1_S3 S_S5 ↔ D1_S4, D2_S8

Quadro 6: Alinhamento

Como pode ser visto no Quadro 6, a sentença 1 do sumário foi alinhada com as sentenças 1 do documento 1, e 1 do documento 2; a sentença 2 do sumário foi alinhada com as sentenças 9 do documento 1, e 3 do documento 2; a sentença 3 do sumário foi alinhada com as sentenças 2 do documento 1, 2 do documento 2, e 4 do documento 2; e assim por diante.

O critério utilizado para o alinhamento é o de sobreposição de conteúdo, dessa forma, são alinhadas duas sentenças se elas contêm a mesma informação, ou parte dessa informação. Por exemplo, no caso do alinhamento de S_S5 com D2_S8, sendo as sentenças alinhadas "Quatro pessoas teriam ficado feridas." e "No começo da rebelião quatro pessoas ficaram feridas, entre elas uma auxiliar de enfermagem e um agente de polícia que trabalham no presídio.", as duas possuem a informação de que pessoas ficaram feridas no incidente.

Como pode ser verificado pelas sentenças sublinhadas nos Quadro 3 e 4, menos da metade das sentenças dos textos foram alinhadas (aproximadamente 38%). Uma sentença do sumário normalmente é alinhada a mais de uma sentença dos textos fonte. Esse tipo de alinhamento é referenciado por 1-N (alinhamento de 1 unidade textual com mais de uma unidade textual). Isso acontece porque os textos a serem sumarizados falam sobre o mesmo assunto, e são as informações mais relevantes (na maioria das vezes mais redundantes) que compõem o sumário. Um exemplo desse tipo de alinhamento é o "S_S1 ↔ D1_S1, D2_S1" ou o "S_S3 ↔ D1_S2, D2_S2, D2_S4", que é do tipo 1-3, especificamente. No caso, o alinhamento do sumário possui 3 alinhamentos do tipo 1-2 ("S_S1 ↔ D1_S1, D2_S1", "S_S2 ↔ D1_S9, D2_S3" e "S_S5 ↔ D1_S4, D2_S8"), 1 alinhamento do tipo 1-3 ("S_S3 ↔ D1_S2, D2_S2, D2_S4") e 1 alinhamento do tipo 1-1 ("S_S4 ↔ D1_S3").

Outro exemplo de alinhamento pode ser visto na Figura 1. Os alinhamentos são referenciados por setas e cada sentença é representada por um círculo, preenchido, se houver um alinhamento, e vazio, se não houver.

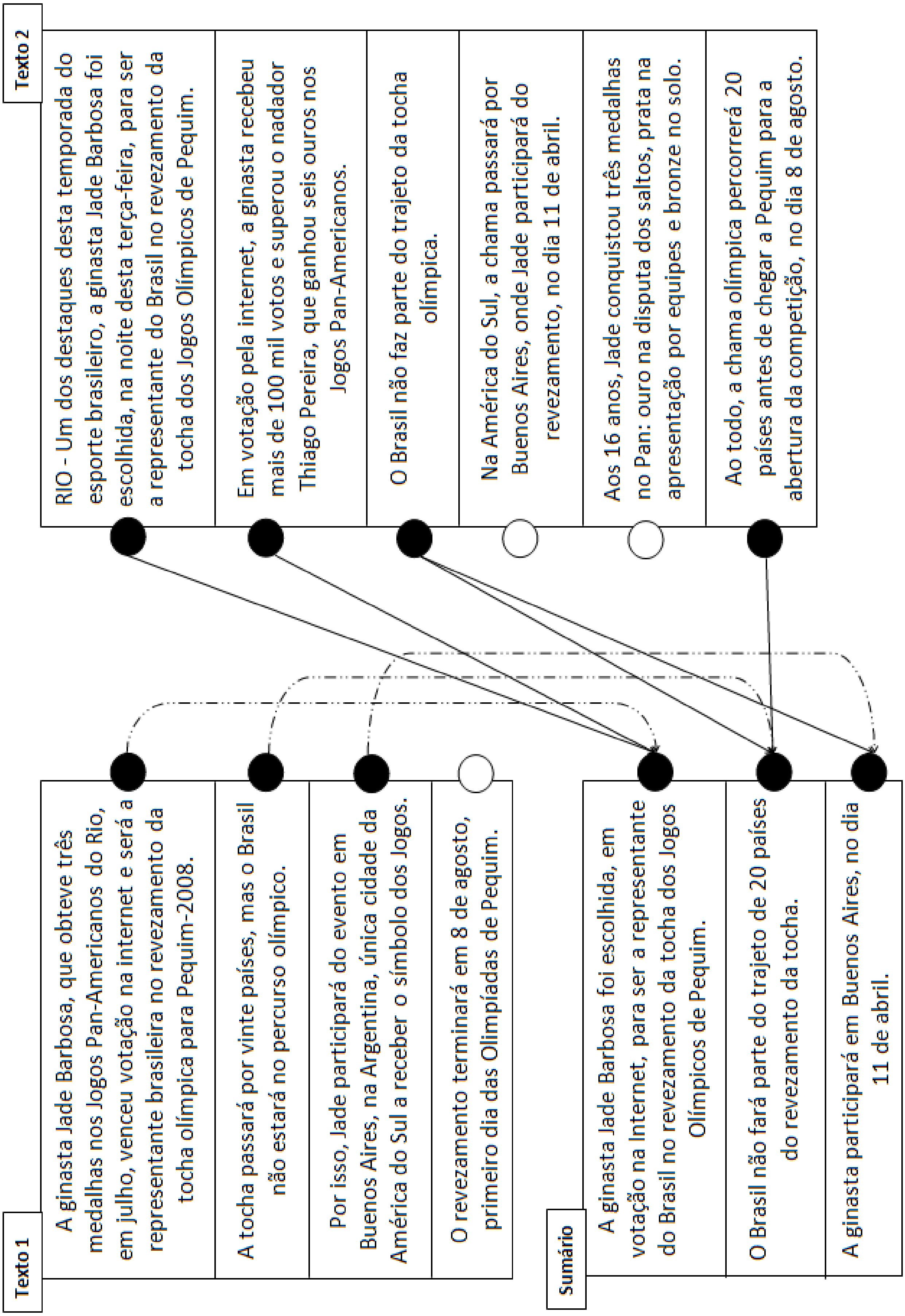


Figura 1: Exemplo de alinhamento

Como já citado, uma vez que foram encontrados esses alinhamentos, é possível obter informações de como sumarizadores humanos realizam a sumarização. Também é possível ajudar a guiar a sumarização automática multidocumento, pois com os alinhamentos pode-se obter regras de como realizar a sumarização. Um exemplo de regra, que pode ser facilmente identificada pelas sentenças grifadas dos textos (nos Quadro 3 e 4), é a qual indica que as primeiras sentenças dos textos fonte devem ser utilizadas para compor o sumário, pois as quatro primeiras sentenças dos dois documentos foram alinhadas. Além disso, utilizando aprendizado de máquina, por exemplo, é possível criar um modelo de alinhamento a partir de sentenças alinhadas. Esse modelo é utilizado para verificar, para cada novo par de sentenças, se existe ou não um alinhamento entre elas. Vê-se, portanto, que realizar tais alinhamentos ajudaria na Sumarização Automática e é nessa área de pesquisa em que se insere este trabalho. Mais especificamente, este trabalho é motivado pela falta de estudos significativos sobre a sumarização multidocumento, e também devido aos sumários multidocumento ainda apresentarem problemas relativos aos fenômenos multidocumento, como redundância, contradição, etc..

Já existem alguns trabalhos que focam no alinhamento de sumários e seus textos fonte, são exemplos os trabalhos de Banko *et al.* (1999), Marcu (1999), Jing e McKeown (1999), Daumé III e Marcu (2004, 2005) e Hirao *et al*. (2004). Neste trabalho de mestrado, foi explorado especificamente o alinhamento entre um sumário multidocumento e seus textos fonte, diferenciando-se das abordagens de Banko *et al*. (1999), Marcu (1999), Jing e McKeown (1999), e Daumé III e Marcu (2004, 2005), e aproximando-se da de Hirao *et al.* (2004). Para isso, foram exploradas tanto técnicas superficiais quanto profundas do processamento de língua natural. As técnicas profundas baseiam-se em uma teoria do nível do discurso, diferenciando-se dos trabalhos anteriores na área.

O objetivo principal deste trabalho foi, portanto, investigar e explorar técnicas de alinhamento sentencial entre sumários e seus textos fonte, tendo sido exploradas técnicas presentes nas duas abordagens de processamento de língua natural: a superficial e a profunda, como mencionado anteriormente. Como consequência, produziu-se um alinhador que realiza tais alinhamentos de forma automática. Os textos explorados são do gênero jornalístico, pois estes são escritos em uma linguagem do dia a dia e, por isso, são bastante comuns.

Uma das hipóteses que nortearam esse trabalho é a de que métodos que contenham mais conhecimento linguístico trazem melhores resultados quando comparados com métodos de alinhamento que contem menos informações linguísticas. Tinha-se também como hipótese que os alinhamentos multidocumento refletem, em certa medida, os fenômenos multidocumento. Como dito anteriormente, os fenômenos multidocumento são a redundância, contradição, diferença de estilo entre os autores, e todos os outros possíveis desafios da sumarização multidocumento em relação à sumarização monodocumento. Com o alinhamento, é possível descobrir, por exemplo, quais as informações mais redundantes, pois, em princípio, estas podem originar (serem alinhadas a) uma única sentença dos sumários.

Além disso, era esperado que houvesse maior quantidade de alinhamentos do tipo 1-N (1-2, 1-3, etc.), em que uma sentença do sumário é alinhada a mais de uma sentença dos textos fonte, do que alinhamentos dos tipos N-1, 1-0, e assim por diante.

Por fim, tinha-se como hipótese que um alinhamento em nível sentencial é suficiente para se obter bons resultados, quando comparados com um alinhamento realizado por humanos.

Para os objetivos serem atingidos, um córpus, chamado CSTNews (Cardoso *et al.*, 2011b), foi utilizado, um alinhamento manual foi feito, e alguns métodos foram desenvolvidos. Uma avaliação comparativa entre os métodos propostos e também entre um método da literatura foi realizada.

Esse trabalho é o primeiro que se tem conhecimento a realizar alinhamentos entre sumários multidocumento e seus textos fonte para a língua portuguesa do Brasil, e o primeiro a utilizar uma teoria discursiva para realizar alinhamentos. Até então, tem-se conhecimento de trabalhos para a língua inglesa (por exemplo, Marcu (1999), Jing e McKeown (1999)) e para a língua japonesa (Hirao *et al*. (2004)).

De fato, foi comprovado que os alinhamentos refletem alguns fenômenos multidocumento, graças a um bom resultado nas técnicas que se baseiam em uma teoria discursiva. Também foi comprovado que existem mais alinhamentos do tipo 1-N, e que um alinhamento em nível sentencial é suficiente para se obter bons resultados.

Esta dissertação está organizada da seguinte maneira: no Capítulo 2, é apresentada a revisão literária sobre trabalhos relacionados ao tema deste; no Capítulo 3, são apresentados os recursos utilizados para o desenvolvimento deste trabalho e também é apresentado o alinhamento manual que foi realizado; no Capítulo 4, são apresentados os métodos que foram desenvolvidos; no Capítulo 5, são apresentadas as avaliações dos métodos e de suas preposições; e, no Capítulo 6, algumas considerações finais são feitas.

Capítulo 2. Revisão Bibliográfica

2.1. Sumarização Automática

Como foi dito anteriormente, a sumarização automática consiste em produzir de forma automática um documento reduzido a partir de um ou mais textos fonte (também chamados aqui de **documentos**, ou **documentos de origem**). Quando um sumário é produzido a partir de apenas um texto, a sumarização automática é **monodocumento**, e, quando o sumário é produzido de mais de um texto, a sumarização automática é **multidocumento**.

Existem vários tipos de sumários. Uma divisão feita entre eles, quanto à forma, é a de **extratos** e ***abstracts***. Extrato é um sumário que consiste inteiramente de material copiado do texto fonte, por exemplo, sentenças inteiras, e *abstract* é um sumário que contém algum material não presente no texto fonte, em geral, produzido por operações de reescrita, como define Mani (2001).

Outra classificação de sumários é feita por Hutchins (1987), que classifica sumários em indicativos, informativos e avaliativos. Sumários indicativos não mostram detalhes, apenas os tópicos essenciais de um texto; sumários informativos são considerados substitutos do texto original; e sumários avaliativos servem como crítica ao texto fonte que o originou. Uma síntese dos tipos de sumário pode ser vista no Quadro 7.

Tipo de sumário	Breve explicação
Extrato	Sumário feito apenas com material copiado do texto fonte
Abstract	Sumário que contém algum material não presente no texto fonte
Indicativo	Sumário que mostra apenas os tópicos essenciais de um texto
Informativo	Sumário que pode substituir o texto fonte
Avaliativo	Sumário que serve como crítica ao texto fonte

Quadro 7: Tipos de sumário

Alguns exemplos de sumários podem ser vistos nos Quadro 8 a 11. No Quadro 8, é possível ver um exemplo de um *abstract* informativo, pois o mesmo contém todas as informações relevantes dos textos que o originaram e foi criado com operações de reescrita; no Quadro 9, é possível ver um sumário formado por sentenças dos textos fonte, caracterizando um extrato, e também um sumário informativo, pois contém todas as informações relevantes dos documentos de origem; no Quadro 10, é possível ver um exemplo de sumário indicativo, em que apenas é apontado o conteúdo do texto que o originou; e, no Quadro 11, é possível ver um exemplo de sumário avaliativo, em que o autor registra sua opinião sobre o discurso dos Estados Unidos. Neste trabalho de mestrado, serão utilizados sumários do tipo *abstract* para realizar o alinhamento.

> A Polícia Federal e o Ministério Público realizam nesta terça-feira operações simultâneas de busca e apreensão nos departamentos de controle de tráfego aéreo de Cumbica, em Guarulhos, de Congonhas, na capital paulista, e no Centro Integrado de Defesa Aérea e Controle de Tráfego Aéreo (Cindacta I), em Brasília.
>
> O objetivo das buscas é garantir a apreensão dos registros de ocorrências que contêm informações sobre as falhas no controle de tráfego aéreo. Esse trabalho permitirá avaliar os riscos aos quais estão expostos os passageiros e tripulantes de aeronaves e tomar medidas necessárias para aumentar a segurança no setor aéreo.

Quadro 8: Exemplo de *abstract* informativo (CSTNews (Cardoso *et al.*, 2011b)

> O Itaú obteve nos primeiros seis meses deste ano o maio lucro já registrado por um banco privado do país nos últimos vinte anos.
>
> O lucro líquido acumulado de janeiro a junho chegou a R$ 4,016 bilhões, 35,7% acima dos R$ 2,958 bilhões dos primeiros seis meses de 2006 e também superior aos R$ 4,007 bilhões anunciados na véspera pelo Bradesco, líder no ranking de bancos do país.
>
> Segundo cálculos da consultoria Economática, o resultado só perde para os R$ 4,032 bilhões (valores atualizados pelo IPCA) registrados pelo Banco do Brasil no primeiro semestre do ano passado.

Esse resultado inclui entre outros efeitos não recorrentes as vendas da participação acionária do banco na empresa de informações de crédito Serasa e da sede do BankBoston em São Paulo e constituição de provisão para créditos de liquidação duvidosa excedente ao mínimo requerido de forma a permitir a absorção de eventuais aumentos de inadimplência ocasionados por forte reversão do ciclo econômico em situações de stress.

Quadro 9: Exemplo de extrato (CSTNews (Cardoso *et al.*, 2011b)

Types of female power in Jane Austen's Pride and Prejudice are discussed. Mrs. Bennet and Charlotte Lucas represent the lack of power possessed by married women of the middle class. Lady Catherine and Caroline Bingley demonstrate the power of wealthy, single women to occasionally flaunt rules of etiquette. Lydia Bennet represents the risks of female power when bestowed upon too immature a woman, but Elizabeth and Jane Bennet characterize the positive personal and social effects of women who recognize their own power over self.

Quadro 10: Exemplo de sumário indicativo[1]

The Gettsyburg address, though short, is one of the greatest American speeches. Its ending words are especially powerful — "that government of the people, by the people, for the people, shall not perish from the earth."

Quadro 11: Exemplo de sumário avaliativo (Mani, 2001)

Assim como acontece com a sumarização humana profissional, o processo de sumarização automática pode ser dividido em fases, sendo elas: análise, transformação e síntese (Sparck Jones, 1999; Mani, 2001), como pode ser visto na Figura 2.

[1] Retirado de http://www.indiana.edu/~wts/pamphlets/abstracts.shtml

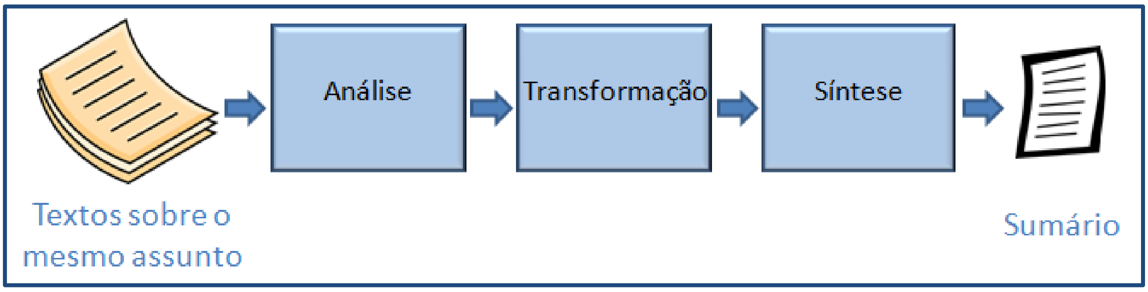


Figura 2: Fases da sumarização automática

A fase de análise consiste em analisar o(s) texto(s) fonte(s), construindo uma representação interna dele(s). A fase de transformação consiste em transformar a representação interna do texto fonte em uma representação de seu sumário, ocorrendo nessa etapa a escolha da informação importante do(s) texto(s) fonte que deverá compor o sumário multidocumento. A fase de síntese, por sua vez, consiste em transformar a representação do sumário novamente para a forma de língua natural. Este trabalho de mestrado, por buscar alinhamentos entre sumários e seus textos fonte, poderia acontecer após um sumário multidocumento já ter sido criado, e os alinhamentos poderiam trazer informações que ajudariam posteriormente, em outro processo de sumarização, a fase de transformação. Porém, é importante ressaltar que não é o foco deste trabalho realizar a sumarização, e sim auxiliá-la construindo recursos com o alinhamento sentencial (como o alinhador automático, um dos produtos deste trabalho de mestrado).

Como dito anteriormente, a sumarização automática pode ser dividida em duas abordagens: a superficial e a profunda (Mani, 2001) e existe ainda a abordagem híbrida, que utiliza técnicas das duas outras abordagens.

No processamento de língua natural, existem diferentes níveis de análise linguística: o nível fonético-fonológico, o nível morfológico, o nível sintático, o nível semântico e o nível pragmático-discursivo. Quando é dito que uma abordagem de sumarização utiliza menos conhecimento linguístico, é porque ela não se aventura além de uma representação em nível sintático e, quando uma técnica utiliza mais conhecimento linguístico, assume-se que possua no mínimo uma representação sentencial em nível semântico (Mani, 2001). Mesmo assim, nem sempre é trivial realizar essa separação.

No nível discursivo, existem algumas teorias, ou modelos, que analisam as relações que segmentos textuais podem possuir entre si. No cenário monodocumento, tem-se como exemplo a *Rhetorical Structure Theory* (RST) (Mann & Thompson, 1987), e no cenário multidocumento, tem-se como exemplo o trabalho de Radev (2000), que propôs a *Cross-Document Structure Theory* (CST). Teorias como essas podem ser utilizadas para descobrir quais segmentos textuais são mais relevantes, como é o caso da RST, na sumarização automática, por exemplo. São exemplos que utilizam RST, para a língua portuguesa, os trabalhos de Pardo e Rino (2002), Seno e Rino (2005), Uzêda *et al*. (2010) e Cardoso *et al*. (2011a). A seguir, a CST será apresentada com mais detalhes, pois a mesma foi utilizada em uma das abordagens deste trabalho.

2.2. CST (*Cross-Document Structure Theory*)

A CST surgiu a partir de trabalhos como Trigg (1983) e Trigg e Weiser (1986), Mann e Thompson (1987) e Radev e McKeown (1998). Essa teoria marca diversas relações possíveis entre partes (palavras, sentenças, blocos de texto, etc.) de mais de um documento, como exemplificado na Figura 3. As relações CST podem ser várias e irão demonstrar se, por exemplo, duas sentenças possuem informações contraditórias entre si, se possuem informações redundantes, etc.. As 24 relações CST originais podem ser vistas no Quadro 12 e as descrições de cada relação no Quadro 13.

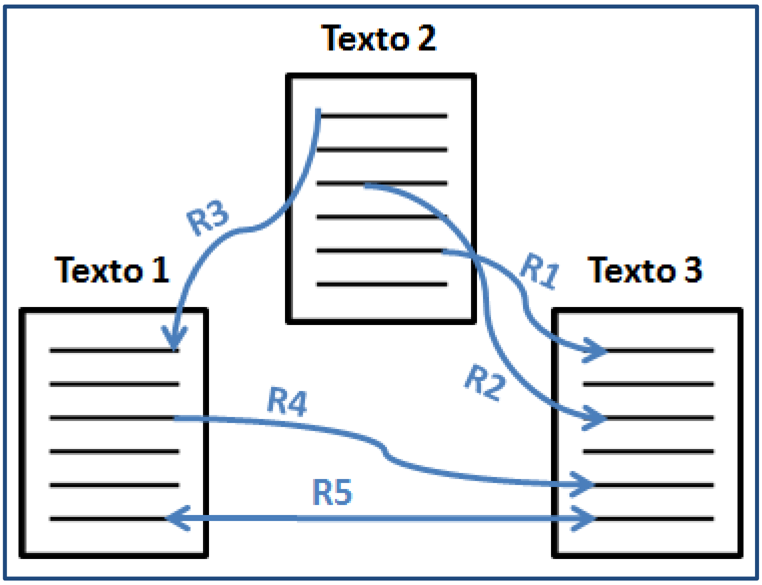


Figura 3: Relações entre mais de um texto

Identity	*Modality*	*Judgement*
Equivalence	*Attribution*	*Fulfilment*
Translation	*Summary*	*Description*
Subsumption	*Follow-up*	*Reader profile*
Contradiction	*Elaboration*	*Contrast*
Historical background	*Indirect speech*	*Parallel*
Cross-reference	*Refinement*	*Generalization*
Citation	*Agreement*	*Change of perspective*

Quadro 12: Relações CST originais

Nome da Relação	**Descrição da Relação**
Identity	Os segmentos textuais são idênticos
Equivalence	Os segmentos textuais possuem a mesma informação, porém expressa com palavras diferentes
Translation	Os segmentos textuais possuem a mesma informação em línguas diferentes
Subsumption	Um segmento textual contém mais informação do que outro segmento textual
Contradiction	Os segmentos textuais contêm informação conflitante
Historical background	Um segmento textual contém informação de conteúdo histórico em relação a outro segmento textual
Cross-reference	A mesma entidade é mencionada nos dois segmentos textuais
Citation	Um segmento textual cita outro documento
Modality	Um segmento textual contém uma informação escrita de uma forma diferente em relação à outra informação de outro segmento textual
Attribution	Um segmento textual possui uma informação e outro segmento textual possui a mesma informação atribuída a algo ou alguém
Summary	Um segmento textual sumariza outro
Follow-up	Um segmento contém uma informação complementar que reflete fatos que aconteceram desde o relato anterior
Elaboration	Um segmento contém uma informação complementar que não foi incluída em outro segmento textual
Indirect speech	Um segmento possui uma mudança de discurso direto para discurso indireto, ou vice-versa
Refinement	Um segmento possui informação complementar mais específica que outra incluída previamente
Agreement	Um segmento textual expressa concordância com outro

Judgment	Um segmento textual especifica uma informação relatada em outro segmento textual
Fulfilment	Um segmento textual possui afirma a ocorrência de um evento previsto em outro segmento textual
Description	Um segmento textual possui uma descrição de uma entidade mencionada em outro segmento textual
Reader profile	Dois segmentos textuais contêm informações semelhantes escritas para um público diferente.
Contrast	Um relato ou fato de um segmento textual é contrastado com um relato ou fato de outro segmento textual
Parallel	Um fato ou relato de um segmento textual é comparado com outro fato ou relato de outro segmento textual
Generalization	Um segmento textual é uma generalização de outro segmento textual
Change of perspective	Um segmento textual apresenta o mesmo fato de forma diferente (em outra perspectiva) que outro segmento textual

Quadro 13: Descrição das relações CST originais

É dito por Radev (2000) que a teoria CST pode ser usada como base para a sumarização automática multidocumento, pois pode fazer o sumário ser guiado por preferências do usuário, como o tamanho do sumário, a fonte das informações, a concordância entre as fontes e a ordem cronológica dos fatos. Alguns exemplos de trabalhos que fizeram uso das relações CST para realizar a sumarização automática, para a língua portuguesa, são os de Castro Jorge e Pardo (2010), de Castro Jorge *et al*. (2011) e de Cardoso *et al*. (2011a).

Na língua portuguesa do Brasil, utilizando essa teoria, o córpus CSTNews (Cardoso et al., 2011b) foi anotado. Esse córpus será apresentado com detalhes na Seção 3.1, pois foi utilizado neste trabalho de mestrado. Outros exemplos de recursos que fazem uso da CST são o córpus CSTBank (Radev *et al*., 2004), para a língua inglesa, e o parser CSTParser (Maziero e Pardo, 2011), para a língua portuguesa.

No trabalho de Zhang *et al*. (2003), as relações CST foram refinadas em seu experimento para a língua inglesa, restando 18 das relações originais. Com a anotação do córpus CSTNews e com uma nova versão do mesmo sendo feita no trabalho de Maziero *et al*. (2010), as relações CST foram também refinadas, por meio da remoção de algumas relações que nunca foram observadas no córpus CSTNews, ou ainda que não eram esperadas de ocorrer nos textos. No refinamento, algumas relações, por serem bastante similares, foram unidas. As 14 relações restantes podem ser vistas no Quadro 14.

Identity	*Modality*
Equivalence	*Attribuition*
Translation	*Summary*
Subsumption	*Follow-up*
Contradiction	*Elaboration*
Historical background	*Indirect speech*
Citation	*Overlap*

Quadro 14: Relações CST refinadas (Maziero *et al*., 2010)

Um exemplo da relação *Historical background* retirada do córpus CSTNews pode ser visto no Quadro 15. Neste exemplo, a sentença (1) contém uma informação de conhecimento histórico em relação à sentença (2).

(1) De quebra, esta conquista iguala o número de medalhas de ouro faturadas em Santo Domingo (2003), quando o Brasil também somou 29.
(2) É a 29ª medalha para o Brasil no Pan.

Quadro 15: Exemplo de relação *Historical Background*

Um exemplo da relação *Subsumption* retirada do córpus CSTNews pode ser visto no Quadro 16. Neste exemplo, a sentença (1) subsume (engloba) a sentença (2), pois a (1) possui mais informação que a (2).

(1) A medalha de prata ficou com a americana April Steiner, com a marca de 4m40 e o bronze foi para a cubana Yarisley Silva, com 4m30.
(2) Já o bronze pertence à cubana Yarisley Silva, com a marca de 4,30m.

Quadro 16: Exemplo de relação *Subsumption*

Um exemplo da relação *Contradiction* retirada do córpus CSTNews pode ser visto no Quadro 17. Neste exemplo, as duas sentenças contêm a informação conflitante de quanto era o antigo recorde pan-americano.

(1) Depois da queda de April Steiner, a brasileira Fabiana Murer leva a medalha de ouro no salto com vara, com <u>4m50</u> - novo recorde pan-americano.
(2) Com a marca de 4m60, Fabiana não só venceu a prova, como também estabeleceu o novo recorde pan-americano, 20cm mais alto do que a antiga marca de <u>4m40</u>.

Quadro 17: Exemplo de relação *Contradiction*

Um exemplo da relação *Elaboration* pode ser visto no Quadro 18. Nesse exemplo, a sentença (1) elabora a sentença (2), revelando uma informação complementar.

(1) A brasileira Joana Costa ficou com a quinta posição, com 4m20 e mostrou, mais uma vez, neste Pan do Rio, que a pressão de competir em casa pode prejudicar os atletas.
(2) Já a outra brasileira que participou da prova, Joana Costa, não subiu ao pódio, uma vez que não alcançou a marca da cubana.

Quadro 18: Exemplo de relação *Elaboration*

Ainda no trabalho de Maziero *et al*. (2010), foi definida uma tipologia das relações CST, que pode ser vista na Figura 4. Nessa tipologia, as relações refinadas são divididas entre relações do tipo “conteúdo” e relações do tipo “apresentação e forma”. O objetivo principal das relações do tipo “conteúdo” é relacionar o conteúdo dos segmentos textuais, e as relações do tipo “apresentação e forma” consideram a forma que o conteúdo foi expresso. Entre um par de segmentos textuais, apenas uma relação do tipo “conteúdo” pode ocorrer. Porém, relações do tipo “apresentação e forma”, por sua vez, eventualmente acontecem com relações do tipo “conteúdo”.

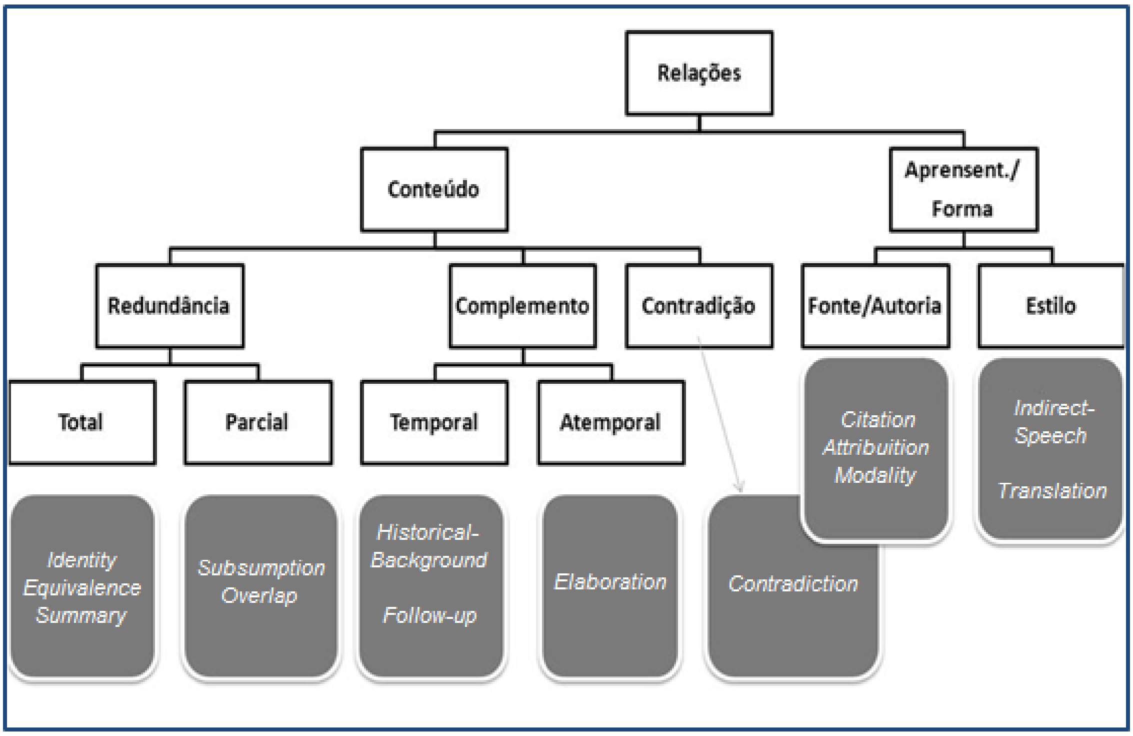


Figura 4: Tipologia das relações CST (Maziero *et al.*, 2010)

A teoria CST pode, inclusive, ajudar a guiar o alinhamento, como será explicado com detalhes na Seção 4.2.

Na seção seguinte, a revisão literária de alinhamento é apresentada, com exemplos na área de sumarização automática, tradução automática, entre outros. Uma vez que alinhamentos são obtidos, é possível aprender como sumarizadores humanos realizam a sumarização. É possível também, a partir dos alinhamentos, a obtenção de regras e modelos[2] que, quando explicitados, podem subsidiar métodos automatizados para a sumarização automática.

2.3. Alinhamento

A técnica de alinhamento de textos surgiu na área de tradução (Brown *et al.*, 1990), em que são alinhados segmentos textuais entre um texto e sua versão traduzida para outra língua. Com esses alinhamentos, é possível obter regras e modelos para auxiliar na tradução automática.

[2] Modelo pode ser entendido como um conjunto de regras para se realizar algum propósito, como realizar a sumarização automática.

Os alinhamentos podem ser feitos de diversas formas, utilizando mais ou menos conhecimento linguístico. Um exemplo que utiliza pouco conhecimento linguístico é o que realiza o alinhamento entre sentenças com o critério do tamanho das mesmas (por exemplo, Gale e Church, 1993), e um exemplo que utilize mais conhecimento linguístico é o que realiza o alinhamento com base na quantidade de substantivos, verbos, etc. que as sentenças possuam (por exemplo, Piperidis *et al.*, 2000).

Primeiro, na Seção 2.3.1, são comentados brevemente exemplos de alinhamentos presentes em áreas correlatas e, na Seção 2.3.2, são descritos detalhadamente trabalhos de alinhamentos relacionados à sumarização automática, foco deste trabalho.

2.3.1. História do Alinhamento

Como citado anteriormente, muitas aplicações do processamento de língua natural usam o alinhamento, seja de palavras, n-gramas, sentenças, entre outros. O alinhamento surgiu na tradução automática, em que são alinhados textos em uma língua e sua versão em outra língua.

Quando uma unidade textual é alinhada a apenas uma unidade textual, o alinhamento é dito ser do tipo 1-1. É possível que ocorra o alinhamento entre uma unidade textual e mais de uma unidade textual, caracterizando um alinhamento 1-N. Na tradução automática, não é incomum que uma sentença da língua fonte origine mais de uma sentença da língua alvo. Isso também pode acontecer com alinhamento de outras granularidades, como o nível de palavras, em que uma palavra é traduzida dando origem a mais de uma palavra na língua alvo. É também possível que uma palavra não possua uma tradução direta no texto fonte, resultando, assim, em um alinhamento 1-0, ou que mais de uma palavra dê origem a apenas uma palavra, caracterizando um alinhamento N-1. Alguns exemplos de alinhamentos sentenciais podem ser vistos no Quadro 19. Na primeira linha, duas sentenças em inglês são alinhadas a duas sentenças em francês; na segunda e na terceira linha, uma sentença é alinhada a uma sentença; e, na quarta linha, duas sentenças em inglês são alinhadas a uma sentença em francês.

English	French
According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above-average growth rates.	Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment.
The higher turnover was largely due to an increase in the sales volume.	La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes.
Employment and investment levels also climbed.	L'emploi et les investissements ont egalement augmenté.
Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.	La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.

Quadro 19: Exemplos de alinhamento (Gale e Church, 1993, p. 77)

Para realizar o alinhamento, menos informações linguísticas podem ser utilizadas, como acontece em alinhamentos que utilizam técnicas empíricas, como alinhar de acordo com o tamanho de uma sentença (por exemplo, Gale e Church, 1993). Esse tipo de alinhamento baseia-se na suposição de que sentenças pequenas teriam sua correspondente em um texto traduzido com um tamanho também pequeno e que sentenças grandes, por sua vez, teriam traduções também grandes.

Outros exemplos de alinhamento que utilizam técnicas empíricas são os que utilizam técnicas de reconhecimento de padrões para realizar o alinhamento (por exemplo, Melamed, 2000). Nesse caso, palavras são alinhadas se forem similares, o que é verificado por uma medida baseada em cognatos, a LCRS (*Longest Common Subsequence Ratio*). Essa medida calcula a divisão entre o tamanho da maior sequência de caracteres comum e o tamanho da maior palavra. Por exemplo, entre as palavras "*automatic*" e "automático", a LCRS é de 7/8. Um número limite é definido para o valor dessa medida, da forma que palavras com LCRS acima do valor são alinhadas.

Na tradução estatística, a tarefa de tradução é modelada como uma função de probabilidades extraídas de exemplos de tradução, o córpus paralelo (Brown *et al.*, 1993). Basicamente, a partir de um córpus alinhado no nível de sentenças ou de palavras, por exemplo, é possível obter um modelo em que é necessário descobrir a probabilidade de uma sentença em uma língua fonte gerar a outra sentença em uma língua alvo, escolhendo a tradução com maior probabilidade (por exemplo, Vogel *et al.*, 1996; Och *et al.*, 1999; Yamada e Knight, 2001).

Ainda, existem os alinhamentos que utilizam mais informações linguísticas. Tais alinhamentos podem fazer uso de informações como o número de substantivos, verbos, advérbios e adjetivos presentes nas duas unidades textuais. (por exemplo, Papageorgiou *et al.*, 1994; Piperidis *et al.*, 2000). Esses alinhamentos utilizam recursos dependentes de língua, que são de difícil construção, além de requererem muito conhecimento das duas línguas envolvidas.

No trabalho de Caseli (2003), no âmbito da tradução automática, alguns métodos das categorias linguísticas e empíricas da língua inglesa foram desenvolvidos para serem avaliados na língua portuguesa do Brasil. Nesse trabalho, uma análise foi feita para comparar os resultados dos métodos, mas não foi possível eleger um método apenas como o melhor, pois, em suas avaliações, os mesmos obtiveram valores próximos.

Similar aos alinhamentos presentes na tradução automática, estão os alinhamentos presentes na simplificação textual. No trabalho de Specia (2010), por exemplo, a tarefa de simplificação textual é abordada como uma tarefa de tradução estatística. Dessa forma, um texto complexo é "traduzido" para sua versão simplificada a partir de córpus de textos originais e simplificados alinhados no nível de sentença. Um alinhamento desse tipo pode ser visto no Quadro 20, em que se pode notar que a sentença simplificada foi produzida em função da sentença original com alguma reescrita, para facilitar o entendimento.

Sentença original	**Sentença simplificada**
Cientistas britânicos detectaram, em adultos, a produção de células hepáticas a partir de células-tronco da medula óssea.	Cientistas britânicos detectaram em adultos que células-tronco da medula óssea produziram células do fígado.

Quadro 20: Exemplo de alinhamento na simplificação textual (Specia, 2010, p. 32)

Para realizar o alinhamento, um anotador humano realiza as simplificações de sentenças de textos previamente coletados e, dessa forma, o alinhamento já é registrado. Nesse tipo de tarefa, também existem os alinhamentos 1-N, pois uma sentença complexa, e muitas vezes grande, pode ser transformada em mais de uma sentença mais simples.

O alinhamento pode ser também utilizado em tarefas de perguntas e respostas, em que são alinhadas sentenças da pergunta com suas sentenças da respectiva resposta. No trabalho de Soricut e Brill (2004), uma arquitetura *noisy-channel* é utilizada para realizar a tarefa de perguntas e respostas. Basicamente, um modelo como esse envolve uma fonte emitindo uma mensagem (no caso, uma pergunta) por um canal, e esse canal, através de ruídos, transforma a mensagem em outra mensagem (no caso, uma resposta para a pergunta). O segredo é calcular a probabilidade no caminho inverso (a probabilidade da resposta dado que se tem a pergunta) de acordo com o teorema de Bayes[3]. Assim, resume-se que, para encontrar a resposta para a pergunta requerida, é necessário computar a probabilidade do modelo de língua da resposta e a probabilidade do modelo de tradução da resposta para a pergunta. O alinhamento acontece em um córpus de pares de perguntas e respostas construído para ser utilizado na arquitetura *noisy-channel*. Para realizar o alinhamento, páginas de perguntas e respostas frequentes (FAQs) são coletadas com uma busca na WEB. Após isso, uma passagem por elas é realizada baseada em alguns indicadores, como pontuação e pistas lexicais, e assim a maioria dos pares de perguntas e respostas é retornada, juntamente com ruído. Depois, os pares passam por filtros, como identificação de língua, restrições de tamanho e pistas lexicais, para reduzir o nível de ruído. Um exemplo de um alinhamento desse tipo pode ser visto no Quadro 21, em que Q é a pergunta e A é a resposta.

Q: *Can I travel with the Ameripass in Mexico?*

[3] No exemplo, a fórmula de Bayes é $P(Q|A) = P(A|Q).P(Q)/P(A)$, em que Q é uma pergunta e A uma resposta.

A: *The Ameripass includes all of USA excluding Alaska. In Cadada, it is valid for travel to Montreal, Quebec, Toronto, Ontario, Vancouver BC, and New Westminster BC. In Mexico it is valid for travel to Matamoros, Nuevo Laredo, Tamaulipas and Tijuana Baja California.*

Quadro 21: Exemplo de alinhamento em tarefas de perguntas e respostas (Soricut e Brill, 2004, p. 63)

A área de interesse neste trabalho que usa o alinhamento é a sumarização automática. Alinhamentos encontrados nessa área terão seus trabalhos apresentados na próxima seção, em mais detalhes.

2.3.2. Alinhamento na Sumarização Automática

Nesta seção, são apresentados os trabalhos de alinhamento relacionados à sumarização automática, foco deste trabalho. Ao contrário de alinhamentos presentes na tradução automática, por exemplo, é esperado que muitos dos alinhamentos presentes na sumarização automática multidocumento sejam do tipo 1-N, pois, nesse caso, um texto reduzido, que contém a informação principal dos textos que o originaram, foi gerado a partir de mais de um documento que versam sobre um mesmo assunto.

O primeiro trabalho na área de alinhamento na sumarização data de 1995. Neste trabalho (Kupiec *et al.*, 1995), é apresentada uma abordagem para a criação de extratos monodocumento. Para compor os mesmos, cada sentença dos textos fonte recebe uma probabilidade obtida de um classificador, que considera cinco características para realizar a classificação, sendo elas: (i) o tamanho da sentença, (ii) a presença de expressões fixas (como "em conclusão"), (iii) a posição do parágrafo, (iv) a presença de palavras temáticas (ou seja, relativas ao assunto do texto), e (v) a presença de palavras em letras maiúsculas. Além do classificador, foi criado um córpus de treino formado por 188 pares de documentos de domínio científico e técnico e seus sumários manuais. Os sumários eram, em sua maioria, sumários indicativos. Para realizar o treinamento desejado pelos autores, era necessário possuir os extratos dos documentos e, para isso, o alinhamento de cada sentença do sumário a sentenças do seu texto de origem foi realizado, sendo feito sempre o melhor casamento possível.

De acordo com os autores, um alinhamento entre duas sentenças pode ocorrer: (i) se houver um “casamento direto” entre elas, ou seja, forem idênticas ou possuírem poucas diferenças, que ainda faça com que as duas tenham o mesmo significado, e (ii) se for óbvio que mais de uma sentença do documento tiver sido utilizada para criar uma sentença do sumário (resultando em uma “junção” das mesmas). Ainda, as sentenças dos sumários manuais poderiam ser classificadas em: (i) “não casável” (*unmatchable*), quando a sentença fora criada pelo autor a partir de uma leitura geral (provavelmente incluindo informações inferidas por ele), ou (ii) “incompleta”, quando há um *overlap* de informação entre o par de sentenças considerado, mas o conteúdo da sentença original não foi preservado na sentença do sumário, ou quando a sentença do sumário inclui uma sentença do documento, porém possui também informação inferida pelo autor. Exemplos desses tipos, retirados do artigo dos autores, podem ser vistos nos Quadro 22, 23 e 24.

Casamento direto (*Direct match*)	
Sentença do sumário	**Sentença do documento**
This paper identifies the desirable features of an ideal multisensory gas monitor and lists the different models currently available.	*The present part lists the desirable features and the different models of portable, multisensor gas monitors currently available.*

Quadro 22: Exemplo de casamento direto (Kupiec *et al.*, 1995, p. 73)

No exemplo do Quadro 22, as duas sentenças contêm exatamente a mesma informação, apesar de possuírem algumas palavras diferentes e nem todas elas manterem a mesma ordem. Por esse motivo, recebem a classificação de “casamento direto”.

Junção direta (*Direct join*)
Sentença do sumário

In California, Caltrans has a rolling pavement management program, with continuous collection of data with the aim of identifying roads that require more monitoring and repair.
Sentenças do documento
Rather than conducting biennial surveys, Caltrans now has a rolling pavement-management program, with data collected continuously.
The idea is to pinpoint the roads that may need more or less monitoring and repair.

Quadro 23: Exemplo de junção direta (Kupiec *et al.*, 1995, p. 73)

No exemplo do Quadro 23, a sentença do sumário engloba as duas sentenças do documento, e por isso os alinhamentos são classificados como uma "junção direta".

Casamento incompleto (*Incomplete match*)
Sentença do sumário
Intergranular fracture of polycrystalline Ni_3Al was studied at 77K.
Sentença do documento
Before discussing the observed deformation and fracture behavior of polycrystalline Ni_3Al at 77K in terms of the kinetics of the proposed environmental embrittlement mechanism, we should ask whether the low temperature by itself significantly affects the brittleness of Ni_3Al.

Quadro 24: Exemplo de casamento incompleto (Kupiec *et al.*, 1995, p. 73)

Ainda, no exemplo do Quadro 24, as sentenças recebem a classificação de "casamento incompleto", pois possuem informações diferentes entre si (especialmente o segmento: "*we should ask whether the low temperature by itself significantly affects the brittleness of Ni_3Al*"), sendo estas não encontradas em nenhuma outra sentença do sumário. Se essa informação extra fosse encontrada no sumário, uma "junção direta" poderia ocorrer.

O alinhamento era, então, feito de forma automática em um primeiro passo, seguindo as classificações que os autores criaram, e depois, o alinhamento resultante era utilizado em um segundo passo, em uma anotação manual.

Para a avaliação, as sentenças escolhidas pelo sumarizador para compor os extratos eram julgadas como corretas: (i) se possuíam um “casamento direto” e estavam presentes no sumário manual, e (ii) se estavam no sumário manual como parte de uma “junção” e todas as outras sentenças da “junção” também foram incluídas no extrato (ou seja, toda a informação da “junção” fora preservada). Como resultado, os autores obtiveram que, quando o sumarizador retorna a mesma quantidade de sentenças que possuem os sumários manuais (498), ele acerta 42% (211). No trabalho de Kupiec *et al*. (1995), portanto, o alinhamento era utilizado para a criação dos extratos e esses extratos (tidos como *gold standard*, ou seja, eram considerados como corretos) eram usados na comparação com os extratos gerados por seu sumarizador.

Outros trabalhos na área surgiram no ano de 1999. Banko *et al*. (1999), por exemplo, nas duas abordagens discutas em seu trabalho, focam-se na ideia de fazer o alinhamento de sentenças usando palavras comuns a esses segmentos textuais. Os autores utilizam esse alinhamento para a geração de um córpus de aproximadamente 25000 extratos. O intuito dessa abordagem é exemplificado na Figura 5.

Tendo descoberto de onde as sentenças do sumário humano vieram do texto fonte, é possível criar um extrato automático a partir dessas sentenças que foram utilizadas para a criação do sumário humano.

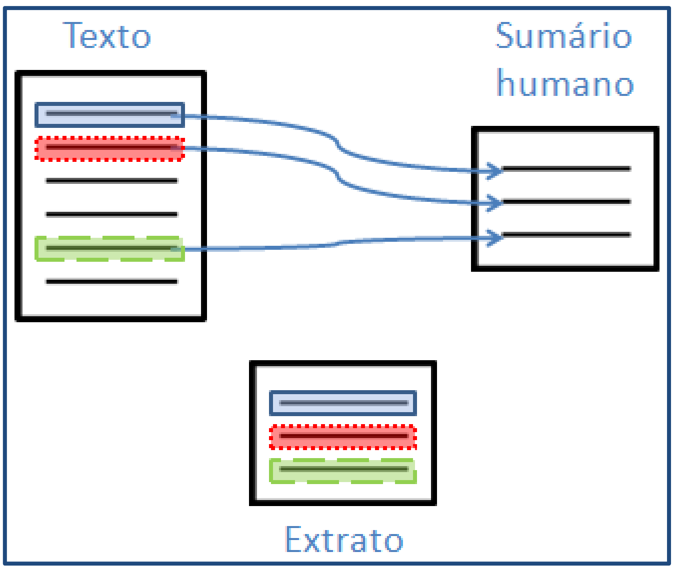


Figura 5: Geração de extratos

No trabalho de Banko *et al*. (1999), é também discutido como esse problema de alinhamento difere de trabalhos anteriores em alinhar textos paralelos. Como já comentado anteriormente, para fazer o alinhamento entre textos paralelos, duas técnicas são principalmente utilizadas: (i) fazer uso do tamanho dos segmentos textuais, alinhando, dessa forma, intervalos de tamanhos aproximados, e (ii) usar pistas lexicais. A última técnica baseia-se na observação de que sequências de caracteres são comuns entre palavras de algumas línguas, como o inglês e o português no exemplo: “*presentation*” e “apresentação”, em que os oito primeiros caracteres da primeira palavra são comuns aos caracteres de 2 a 9 na segunda palavra. Em outras línguas, como o japonês, por exemplo, em relação ao português, não existe essa correspondência de caracteres, como no exemplo “家” e “casa”, ou “薔薇” e “rosa”. Essa segunda técnica pode ser aproveitada para o alinhamento entre textos e sumários, pois as duas línguas são as mesmas e é baseado nessa segunda técnica que surgiram as abordagens de Banko *et al*. (1999).

Na primeira abordagem, é usado o algoritmo *Term Lenght Term Frequency*[4] (TLTF) para realizar o alinhamento. O TLTF é baseado na suposição que palavras mais frequentes são menores e palavras que aparecem com menos frequência nos textos, por sua vez, tendem a ser maiores. As palavras menos frequentes são as mais prováveis de distinguir segmentos textuais, como sentenças, frases ou documentos inteiros. No trabalho de Banko *et al*. (1999), o TLTF é utilizado para identificar as palavras que são mais propensas a serem comuns a um par de sentenças relacionadas.

A primeira abordagem resume-se ao algoritmo apresentado a seguir.

(1) Para cada sentença h_i pertencente a um sumário H,

(a) Computar o TLTF para todos os termos t pertencentes à sentença h_i,

(b) Fazer com que os *top* n termos representem uma *query* Q para h_i, em que $n =$ número de termos únicos em $h_i/3$

[4] O algoritmo TLTF multiplica uma função monótona do tamanho do termo e uma função monótona da frequência do termo (Kantrowitz, 2000). Uma função monótona é uma função que sempre varia no mesmo sentido.

(c) Para cada sentença d_j pertencente ao documento D, fazer com que o score de d_j seja igual à frequência relativa cujos termos em Q casam (alinham) com os temos em d_j,

(d) Selecionar sentença(s) do documento D da forma que sejam os melhores alinhamentos para a sentença h_i:

(i) Se existir um *score* de d_j que seja maior ou igual a 0,75, escolher $max_j(d_j)$.

(ii) Caso contrário, escolher não mais que duas *top* sentenças d_j com maiores *scores* em que o *score* de d_j é maior ou igual a 0,5.

(iii) Caso contrário, escolher não mais que duas *top* sentenças d_j com maiores *scores* em que o *score* de d_j é maior ou igual a α e α é um limitante escolhido empiricamente.

Qualquer sentença d_j pode ser escolhida por no máximo 1 sentença h_i.

De acordo com o algoritmo anterior, nos passos (a) e (b), o TLTF das palavras das sentenças do sumário é calculado para que sejam consideradas apenas as palavras que interessam (não as mais frequentes e nem as muito específicas). Com as n *top* palavras[5], é representada uma *query*. Depois, no passo (c), é atribuída uma pontuação para cada sentença do documento da forma que, quanto mais termos a sentença do documento tiver que casem (alinhem) com a *query* da sentença do sumário, maior a pontuação dela será. Por fim, no passo (d), escolhem-se as sentenças com maior pontuação, garantindo-se que o extrato criado a partir da união dessas sentenças seja o mais próximo possível do sumário humano.

Com o algoritmo anterior, aproximadamente 25000 extratos foram produzidos a partir de um conjunto de notícias da *Reuters*[6] e do *Los Angeles Times*[7]. As notícias são de um período de seis meses dos anos 1997 e 1998 e abordavam vários tópicos, como política, esportes, saúde e negócios, entre outros. Para a avaliação, centenas de

[5] Os n top termos são os n termos únicos na sentença do sumário divido por três. Foi dividido por esse valor baseado na observação que as palavras chaves mais importantes do texto representam 1/4 ou 1/3 dos termos únicos da sentença.

[6] http://www.reuters.com/

[7] http://www.latimes.com/

notícias e seus sumários foram selecionadas para estudar possíveis problemas da abordagem criada. Um problema chave que foi encontrado foi que o algoritmo falha em alinhar duas sentenças que contenham nomes próprios com tamanhos pequenos. Isso acontece por causa do algoritmo TLTF, pois, da forma que foi utilizado na primeira abordagem, ele tende a selecionar palavras de tamanho maior. Para contornar o problema, uma segunda abordagem foi proposta.

A segunda abordagem proposta por Banko *et al*. (1999) estende a anterior marcando as classes das palavras[8] (*part of speech*) nos documentos para que nomes não sejam perdidos. A segunda abordagem segue o algoritmo a seguir.

> Para um documento D e seu correspondente sumário manual H, computar o TLTF como descrito anteriormente, juntamente com as funções morfossintáticas das palavras como se segue:
>
> (1) Usando um *part-of-speech tagger,* marcar substantivos próprios e substantivos comuns que ocorrem em cada sentença h_i pertencente a H e cada sentença d_j pertencente a D.
>
> (2) Para cada sentença h_i pertencente a H:
>
> (a) Medir a sobreposição (*overlap*) entre cada d_j pertencente a D e h_i.
>
> Atribuir pontos para o *overlap* de substantivo da seguinte maneira:
>
> (i) Se casar (alinhar) com um substantivo próprio, incrementar $overlap(h_i, d_j)$ em dois pontos.
>
> (ii) Se casar (alinhar) com um substantivo comum, incrementar $overlap(h_i, d_j)$ em um ponto.
>
> (b) Selecionar $argmax\left(overlap(h_i, d_j)\right)$ como o melhor alinhamento para h_i, onde $overlap(h_i, d_j)$ for maior ou igual a α, em que α é um limitante empiricamente escolhido.
>
> Qualquer sentença d_j pode ser escolhida por no máximo 1 sentença h_i.

[8] Função morfossintática é a classe que a palavra possui em uma sentença, como substantivo, adjetivo, verbo, ou ainda mais específicas, como substantivo próprio, substantivo comum, etc..

Nesse segundo algoritmo, após calcular o TLTF novamente, no passo (1), são marcados os substantivos próprios e substantivos comuns em todas as sentenças fonte e do sumário. No passo (2a), são medidas as sobreposições[9] entre todas as sentenças fonte e do sumário, e são atribuídos pontos para essas sobreposições para que sejam favorecidas as sentenças que casam substantivos próprios e substantivos comuns. Por fim, no passo (2b), é escolhida a sentença que maximizar essa sobreposição.

Com uma inspeção visual de um grande conjunto aleatoriamente amostrado, os autores relataram não ter observado grandes problemas, e os problemas que ainda restaram ou estão fora do escopo da abordagem, como o fato de alguns sumários serem mais uma análise crítica do que uma sumarização, ou são problemas no pré-processamento, como determinar o limite de sentenças, entre outros.

Uma desvantagem para esse método é o fato dos autores não terem realizado uma avaliação mais objetiva. Dessa forma, fica difícil mensurar e avaliar se os resultados foram bons ou ruins.

Outro trabalho na área foi feito por Marcu (1999). O algoritmo desenvolvido pelo autor tem basicamente a mesma finalidade das duas abordagens de Banko *et al*. (1999): a construção de um grande córpus de forma automática para a pesquisa de sumarização. Tendo como entrada um texto e seu sumário humano, o algoritmo retorna seu extrato correspondente, ou seja, o conjunto de orações (*clauses*) e sentenças do texto que foram utilizadas para escrever o sumário. A ideia chave de sua abordagem é determinar qual extrato tem a maior similaridade com o sumário correspondente e, para isso, em vez de ser respondida a pergunta "Deve ser inserida essa oração no sumário?" é respondida uma questão complementar "Se for removida esta oração do texto, ainda é possível escrever o sumário?".

Utilizando uma abordagem gulosa[10], orações (*clauses*) do texto são repetidamente excluídas até que o extrato resultante seja o mais similar possível, verificado com uma medida de similaridade, com o sumário humano. A medida de similaridade utilizada e o algoritmo completo podem ser vistos a seguir.

[9] Por sobreposição (*overlap*) entende-se encontrar o que é comum às duas sentenças.
[10] Um algoritmo guloso é aquele que sempre faz a melhor escolha local enquanto procura por uma resposta (Black, 2005).

Eventualmente, o extrato convergirá para um estado em que a próxima oração excluída faria diminuir a similaridade com o sumário.

$$sim(E_M, A) = \frac{\sum_{t \,\in\, E_M \,\cup\, A} w(t)_{E_M}\, w(t)_A}{\sqrt{\sum_{t \,\in\, E_M} w(t)^2{}_{E_M} \sum_{t \,\in\, A} w(t)^2{}_A}}$$

Na medida de similaridade $sim(E_M, A)$, E_M é o extrato cuja similaridade com o *abstract* correspondente é máxima, A é o *abstract* correspondente, t é um *token* e $w(t)$ é o seu peso, e $w(t)_A$ e $w(t)_{E_M}$ representam os pesos do *token* t no *abstract* A e no extrato E_M, respectivamente. O peso dos *tokens* é dado por suas frequências no extrato e no *abstract* respectivamente.

Entrada: Uma tupla $[Abstract, Texto]$

Saída: Uma tupla $[Abstract, Extrato, Texto]$

(1) Segmentar o $Abstract$ e o $Texto$ em orações;

(2) Realizar lematização e excluir as *stopwords*[11] em ambos os conjuntos de orações;

(3) E_M = ClausesOf($Texto$);

(4) **Enquanto**

$\left(E = E_M \setminus C_i \;\middle|\; C_i \in E_M \wedge (\forall C_j \in E_M)\left(i \neq j \rightarrow sim(E_M \setminus C_i, Abstract) \geq\right.\right.$ $simEM \mid Cj, Abstract \wedge E>EM$

(5) $E_M = E$;

(6) **Fim do enquanto**

(7) Eliminar de E_M as orações que possuam categoria retórica de satélite fraco;[12]

[11] *Stopwords* são palavras muito frequentes que não carregam muito significado, como artigos e preposições.

(8) Eliminar de E_M as orações pequenas que possuam categoria retórica de satélite forte;

(9) Eliminar de E_M os subtítulos;

(10) Eliminar de E_M as orações que não sejam similares a nenhuma outra

(11) oração no $Abstract$;

(12) Adicionar a E_M as orações no $Texto$ que são mais similares com cada oração no abstract;

(13) Adicionar a E_M as únicas orações do $Texto$ que contenham pelo menos duas palavras do $Abstract$ que não são usadas em nenhuma outra oração;

(14) Eliminar de E_M todas as orações redundantes pequenas;

Retornar $Extrato = E_M$;

O passo mais importante do algoritmo anterior compreende as linhas 3 a 6, em que o extrato é construído seguindo a ideia de excluir orações do documento até que a similaridade com o sumário deixe de aumentar. Nas linhas anteriores (1 a 2), ocorre a preparação dos textos e, nas linhas seguintes (7 a 14), o extrato criado nas linhas 3 a 6 é "limpo".

Para avaliar o algoritmo de extração desenvolvido por Marcu (1999), o autor fez o experimento a seguir. De forma aleatória, foram selecionados 10 textos acompanhados de seus sumários do córpus Ziff-Davis[13], uma coleção de artigos de jornal na língua inglesa anunciando produtos relacionados a computadores. Cada texto e cada sumário foram divididos em orações e cada oração foi numerada. Os textos e sumários divididos foram apresentados a 14 juízes para que fossem determinadas quais eram as unidades do texto cuja semântica estava refletida por unidades no sumário, ou seja, quais unidades textuais foram utilizadas para a criação do sumário. 11 juízes analisaram os 10 textos e sumários, enquanto 3 juízes analisaram apenas parte do total de textos e sumários, resultando no total em 125 julgamentos independentes.

[12] São os segmentos (orações) que não são a informação mais importante da sentença (núcleo), mas sim um satélite.

[13] http://www.ziffdavis.com/

As médias dos resultados obtidos, aplicando-se o algoritmo de extração no nível de oração, juntamente com a avaliação humana, podem ser vistas na **Erro! Fonte de referência não encontrada.**. Também foi aplicado o algoritmo em nível sentencial. As médias dos resultados obtidos juntamente com a avaliação humana podem ser vistas na **Erro! Fonte de referência não encontrada.**.

Tabela 1: Média dos resultados em nível de oração (Marcu, 1999)

	Precisão	**Cobertura**	**Medida-F**
Juízes humanos	80,94%	88,01%	83,34%
Algoritmo	74,27%	80,29%	76,47%

Tabela 2: Média dos resultados em nível de sentença (Marcu, 1999)

	Precisão	**Cobertura**	**Medida-F**
Juízes humanos	84,42%	88,73%	85,71%
Algoritmo	77,45%	80,06%	78,15%

A medida "precisão" é utilizada para verificar a quantidade de orações ou sentenças do texto similares aos do sumário que foram retornadas dentre todas as retornadas (em outras palavras: as instâncias identificadas de forma correta dentre todas as retornadas); a medida "cobertura" é utilizada para verificar a quantidade de orações ou sentenças retornadas (em outras palavras: as instâncias identificadas de forma correta dentre todas as que deveriam ter sido retornadas); e a medida-F combina as duas anteriores em uma única medida. As fórmulas das medidas podem ser vistas a seguir.

$$precisão = \frac{resultados\ corretos}{resultados\ corretos + resultados\ inesperados}$$

$$cobertura = \frac{resultados\ corretos}{resultados\ corretos + resultados\ corretos\ ausentes}$$

$$medidaF = 2\frac{precisão\ .\ cobertura}{precisão + cobertura}$$

É interessante ressaltar que os resultados para o nível de oração e para o nível de sentença ficaram próximos. Isso provavelmente acontece por se tratarem de sentenças pequenas, sendo que os textos escolhidos foram textos de notícias. Em comparação com a abordagem de Banko *et al*. (1999), esse trabalho já possui uma avaliação mais objetiva, que utiliza medidas conhecidas da literatura. Isso possibilita que comparações possam ser feitas mais claramente. Inclusive, vários dos trabalhos que ainda serão citados nesta monografia refazem o experimento de Marcu (1999) para comparações com seus métodos. Neste trabalho de mestrado, as mesmas medidas (precisão, cobertura e medida-F) são utilizadas, a fim de realizar uma avaliação comparativa. Porém, os textos utilizados foram escritos na língua portuguesa do Brasil.

As abordagens dos trabalhos de Banko *et al*. (1999) e Marcu (1999) podem ser consideradas modelos de *bag of words*. Em modelos de *bag of words*, um texto é representado como uma coleção não ordenada de palavras, sendo a gramática e a ordem das palavras ignoradas.

Jing e McKeown (1999) propuseram uma abordagem que faz uso de um Modelo oculto de Markov (*Hidden Markov Model*) (HMM) (Baum, 1972) para resolver o problema de decompor sentenças de sumários em função do texto fonte. O objetivo do programa de decomposição desenvolvido é determinar as relações entre n-gramas no sumário humano e n-gramas no texto que originou o sumário. Diferentemente dos trabalhos de Banko *et al*. (1999) e de Marcu (1999), em que a saída do programa era um extrato (o conjunto de sentenças e/ou orações do texto fonte que foram utilizadas para compor o sumário humano), no trabalho de Jing e McKeown (1999), os alinhamentos podem ser evidenciados na saída do programa.

Esse problema da decomposição é reduzido a encontrar a posição mais provável de uma palavra em uma sentença do sumário, no texto que ela foi originada, da seguinte maneira: uma sentença de um sumário de entrada pode ser representada com uma sequência de palavras $(I_1, \dots, I_N)$, em que I_1 é a primeira palavra da sentença e I_N a última. A posição de uma palavra em um documento ($(SNUM, WNUM)$) pode ser identificada pela posição da sentença no texto ($SNUM$) e a posição da palavra na sentença ($WNUM$) (por exemplo, (2, 4) refere-se à quarta palavra na segunda sentença). Várias ocorrências de uma mesma palavra no documento podem ser representadas como uma lista de posições $\{(SNUM_1, WNUM_1), \dots, (SNUM_m, WNUM_m)\}$. Então, dada uma sequência de palavras $(I_1, \dots, I_N)$ e para cada palavra da sequência sua posição (posições) no texto original, ou seja, $\{(SNUM_1, WNUM_1), \dots, (SNUM_m, WNUM_m)\}$, é determinada para cada palavra na sequência sua posição mais provável no texto.

É apontado pelos autores que a posição de uma palavra em um texto depende das posições das palavras que estão a sua volta. Sendo $PROB(I_{i+1}|I_i)$, em que I_{i+1} e I_i são duas palavras adjacentes no sumário, a probabilidade de uma palavra I_{i+1} vir de uma palavra $W1$ e da sentença $S1$ e de I_i vir de uma palavra $W2$ e da sentença $S2$, o seguinte processo de decomposição (HMM) é criado: se duas palavras pertencentes a uma mesma sentença forem adjacentes no documento, a $PROB(I_{i+1}|I_i)$ recebe o valor máximo P1; se uma palavra for posterior a outra na mesma sentença do documento, a $PROB(I_{i+1}|I_i)$ recebe o segundo valor máximo P2; se uma palavra for anterior a outra na mesma sentença do documento, a $PROB(I_{i+1}|I_i)$ recebe o terceiro valor máximo P3; se a palavra pertencer a uma sentença anterior a sentença que pertence, a $PROB(I_{i+1}|I_i)$ recebe o quarto valor máximo P4 (sendo que a distância das duas sentenças não pode ser maior que i, em que $1 < i < CONST$) (sendo $CONST$ um número constante pequeno, como 3 ou 5); se a palavra pertencer a uma sentença posterior a sentença que pertence, a $PROB(I_{i+1}|I_i)$ recebe o quinto valor máximo P5 (sendo que a distância das duas sentenças não pode ser maior que i, em que $1 < i < CONST$); e, por fim, a $PROB(I_{i+1}|I_i)$ receberá o valor P6 se a palavra pertencer a uma sentença com uma distância maior que i em relação à sentença que pertencer. Esse processo pode ser visto sintetizado na Figura 6. Os valores de P1 a P6 são atribuídos de forma experimental. No experimento dos autores, o valor máximo atribuído é 1 e os outros são atribuídos decrescendo os valores em 0,9; 0,8, e assim por diante.

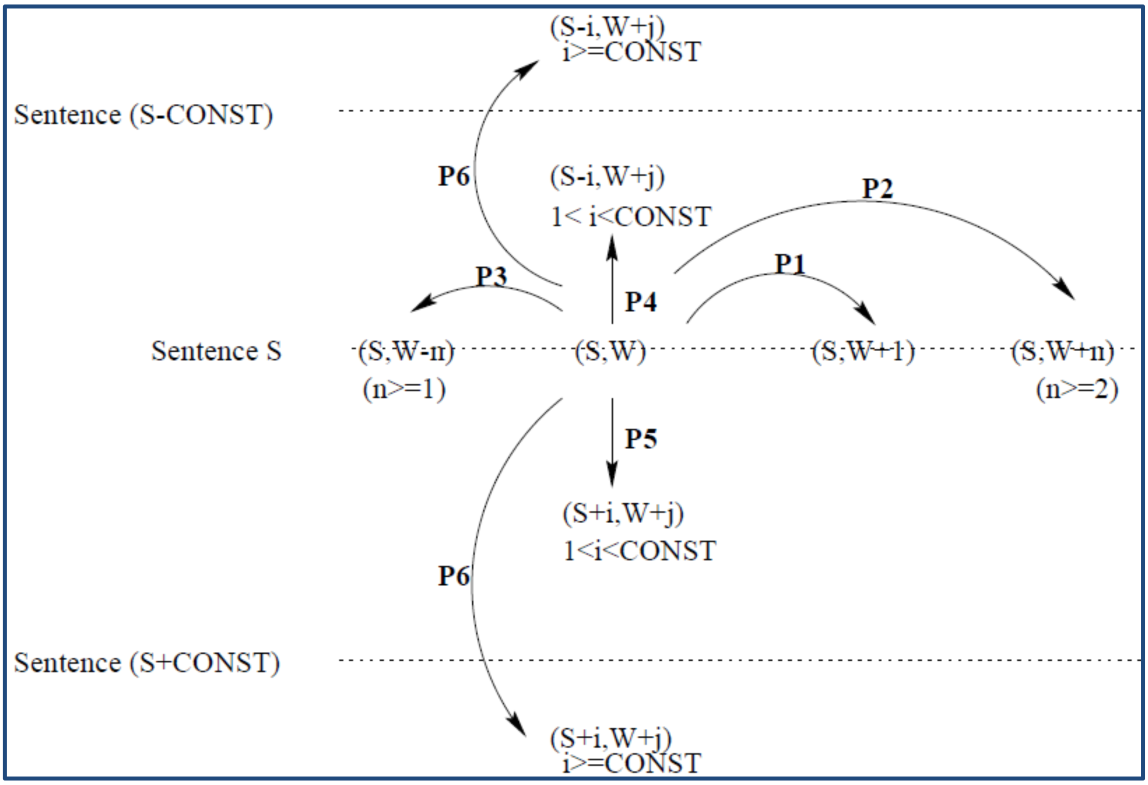


Figura 6: O HMM (Jing e McKeown, 1999)

O HMM descrito anteriormente é criado a partir de um conjunto de heurísticas baseadas em técnicas de corte e cola que os humanos usam para construir sumários, que foram evidenciadas pelos autores. Jing e McKeown (1999) evidenciaram 6 principais técnicas baseado em uma análise manual de mais de 120 sentenças em 15 sumários humanos: redução sentencial, combinação sentencial, transformação sintática, parafraseamento lexical, generalização/especificação, e reordenação.

A intuição desse método segue as seguintes regras produzidas heuristicamente a partir das operações de corta e cola: (1) duas palavras adjacentes no sumário são provavelmente provenientes de duas palavras adjacentes no documento; (2) palavras adjacentes do sumário são provavelmente provenientes de uma mesma sentença no documento, mantendo a ordem relativa entre elas; (3) palavras adjacentes em um sumário são provavelmente provenientes de uma mesma sentença no documento, invertendo a ordem relativa entre elas; (4) palavras adjacentes em um sumário podem vir de sentenças próximas no documento e manter sua ordem relativa; (5) palavras adjacentes em um sumário podem vir de sentenças próximas no documento e inverter sua ordem relativa; e (6) palavras adjacentes em um sumário provavelmente não são provenientes de sentenças distantes no documento.

O algoritmo de Viterbi (Viterbi, 1967) é utilizado para encontrar, de forma incremental, a sequência de posições mais provável no documento para cada palavra na sentença do sumário, ou seja, que maximiza a probabilidade $PROB(I_1, \ldots, I_N)$. Primeiro, encontra a sequência mais provável para $(I_1 I_2)$, para cada posição possível de I_2. Essa informação é então utilizada para computar a sequência mais provável para $(I_1 I_2 I_3)$, para cada possível posição de I_3. Esse processo é repetido até que todas as palavras da sequência tenham sido consideradas.

Dois experimentos foram realizados. No primeiro, foram utilizados 10 documentos do córpus Ziff-Davis do mesmo experimento feito por Marcu (1999). O programa construído fornece um conjunto de sentenças relevantes do documento para cada sentença do sumário, como pode ser visto na saída do programa, no Quadro 25, e, tomando a união das sentenças selecionadas, um extrato pode ser feito. O extrato foi comparado aos extratos de referência feitos baseados na maioria dos julgamentos humanos. As médias dos desempenhos atingidas pelos 14 juízes humanos e as médias atingidas pelo programa descrito podem ser vistas na Tabela 3.

Summary sentence:
(F0:S1 **arthur b sackler vice president for law and public policy of time warner inc**) (F1:S-1 *and*) (F2:S0 **a member of the direct marketing association told**) (F3:S2 **the communications subcommittee of the senate commerce committee**) (F4:S-1 *that legislation*) (F5:S1**to protect**) (F6:S4 **children' s**) (F7:S4 **privacy**) (F8:S4 **online**) (F9:S0 **could destroy the spontaneous nature that makes the internet unique**)

Source document sentences:
<u>Sentence 0:</u> a proposed new law that would require web publishers to obtain parental consent before collecting personal information from children (**F9 could destroy the spontaneous nature that makes the internet unique**) (**F2 a member of the direct marketing association told**) a senate panel thursday
<u>Sentence 1:</u> (**F0 arthur b sackler vice president for law and public policy of time warner inc**) said the association supported efforts (**F5 to protect**) children online but he urged lawmakers to find some middle ground that also allows for interactivity on the internet
<u>Sentence 2:</u> for example a child's e-mail address is necessary in order to respond to inquiries such as updates on mark mcguire's and sammy sosa's home run figures this year or updates of an online magazine sackler said in testimony to (**F3 the communications subcommittee of the senate commerce committee**)
<u>Sentence 4:</u> the subcommittee is considering the (**F6 children's**) (**F8 online**) (**F7 privacy**) protection act which was drafted on the recommendation of the federal trade commission

Quadro 25: Saída do programa (Jing e McKeown, 1999)

Tabela 3: Resultados do experimento 1 (Jing e McKeown, 1999)

	Precisão	**Cobertura**	**Medida-F**
Juízes humanos	88,8%	84,4%	85,7%
Programa	81,5%	78,5%	79,1%

Os resultados obtidos pelo programa ficaram relativamente próximos ao julgamento dos juízes humanos. Nota-se que mesmo o resultado dos juízes não alcançou valores acima de 90%. Isso pode indicar que a tarefa de alinhamento é complexa e/ou subjetiva.

Para o segundo experimento, foram selecionados 50 sumários do córpus para serem passados para o programa e, em seguida, um humano deveria julgar os resultados da decomposição. Das 305 sentenças nos 50 sumários, 18 (6,2%) sentenças foram decompostas de forma errada, então a acurácia atingida foi de 93,8%. A maioria dos erros ocorreu quando as sentenças do sumário não eram construídas por operações de corta e cola.

O programa de decomposição desenvolvido em Jing e McKeown (1999) também foi utilizado para alinhar 300 sumários humanos de artigos de notícias. Os sumários continham 1642 sentenças no total, variando de 2 sentenças por sumário para 21 sentenças por sumário. Os resultados relatados mostraram que 315 (19%) das sentenças não tinham sentenças correspondentes no documento, 686 (42%) sentenças alinharam com uma sentença única no documento, 592 (36%) sentenças alinharam com 2 ou 3 sentenças no documento e, apenas 49 (3%) alinharam com mais de 3 sentenças no documento. Essas estatísticas comprovadas por Jing e McKeown (1999) também são relativas à sumarização multidocumento, pois, quando se trata de mais de um texto gerando um único texto reduzido, espera-se que existam mais alinhamentos 1-N do que alinhamentos 1-1.

O programa possui dois tipos de erros: (i) pode falhar em encontrar sentenças equivalentes semanticamente que possuem palavras bastante diferentes, e (ii) pode identificar uma sentença não relevante do documento como relevante por conter algumas palavras em comum com a sentença do sumário. Uma edição posterior é feita para cancelar falsos casamentos como esses, mas essa edição não os remove completamente.

No quesito do alinhamento em si, a abordagem descrita pelos autores torna-se bastante positiva, pois com ela é possível observar os alinhamentos retornados pelo programa desenvolvido.

Hatzivassiloglou *et al.*, (1999) propuseram uma abordagem próxima ao alinhamento entre textos e sumários. A abordagem proposta foi utilizada para encontrar segmentos textuais, como parágrafos ou sentenças, que possuem a mesma informação, porém os textos utilizados foram textos de notícias previamente separados em subconjuntos que possuem o mesmo tópico. Como os textos possuíam o mesmo tópico, a abordagem proposta poderia ser utilizada para realizar o alinhamento entre textos e sumários, pois estes tratam de um assunto único. Para medir a distância semântica entre pares de segmentos textuais, os autores apresentam uma medida de similaridade composta que combina informação de múltiplos indicadores linguísticos. Para isso, vários atributos são investigados e, para selecionar a combinação ótima deles, usam aprendizado de máquina em um aprendizado supervisionado. Um vetor de atributos é computado sobre um par de unidades textuais, em que os atributos são ou primitivos, consistindo de uma característica, ou compostos, que consistem de pares de atributos primitivos.

Os atributos primitivos são cinco: (i) co-ocorrência de palavra (*word co-occurrence*), em que é feito o casamento da mesma palavra entre segmentos textuais, (ii) casamento de sintagmas nominais (*noun phrase matching*), em que é feito o casamento entre sintagmas nominais simples que possuem o mesmo núcleo, (iii) sinônimos da WordNet[14] (Miller, 1995; Fellbaum, 1998) (*WordNet synonyms*), em que são casadas palavras que estejam no mesmo conjunto de sinônimos (*synset*) da WordNet, (iv) classes semânticas comum para verbos (*common semantic classes for verbs*), em que são casados dois verbos que tenham a mesma classe semântica, e (v) nomes próprios compartilhados (*shared proper nouns*), em que é feito o casamento

[14] Wordnet é uma base lexical em inglês em que substantivos, verbos, adjetivos e advérbios estão agrupados em conjuntos de sinônimos denominados synsets, e cada *synset* está interligado por meio de relações conceituais-semânticas e lexicais a outros *synsets*.

entre nomes próprios fazendo algumas restrições sobre eles como, por exemplo, restringindo-os ao tipo do nome próprio de pessoa, de lugar ou de uma organização.

Os atributos compostos são três: (i) ordenação (*ordering*), em que dois pares de elementos primitivos, ou seja, palavras/sintagma nominais que foram consideradas como os atributos primitivos, precisam ter a mesma ordem relativa em ambos os segmentos textuais, (ii) distância (*distance*), em que dois pares de elementos primitivos precisam ocorrer com certa distância nos dois segmentos textuais, e (iii) primitiva (*primitive*), em que cada elemento do par de elementos primitivos pode ser restrito a uma primitiva específica, permitindo mais expressividade nos atributos compostos. Exemplos dos atributos podem ser vistos a seguir, nos Quadro 26, 27 e 28.

(a) *An OH-58 helicopter, carrying a crew of* ***two****, was on a routine training orientation when* ***contact*** *was lost at about 11:30 a.m. Saturday (9:30 p.m. EST Friday).*
(b) *"There were* ***two*** *people on board," said Bacon. "We lost radar* ***contact*** *with the helicopter about 9:15 EST (0216 GMT)."*

Quadro 26: Exemplo de atributo composto ordenação (Hatzivassiloglou, 1999, p. 206)

No Quadro 26, é possível ser visto um exemplo de atributo composto. Com uma restrição sobre a ordem dos elementos (atributo "ordenação"), o par "*two*" e "*contact*" seria considerado um casamento, pois as duas palavras ocorrem com a mesma ordem relativa nos dois segmentos textuais.

(a) *An OH-58 helicopter, carrying a crew of two, was on a routine training orientation when* ***contact*** *was* ***lost*** *at about 11:30 a.m. Saturday (9:30 p.m. EST Friday).*
(b) *"There were two people on board," said Bacon. "We* ***lost*** *radar* ***contact*** *with the helicopter about 9:15 EST (0216 GMT)."*

Quadro 27: Exemplo de atributo composto distância (Hatzivassiloglou, 1999, p. 206)

No exemplo do Quadro 27, o par "*lost*" e "*contact*" seria considerado um casamento de acordo com o atributo "distância", pois os dois elementos ocorrem com a mesma distância nos dois segmentos textuais.

(a) *An OH-58 helicopter, carrying a crew of two, was on a routine training orientation when **contact** was **lost** at about 11:30 a.m. Saturday (9:30 p.m. EST Friday).*
(b) *"There were two people on board," said Bacon. "We **lost** radar **contact** with the helicopter about 9:15 EST (0216 GMT)."*

Quadro 28: Exemplo de atributo composto primitiva (Hatzivassiloglou, 1999, p. 206)

No Quadro 28, pode-se ver um exemplo do atributo composto "primitiva", tendo sido restringido que a primeira primitiva deveria ser um sintagma nominal que possui um casamento com outro sintagma nominal (ou seja, ambos possuem a mesma "*head*") e a outra primitiva ser um verbo que possui um casamento, sendo que os verbos devem possuir um casamento de acordo com a primitiva de verbos (ou seja, os dois verbos devem possuir a mesma classe semântica).

Como dito anteriormente, para cada par de unidades textuais, é computado um vetor de atributos primitivos e compostos. Para saber se as unidades casam totalmente, é usado um algoritmo de aprendizado de máquina, o RIPER (Cohen, 1996), um efetivo sistema de indução de regras.

Devido aos textos do córpus utilizado serem de notícias, que costumam ter parágrafos pequenos (com uma ou duas sentenças normalmente), os segmentos textuais escolhidos foram os parágrafos, mas sentenças também podem ser igualmente utilizadas. No total, restaram 264 unidades textuais e 10345 comparações entre unidades. Os textos foram retirados da parte *Reuters* do córpus TDT[15] (*Topic Detection and Tracking*) de 1997. Esse córpus possui 16000 notícias da *Reuters* e da CNN[16] em que vários dos artigos estão agrupados manualmente em 25 categorias.

Foi usado *Three-fold Cross-validation*, dividindo-se aleatoriamente os 10345 pares de parágrafos em três subconjuntos de tamanhos quase iguais. Em cada um dos três turnos, dois subconjuntos foram usados para treinamento e o outro para teste. Para criar um padrão de referência, toda a coleção de 10345 pares de parágrafos foi marcada por dois anotadores.

Foi realizado um experimento para validar a definição de similaridade do trabalho, dando a três juízes 40 pares de parágrafos para serem marcados como

[15] http://projects.ldc.upenn.edu/TDT/
[16] http://www.cnn.com/

similares ou não similares. Eles concordaram em 97,6%. Apesar da grande concordância que tiveram, pelo fato da maioria dos segmentos (aproximadamente 97%) ter sido considerado não similar, é necessário realizar uma concordância que considere não só essa concordância esperada pelo acaso, por isso, também foi avaliada a concordância com a medida kappa[17] (Cohen, 1960; Carletta, 1996), resultando em um valor de 0,58. O sistema conseguiu recuperar 36,6% dos parágrafos similares com 60,5% de precisão e teve uma acurácia de 98,8%, sendo acurácia a porcentagem total de respostas corretas.

A “desvantagem” do experimento realizado pelos autores foi a utilização de vários textos. Apesar de serem textos que tratavam do mesmo assunto, resultava que poucos segmentos eram alinhados. Com a utilização de um texto e um sumário, os valores de cobertura e precisão devem ser melhorados, por diminuir a diferença entre os textos considerados.

Em 2001, Hatzivassiloglou, *et al*. relataram melhorias na abordagem descrita por Hatzivassiloglou *et al*. (1999), vindo a chamar o produto da abordagem de SimFinder, uma ferramenta que utiliza medição estatística de similaridade e é capaz de organizar pequenos segmentos textuais de um ou mais documentos em grupos. Nesse trabalho, os autores especificaram o uso da ferramenta para a sumarização automática, apontando que descobrir informações de similaridade entre textos auxilia na sumarização automática multidocumento[18]. Entre as melhorias na abordagem, estão incluídos o refinamento de atributos e a inclusão de novos atributos. Da mesma forma que a abordagem de Hatzivassiloglou *et al*. (1999), a abordagem de Hatzivassiloglou *et al*. (2001) foi utilizada para encontrar alinhamentos entre textos, procedimento para encontrar as possíveis informações relevantes dos textos que poderão compor o sumário multidocumento. Para o treinamento e a avaliação, foram utilizados, novamente, textos da parte *Reuters* do córpus TDT (*Topic Detection and Tracking*) de 1997, resultando, dessa vez, em 10535 (190 a mais que no experimento em

[17] A medida kappa é uma medida baseada no número de casos em que o resultado é igual entre os juízes, calculada de acordo com a fórmula: $Kappa = \frac{O-E}{1-E}$, em que O é a concordância observada entre os juízes e E é a concordância esperada, ou seja, E representa a probabilidade de os juízes concordarem pela sorte. A vantagem dessa medida é retirar a concordância pela sorte.

[18] Isso parte da pressuposição de que segmentos textuais recorrentes nos textos são provavelmente os mais centrais, que contêm as informações mais importantes dos textos, e, portanto, poderiam ser utilizados para compor o sumário multidocumento.

Hatzivassiloglou *et al.*, 1999) . Os pares foram manualmente anotados por 2 juízes. Os resultados para esse córpus podem ser vistos na Tabela 4.

Tabela 4: Resultados (Hatzivassiloglou *et al.*, 2001)

	Precisão	**Cobertura**	**Medida-F**
Hatzivassiloglou *et al.* (1999)	44,1%	44,4%	44,2%
Hatzivassiloglou *et al.* (2001)	49,3%	52,9%	51,0%

Nota-se que, em relação à abordagem anterior, e com a utilização do córpus também da abordagem anterior, a precisão mudou de 60,5% para 44,1%, e a cobertura mudou de 36,6% para 44,4%. Isso decorre dos textos do córpus terem sido novamente escolhidos e, por isso, resultou na diferença dos valores. Nota-se também que, com a utilização do córpus da abordagem descrita em Hatzivassiloglou *et al.* (2001), todos os valores aumentaram.

Barzilay e Elhadad (2003) propõem uma abordagem para alinhar textos comparáveis monolíngue[19] no nível sentencial. Essa abordagem poderia servir para qualquer aplicação de geração de texto a partir de um texto, como a sumarização automática e a simplificação textual. O método enfatiza a busca por um alinhamento global enquanto depende de uma função de similaridade local.

O algoritmo desenvolvido segue quatro passos principais. Os passos 1 e 2 acontecem em tempo de treinamento, e os passos 3 e 4 acontecem quando um novo par de textos deve ser alinhado. O passo 1, chamado de indução da estrutura topical (*topical structure induction*), envolve a análise de múltiplas instâncias de parágrafos dentro dos textos de cada coletânea, e as características topicais de cada coleção são identificadas por meio de agrupamento. Cada parágrafo no conjunto de treinamento é atribuído ao tópico que ele verbaliza. Para o passo 2, chamado de aprendizado de regras de mapeamento estrutural (*learning of structural mapping rules*), usa-se o conjunto de treinamento para que regras para o mapeamento de parágrafos sejam

[19] Textos comparáveis monolíngues são textos que possuem a mesma informação a ser transmitida em uma mesma língua.

aprendidas (de uma forma supervisionada). Para o aprendizado das regras, é utilizado um conjunto de pares de parágrafos como instâncias de treinamento, sendo que os atributos são a similaridade entre os dois parágrafos, e os números dos clusters (os clusters são obtidos na etapa anterior) dos parágrafos. Para medir a similaridade uma simples medida de cosseno é utilizada, que é baseada na sobreposição das palavras dos parágrafos.

O passo 3 do algoritmo, chamado de alinhamento macro (*macro alignment*), acontece quando um novo par de textos deve ser alinhado. Então, dado um novo par, cada parágrafo recebe um tópico, e os parágrafos são mapeados seguindo-se as regras aprendidas. Dessa etapa, resulta um parágrafo de um texto mapeado para possíveis parágrafos do outro texto. Por fim, o passo 4 do algoritmo é o alinhamento micro (*micro alignment*), em que, para cada parágrafo mapeado, um alinhamento local é computado em nível de sentença e, para isso, é novamente utilizada a medida de cosseno. O alinhamento final do par é a união de todos os alinhamentos sentenciais dos pares. Resumindo-se os passos anteriores, tem-se que, no passo 1, são agrupados os parágrafos, no passo 2, regras são aprendidas, no passo 3, um novo par de parágrafo é agrupado, e, no passo 4, o alinhamento de sentenças é encontrado.

Os arquivos de dados escolhidos para a avaliação foram duas coletâneas da *Encyclopedia Britannica*[20] e da *Britannica Elementary*. Os artigos da primeira enciclopédia são longos e detalhados e os da segunda contêm entradas de uma a duas páginas voltadas para crianças. Foram coletados 103 pares de descrições de cidades; 11 pares foram selecionados para teste e o resto (92 pares) foi utilizado para o agrupamento. Cada par dos conjuntos de treino e teste foi anotado manualmente por dois anotadores, e um terceiro anotador tinha a função de decidir o destino dos pares de sentenças que provocaram desacordo aos anotadores iniciais. Por fim, 320 pares de sentença foram alinhados no conjunto de treinamento e 281 no conjunto de teste. Os pares de outras sentenças que não foram alinhados serviram como exemplos negativos, gerando um total de 4.192 instâncias de treinamento e 3.884 instâncias de teste.

[20] http://www.britannica.com/

Com a abordagem utilizada, os autores chegaram aos seguintes resultados: a precisão do método completo atingiu o valor de 76,9% quando a cobertura atingia o valor de 55,8%.

Uma possível desvantagem dessa abordagem é a utilização da medida do cosseno para medir a similaridade dos segmentos. Se os dois segmentos possuírem a mesma informação, dita com palavras bastante diferentes, o alinhamento poderá não ocorrer.

Daumé III e Marcu (2004, 2005) propõem um modelo para criar alinhamentos palavra-palavra e sintagma-sintagma (*phrase-phrase*) entre documentos e seus *abstracts* e, para isso, usam uma extensão de um HMM. Um exemplo de duas sentenças alinhadas em nível de palavra e sintagma pode ser visto na Figura 7.

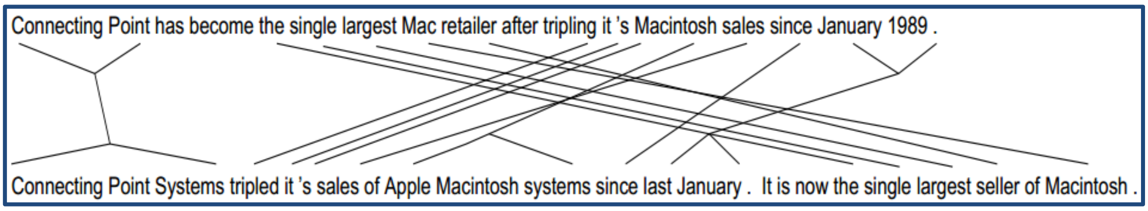


Figura 7: Exemplo de alinhamento em nível de palavra e sintagma (Daumé e Marcu, 2004, 2005)

No exemplo da Figura 7, há alinhamentos entre palavras (por exemplo, os alinhamentos (i) "*the*" e "*the*", (ii) "*Mac*" e "*Macintosh*", e (iii) "*retailer*" e "*seller*") e há alinhamentos entre sintagmas ((i) "*Connecting Point*" e "*Connecting Point Systems*", (ii) "*Macintosh*" e "*Apple Macintosh systems*", e (iii) "*January 1989*" e "*last January*").

Para realizar o alinhamento, os autores, a partir de observações, postulam uma história gerativa[21] de como um sumário é produzido a partir de um documento. A história gerativa criada pode ser vista a seguir.

(1) Repetir até que todo o sumário tenha sido gerado

(a) Escolher uma posição j no documento e pular ($jump$) para ela.

(b) Escolher um sintagma de tamanho l no documento.

(c) Gerar um sintagma no sumário baseado no vão de posições j a $j + l$ no documento.

[21] Uma história gerativa é a história de como os dados foram produzidos/formados.

(2) Pular ($jump$) para o final do documento.

Para que um sintagma no sumário possa ter sido gerado por uma palavra nula, é admitido que seja possível pular para um estado nulo (*null*). Um exemplo desse processo pode ser visto no Quadro 29, em que é demonstrado o começo e o final do processo gerativo que resulta no alinhamento da Figura 7.

- Pule para a primeira palavra do documento e escolha tamanho (*length*) 3; - Gere o sintagma do sumário "*Connecting Point*" baseado no sintagma do documento "*Connecting Point Systems*"; - Pule para um estado nulo (*null*); - Gere a palavra do sumário "*has*" a partir de nulo; - Gere a palavra do sumário "*become*" a partir de nulo; - Pule de nulo para a quarta palavra do documento e escolha tamanho 1; - Gere o sintagma do sumário "*tripling*" dado "*tripled*"; - Pule para a quinta palavra do documento e escolha tamanho 1; ... - Pule para a décima terceira palavra do documento ("*last*") e escolha tamanho 2; - Gere o sintagma do sumário "*January 1989*" dado "*last January*"; - Pule para um estado nulo; - Gere a palavra do sumário "." a partir de nulo; - Pule de nulo para o final do documento

Quadro 29: Começo e fim de um processo gerativo (Daumé e Marcu, 2005, p. 8)

Como pode ser visto no exemplo do Quadro 29, a partir da história gerativa criada, é induzido o alinhamento entre o texto fonte e o sumário, sendo considerado que um sintagma do sumário é alinhado ao sintagma do documento que o "gerou".

Baseado na história gerativa anterior, modela-se o processo de geração de um sumário de acordo com duas distribuições: $jump(j' \mid j + l)$, a probabilidade de pular para uma posição j' no documento quando o sintagma anterior termina na posição $j + l$, e $rewrite(s \mid d_{j:j+l})$, a probabilidade de reescrita de gerar um sintagma s do sumário, dado que se consideram sub-sintagmas (*sub-phrases*) do documento d começando na posição j e terminando na posição $j + l$.

A partir então da história gerativa, é construído um HMM para calcular precisamente as probabilidades dos alinhamentos especificadas pelo modelo. A diferença com um HMM normal é o fato de em vez de apenas uma palavra ser gerada em cada transição, um sintagma inteiro é gerada. Esse HMM funciona da mesma maneira que um HMM tradicional: começando no estado inicial (*start state*), transições são feitas pelo espaço de transições de acordo com probabilidades de transição. A cada passo, uma ou mais observações são geradas. O processo termina quando é atingido o estado final (*end state*). O HMM para um documento composto das palavras “a” e “b” pode ser visto na Figura 8. Supondo-se que o sumário desse documento fosse formado pelas palavras “c” e “d”, e que o alinhamento entre as palavras fosse “a” alinhado com “cd”, e “b” restaria sem alinhamento, o caminho pelo HMM seria: estágio inicial → a → estágio final.

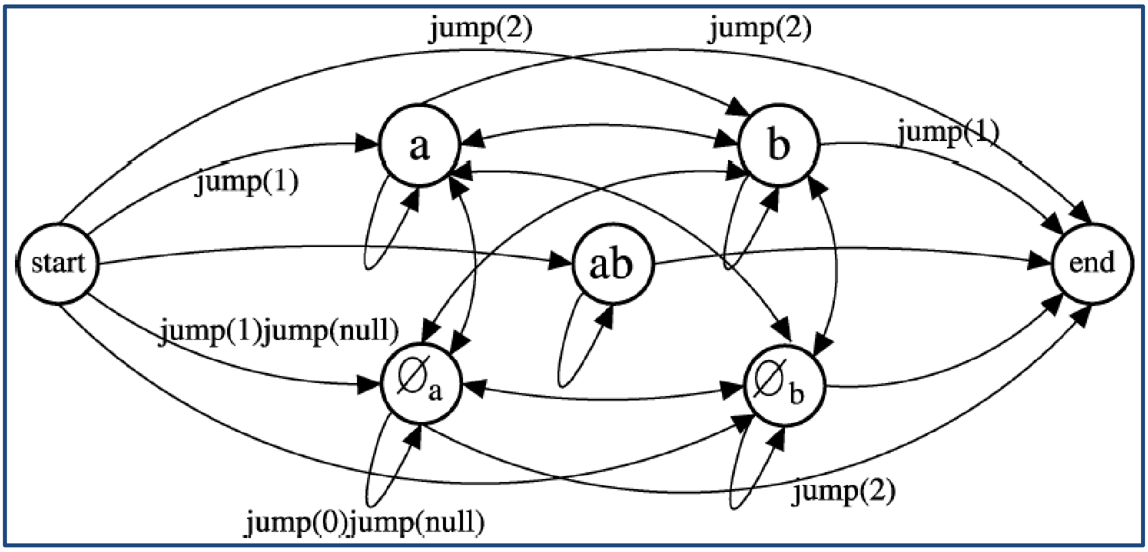


Figura 8: Desenho esquemático do HMM para o documento "ab" (Daumé e Marcu, 2005, p. 9)

Para formar uma tabela que conterá as probabilidades de $jump$ e $rewrite$, os autores utilizam a técnica de EM (*Expectation maximization*) (Dempster *et al.*, 1977). A ideia básica dessa técnica é fazer um palpite dos alinhamentos e utilizar esse palpite para estimar os parâmetros para as distribuições. Assim, são utilizadas essas distribuições re-estimadas para fazer palpites melhores de alinhamentos, e assim utilizar os alinhamentos melhores para re-estimar os parâmetros novamente. O processo continua até ser interrompido quando, por exemplo, os valores de probabilidade pararem de variar de forma significativa (convergirem).

Para a avaliação do modelo, os experimentos foram realizados sobre o córpus Ziff-Davis e o conjunto de alinhamentos de referência foi construído manualmente por dois anotadores. Foram selecionados de forma aleatória 45 pares de documentos-sumários. Os primeiros 5 foram anotados e discutidos de forma conjunta e os 40 restantes foram anotados de forma individual. Os anotadores alinharam no nível de sintagmas, decidindo entre alinhamentos possíveis ou certos, e depois esses alinhamentos foram convertidos para o nível de palavras também. O valor para a concordância kappa entre os anotadores foi de 0,63 para os alinhamentos considerados como certos.

Foram selecionados os 2000 menores pares documento-sumário do córpus Ziff-Davis para treinamento, porém, apenas 12 dos pares anotados manualmente estavam contidos nesse conjunto, então o restante dos pares foi inserido, resultando em 2033 pares documento-sumário para treino. Os melhores resultados obtidos para a abordagem podem ser vistos na Tabela 5.

Tabela 5: Resultados (Daumé III e Marcu, 2005)

Precisão	**Cobertura**	**Medida-F**
52,2%	71,2%%	60,6%

Observa-se, em relação ao trabalho de Marcu (1999), um dos autores dessa abordagem, que os valores dos resultados foram piores. Isso deve acontecer devido à complexidade maior para se alinhar sintagmas em relação a orações ou sentenças.

Até então, nenhuma das abordagens discutidas nos trabalhos de Kupiec *et al*. (1995), Banko *et al*. (1999), Marcu (1999), Jing e McKeown (1999), Hatzivassiloglou *et al*. (1999, 2001), Barzilay e Elhadad (2003), e Daumé III e Marcu (2004, 2005) eram explicitamente utilizadas para realizar alinhamentos entre um sumário multidocumento e seus documentos de origem. Em 2004, Hirao *et al*. propuseram uma abordagem e fizeram experimentos para córpus que continham tanto sumários monodocumento como sumários multidocumento. A língua utilizada em sua abordagem foi o japonês. Foi proposta uma abordagem que utiliza caminhos da árvore de dependência (*Dependency Tree Path*) (DTP) e uma medida de similaridade chamada *Extended String Subsequence Kernel* (ESK) para resolver o problema do alinhamento entre vários documentos e um sumário multidocumento. O DTP é o caminho de uma folha até a raiz de uma árvore de dependência. Um exemplo de uma árvore de dependência pode ser visto na Figura 9. Essa árvore de dependência é comum a duas sentenças que possuem o mesmo sentido e que podem ser vistas na Figura 10.

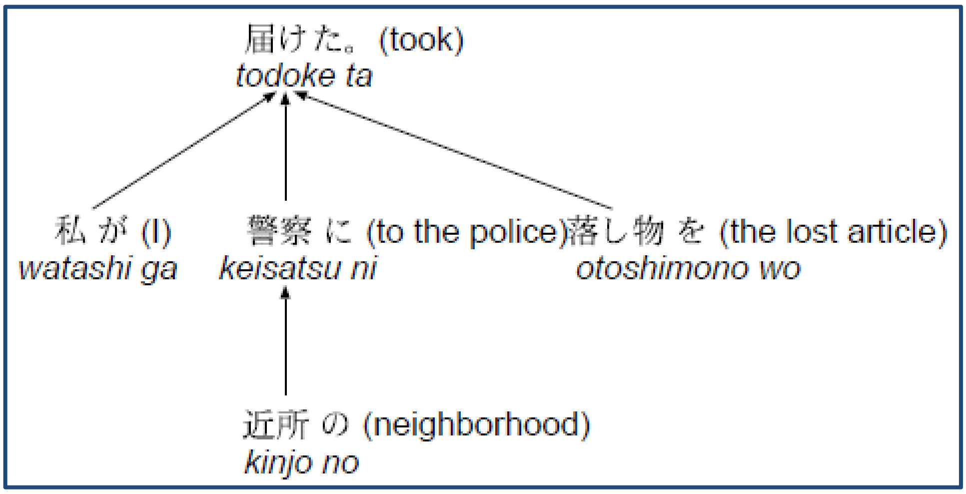


Figura 9: Árvore de dependência (Hirao et al., 2004)

Sentence 1: 私が近所の警察に落し物を届けた。
watashi ga kinjo no keisatsu ni otosimono wo todoke ta.

Sentence 2: 近所の警察に落し物を私が届けた。
kinjo no keisatsu ni otoshimono wo watashi ga todoke ta.

Figura 10: Sentenças cuja árvore de dependência é a mesma (Hirao et al., 2004)

A abordagem baseia-se no fato apontado pelos autores, e evidenciado pelas Figura 9 e 10, de que sentenças, quando refraseadas, têm a estrutura de suas árvores de dependência quase sempre iguais. A ideia do algoritmo é comparar as DTPs de todas as sentenças fonte com todas as sentenças do sumário utilizando uma medida de similaridade baseada em co-ocorrências de palavras, a ESK. A ESK é uma extensão de uma medida chamada *Word Sequence Kernel* (WSK) (Cancedda *et al*., 2003) e, por sua vez, a WSK é uma extensão de uma medida baseada em n-gramas usada para categorização textual. A árvore de dependência da Figura 9, por exemplo, possui três DTPs, retiradas do trabalho de Hirao *et al*. (2004), a saber:

- 私が届けた (*I took*)
- 近所の警察に届けた (*took to the neighborhood police*)
- 落し物を届けた (*took the lost article*)

O algoritmo utilizado para alinhar sentenças fonte com sentenças de seu *abstract* segue quatro passos. O primeiro passo é transformar todas as sentenças fonte em DTPs; o segundo passo é para cada sentença α no *abstract* aplicar os passos 3 e 4; o terceiro passo é transformar α em um conjunto de DTPs; e o quarto passo é, sendo $F(\alpha)$ o conjunto de DTPs de α, e $F(s_i)$ o conjunto de DTPs das i-th sentenças fonte, para cada $P_\alpha\big(\in F(\alpha)\big)$, alinhar uma sentença fonte como se segue: define-se $sim(P_\alpha, s_i) = max\ sim(P_\alpha, P)$, sendo $P\ \in F(s_i)$, em que, para $P(\alpha)$, alinha-se uma sentença fonte que satisfaz $argmax_{s_i \in Source} sim(P_\alpha, s_i)$. Com esse procedimento torna-se possível derivar correspondências N-N.

Os autores informam que, para chegar a um alinhamento robusto, informações sintáticas e semânticas são requeridas, e a DTP traz informações sintáticas enquanto a ESK traz informações semânticas. Com a ESK, é possível adicionar sentidos para cada palavra, e a utilização do sentido da palavra permite casamentos flexíveis até mesmo quando um parafraseamento foi utilizado nas sentenças do sumário.

No trabalho de Hirao *et al*. (2004), foram utilizados textos provenientes da TSC2 (*Text Summarization Challenge*) (Okumura *et al*., 2003), que possui dados de sumarização monodocumento (30 documentos) e multidocumento (224 documentos em 30 grupos). Para cada conjunto de dados, três especialistas fizeram *abstracts* pequenos e *abstracts* longos. Para cada conjunto, sentenças do sumário humano foram alinhadas às sentenças fonte.

Foi avaliado o desempenho do método para um documento e para múltiplos documentos, e os melhores resultados relatados, de medida-F, foram: 97.7% para um documento e 80.8% para múltiplos documentos.

Como pode ser visto, para mais de um documento, o valor de medida-F cai de 97.7% para 80.8%, demonstrando o quanto é mais difícil realizar alinhamentos entre mais de um texto fonte. Os autores apontam que em sumários multidocumento é comum a compactação, a combinação e a integração de sentenças e que, por isso, o alinhamento torna-se mais difícil.

Uma possível desvantagem desse método é que ele pode ser dependente de língua, pois é possível que as DTPs das árvores de sentenças semelhantes apenas sejam bastante próximas em línguas como o japonês.

Ainda na área da sumarização automática, existe um trabalho que propõe um método capaz de encontrar informações relevantes em segmentos textuais, o trabalho de Seno e Nunes (2008, 2009). Nesse trabalho, é apresentado o SiSPI, um sistema utilizado no processo de agrupamento de sentenças que contêm informações em comum, na tarefa de fusão sentencial. A fusão de sentenças é uma tarefa de geração textual em que, a partir de duas sentenças que tenham seu sentido relacionado, uma única sentença é gerada que preserva as informações comuns entre elas (Barzilay, 2003; Barzilay e Mckeown, 2005). Após o agrupamento de sentenças, um alinhamento em nível de *phrases* é realizado para que possa ser feita a fusão das sentenças que contêm informação em comum.

Um córpus foi criado para realizar a avaliação do método. Para sua criação, foram selecionados 50 conjuntos de textos jornalísticos na língua portuguesa do Brasil. Cada coleção possuía aproximadamente 4 documentos relacionados a um mesmo assunto, resultando em 1153 sentenças em 71 documentos. Cada coleção de documentos foi então agrupada utilizando o sistema SiSPI, desenvolvido no trabalho de Seno e Nunes (2008).

O sistema SiSPI é baseado em um método de agrupamento incremental. Dado um conjunto de documentos como entrada, o primeiro grupo é criado selecionando-se a primeira sentença do primeiro documento do conjunto. A cada iteração, o algoritmo verifica se é necessário incluir a nova sentença em um grupo já criado ou se outro grupo precisa ser criado para a sentença analisada. A decisão que o algoritmo faz baseia-se em duas funções de similaridade que calculam a distância entre uma sentença e um grupo de sentenças: uma baseada na medida *Word Overlap* (Radev *et al.*, 2008), que calcula o número de palavras em comum entre uma sentença S e um grupo C normalizado pelo total de palavras de S e C; e outra baseada na distância cosseno, aplicada entre o vetor de frequência de termos de uma sentença e o vetor que representa os termos mais importantes de um grupo.

Para avaliar o método, foram selecionadas 20 coleções de documentos do córpus de forma aleatória e cada sentença de uma coleção foi manualmente classificada/associada a um grupo para a construção de um córpus de referência. A maior média de medida-F obtida foi 88,6%.

O alinhador de *phrases* recebe como entrada as sentenças já agrupadas como similares em seu conteúdo pelo sistema SiSPI. São então identificados todos os possíveis alinhamentos entre as sentenças em nível de *phrase* e, por fim, as sentenças são unidas em uma única estrutura de dependência sintática. O resultado final, uma única estrutura de dependência sintática representando todas as sentenças do conjunto, pode ser visto exemplificado na Figura 11. Com esse sistema, são apenas alinhadas palavras de classes abertas (como substantivos, verbos, advérbios e adjetivos). No exemplo, as sentenças que originaram a estrutura são: “O Airbus A320, vôo JJ 3054, partiu de Porto Alegre, às 17h16 da terça-feira e chegou a São Paulo às 18h45” e “A aeronave da TAM Airbus A320, vôo JJ 3054, partiu de Porto Alegre, às 17h16 com destino a Congonhas”.

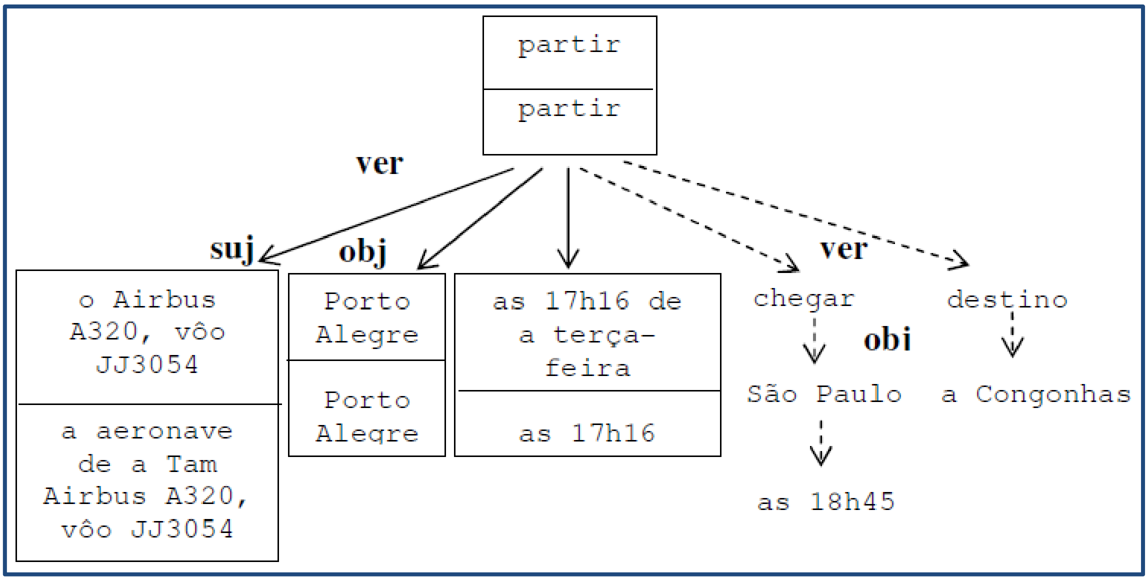


Figura 11: Exemplo - estrutura de dependência final (Seno e Nunes, 2009, p. 81)

A técnica de agrupamento inicial de Seno e Nunes (2008, 2009) pode ser utilizada no alinhamento na sumarização automática, por se tratar de um agrupamento de sentenças que contenham parte da mesma informação em comum.

Além das medidas utilizadas no trabalho de Seno e Nunes (2008, 2009), existem outras medidas de similaridade, como a medida min/max (Dagan, 2000), a medida Dice (Curran, 2003) e a divergência Jenson-Shannon (Lee, 1999), como apontado em Jurafsky e Martin (2009). Essas medidas são utilizadas para computar a distância entre dois vetores de, por exemplo, palavras.

Na Tabela 6, encontra-se um resumo dos trabalhos aqui descritos.

Tabela 6: Resumo dos trabalhos de alinhamento na sumarização automática

Autores	**Abordagens**
Kupiec *et al.*, (1995)	Utilizaram regras para realizar o alinhamento entre <u>um texto e seu sumário humano</u>.
Banko *et al.* (1999)	Utilizaram um modelo de *bag of words* para realizar o alinhamento entre <u>um texto e seu sumário humano</u>.
Marcu (1999)	Utilizou um modelo de *bag of words* para realizar o alinhamento entre <u>um texto e seu sumário humano</u>.
Jing e McKeown (1999)	Utilizaram um HMM para realizar o alinhamento entre

	um texto e seu sumário humano.
Hatzivassiloglou *et al*. (1999, 2001)	Utilizaram aprendizado de máquina para realizar o alinhamento de textos que falem sobre um mesmo assunto.
Barzilay e Elhadad (2003)	Utilizaram uma função de similaridade para realizar o alinhamento de textos que falem sobre um mesmo assunto.
Daumé III e Marcu (2004, 2005)	Utilizaram um HMM juntamente com a técnica EM para realizar o alinhamento entre um texto e seu sumário humano.
Hirao *et al*. (2004)	Utilizaram uma medida de similaridade para comparar o caminho das árvores de dependência de um texto com seu sumário humano e de mais de um texto com seu sumário humano multidocumento.
Seno e Nunes (2008, 2009)	Utilizaram medidas de similaridade para agrupar sentenças que contenham parte da informação em comum.

A seguir, no Capítulo 3, são apresentados os recursos que foram utilizados juntamente com o alinhamento manual entre os sumários humanos multidocumento e seus textos de origem do córpus CSTNews, bem como a tipificação dos mesmos.

Capítulo 3. Recursos

Nesse capítulo, são apresentados todos os recursos que foram utilizados para desenvolver este trabalho de mestrado.

3.1. Córpus CSTNews

O CSTNews (Cardoso *et al.*, 2011b) é um córpus anotado de acordo com a teoria CST descrita na Seção 2.2. O córpus é composto por 50 grupos de textos contendo cada um (i) 2 ou 3 textos sobre um determinado assunto, (ii) sumários humanos monodocumento e sumários humanos e automáticos multidocumento dos textos, (iii) as relações CST entre os textos, entre outras anotações. Quanto aos sumários, é interessante salientar que eles são abstracts, ou seja, foram criados com operações de reescrita, e sua taxa de compressão é de 70%. Os textos são provenientes dos jornais *online* Folha de São Paulo, Estadão, Jornal do Brasil, O Globo e Gazeta do Povo, e foram coletados no período de agosto a setembro de 2007. Cada grupo de texto possui uma seção, sendo elas: política, mundo, cotidiano, esporte, dinheiro e ciência. A distribuição de grupos de textos dentro de cada seção pode ser vista na Figura 12. A distribuição desbalanceada dos textos se deve a dificuldade de encontrar textos, que cobriam a mesma notícia, publicado por mais de um jornal *online*.

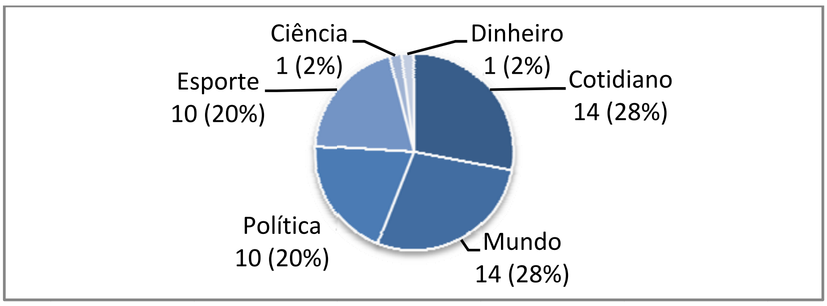


Figura 12: Distribuição das seções no córpus CSTNews

Os sumários possuem, em conjunto, 331 sentenças e os textos fonte 2067, sendo a média de sentenças de 6,62 e 41,34, respectivamente. Todos os dados estatísticos do córpus podem ser vistos na Tabela 7.

Tabela 7: Estatísticas do córpus CSTNews

Coleção	Categoria	Número de documentos	Número de sentenças por documentos	Número de sentenças por sumários
C1	Mundo	3	24	5
C2	Política	3	51	7
C3	Cotidiano	3	50	10
C4	Cotidiano	3	39	5
C5	Cotidiano	2	23	5
C6	Cotidiano	3	36	5
C7	Ciência	2	23	4
C8	Esportes	3	25	6
C9	Política	3	36	6
C10	Mundo	3	38	10
C11	Cotidiano	3	56	11
C12	Mundo	3	34	4
C13	Mundo	3	37	6
C14	Mundo	3	25	5
C15	Mundo	3	26	6
C16	Política	3	47	6
C17	Política	2	41	6
C18	Mundo	3	70	9
C19	Esportes	2	13	4
C20	Política	3	42	8
C21	Cotidiano	3	41	3
C22	Cotidiano	3	50	9
C23	Mundo	2	25	6
C24	Esportes	3	24	5
C25	Esportes	3	88	8
C26	Mundo	3	58	10
C27	Esportes	3	89	12
C28	Esportes	3	35	4
C29	Mundo	3	48	6
C30	Dinheiro	3	46	4
C31	Esportes	2	10	3
C32	Mundo	3	66	9
C33	Cotidiano	3	68	13
C34	Cotidiano	3	59	8
C35	Mundo	3	36	7
C36	Cotidiano	3	74	14
C37	Cotidiano	2	26	5

C38	Esportes	3	26	3
C39	Cotidiano	3	34	3
C40	Política	3	28	4
C41	Esportes	3	45	6
C42	Política	2	39	5
C43	Política	3	49	7
C44	Política	2	26	9
C45	Cotidiano	3	47	6
C46	Mundo	3	23	5
C47	Mundo	3	43	6
C48	Esportes	2	43	9
C49	Cotidiano	3	23	6
C50	Política	3	62	8
Total	—	140	2067	331
Média	—	2,8	41,34	6,62

A anotação CST foi realizada por quatro anotadores e, para auxiliar a anotação, foi utilizada a ferramenta CSTTool (Aleixo e Pardo, 2008). A CSTTool foi criada para fazer a segmentação textual e a detecção de pares de segmentos textuais candidatos a serem relacionados. Foram escolhidas as sentenças como segmentos textuais. Durante o treinamento, as relações foram refinadas, resultando nas já citadas 14 relações (Quadro 14, Página 19). A quantidade de cada relação encontrada na anotação pode ser vista na Figura 13, em que é possível notar que a maioria das relações é do tipo *Overlap*, e *Elaboration*, e assim por diante, e que outras relações tiveram nenhuma, ou quase nenhuma, aparição no córpus (como a relação *Citation*).

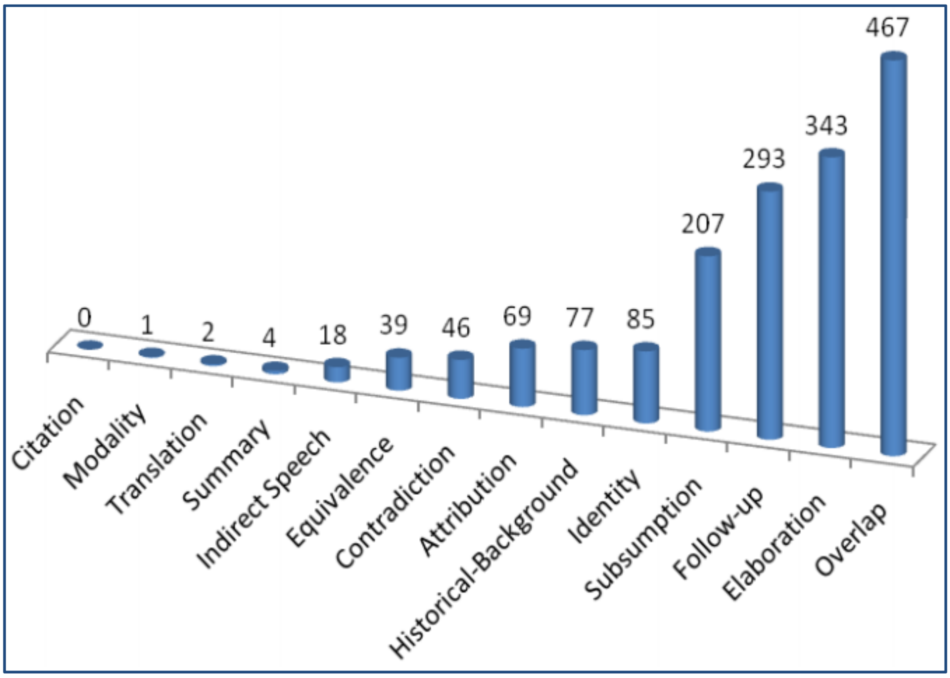


Figura 13: Relações presentes no córpus CSTNews (Cardoso *et al.*, 2011b, p. 101)

Para computar a concordância entre os anotadores, informação crucial para se ter noção da dificuldade e subjetividade de uma tarefa, foram utilizadas duas medidas: a kappa, explicada na Página 45 desta dissertação, e a concordância porcentual. A concordância porcentual foi computada no número de vezes que (i) os anotadores concordaram totalmente com a escolha da relação, (ii) os anotadores concordaram parcialmente com a escolha da relação e (iii) os anotadores não concordaram com nenhuma escolha da relação CST. Os valores de concordância obtidos pelos anotadores na anotação podem ser vistos nas Tabela 8 e 9. Foram avaliadas concordâncias para as relações em si, para a direcionalidade entre elas, e para os tipos das relações, presentes na tipologia da Figura 4 (Página 21). Para tarefas mais subjetivas como essa anotação (uma anotação discursiva), é esperado um valor de kappa não muito alto.

Tabela 8: Concordância kappa para a anotação CST (Cardoso *et al.*, 2011b)

Parâmetros de concordância	**Valores de concordância**
Relações	0,50
Direcionalidade	0,44
Tipos das relações	0,61

Tabela 9: Concordância porcentual para a anotação CST (em %) (Cardoso *et al.*, 2011b)

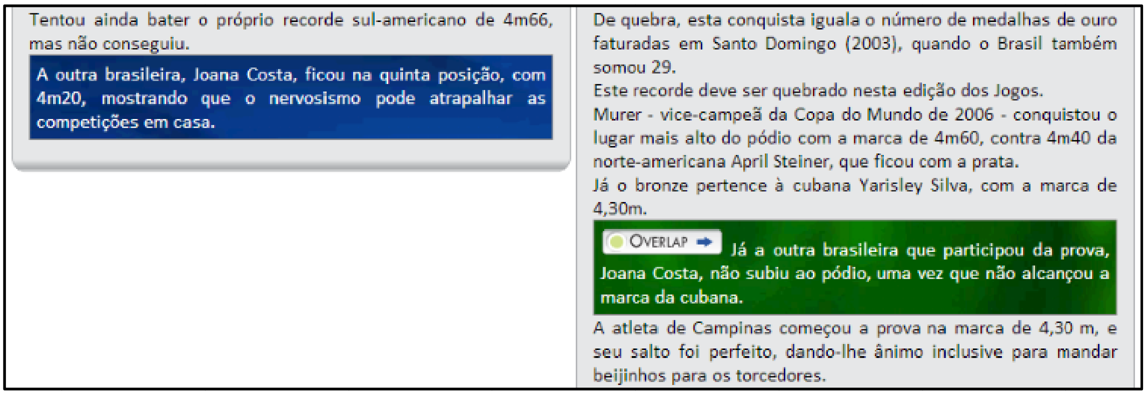
Parâmetros de concordância	Concordância total	Concordância parcial	Concordância nula
Relações	54	27	18
Direcionalidade	58	27	14
Tipos das relações	70	21	9

Do córpus CSTNews, foram utilizados os textos e os sumários multidocumento para o desenvolvimento deste trabalho de mestrado.

3.2. CST Parser

O CSTParser (Maziero e Pardo, 2011) é uma ferramenta cuja função é encontrar as relações CST entre textos. O *parser* recebe como entrada um grupo de textos que versam sobre um mesmo assunto e retorna um grafo de relações CST entre as sentenças dos textos, em que os nós do grafo são as sentenças e as arestas são as relações CST entre elas. Para cumprir essa tarefa, a ferramenta faz uso de aprendizado de máquina.

Analisando com mais detalhes, o primeiro passo que o CSTParser realiza é dividir os textos em sentenças e em seguida filtrar pares de sentença que contenham algumas palavras em comum utilizando uma medida de sobreposição de palavras, a *Word overlap*, (que será apresentada detalhadamente na Seção 4.1 de métodos). Essa filtragem é realizada para diminuir a quantidade de pares a serem classificados. Depois, cada par restante é analisado por várias ferramentas para extrair atributos importantes e esses atributos são utilizados no aprendizado de máquina, para poder realizar a classificação dos pares. O *parser* faz uso de classificadores para identificar as relações *Elaboration*, *Equivalence*, *Follow-up*, *Historical-background*, *Overlap* e *Subsumption*, e utiliza regras para identificar as relações *Attribution*, *Contradiction*, *Indirect-speech* e *Translation*, tendo uma acurácia geral de 68,57%. Na Figura 14, é possível ver um exemplo de uma porção da saída do CSTParser na WEB[22]. No exemplo, a sentença "*A outra brasileira, Joana Costa, ficou na quinta posição, com 4m20, mostrando que o nervosismo pode atrapalhar as competições em casa*." tem uma relação do tipo *Overlap* com a sentença "*Já a outra brasileira que participou da prova, Joana Costa, não subiu ao pódio, uma vez que não alcançou a marca da cubana.*".


Tentou ainda bater o próprio recorde sul-americano de 4m66, mas não conseguiu.
A outra brasileira, Joana Costa, ficou na quinta posição, com 4m20, mostrando que o nervosismo pode atrapalhar as competições em casa.

De quebra, esta conquista iguala o número de medalhas de ouro faturadas em Santo Domingo (2003), quando o Brasil também somou 29.
Este recorde deve ser quebrado nesta edição dos Jogos.
Murer - vice-campeã da Copa do Mundo de 2006 - conquistou o lugar mais alto do pódio com a marca de 4m60, contra 4m40 da norte-americana April Steiner, que ficou com a prata.
Já o bronze pertence à cubana Yarisley Silva, com a marca de 4,30m.
OVERLAP Já a outra brasileira que participou da prova, Joana Costa, não subiu ao pódio, uma vez que não alcançou a marca da cubana.
A atleta de Campinas começou a prova na marca de 4,30 m, e seu salto foi perfeito, dando-lhe ânimo inclusive para mandar beijinhos para os torcedores.

Figura 14: Exemplo - CSTParser

Para os usuários, um *parser* como esse pode ser utilizado para organizar melhor notícias, por exemplo, sendo possível observar onde existem informações contraditórias, redundantes, ou ainda ter um melhor conhecimento sobre a ordem dos eventos relatados no tempo. Para aplicações computacionais, o CSTParser pode ser utilizado, por exemplo, na sumarização automática. Neste trabalho, o CSTParser é utilizado para descobrir as relações CST entre os textos provenientes do CSTNews e seus sumários multidocumento, sendo que essas relações foram utilizadas em dois dos três métodos desenvolvidos neste trabalho.

Além dos recursos descritos, também foi utilizado o **alinhamento manual** (Agostini *et al.*, 2012, 2014) dos sumários multidocumento e seus textos fontes, disponível no córpus CSTNews, que será apresentado na seção seguinte.

3.3. Alinhamento Manual

Nessa seção, o alinhamento manual é apresentado com detalhes. Em relação a este trabalho de mestrado, a anotação foi utilizada como *Gold standard*, ou seja, foi

[22] Disponível em: http://www.nilc.icmc.usp.br/CSTParser

vista como a versão correta dos alinhamentos, para se comparar com os resultados dos alinhamentos produzidos automaticamente e assim julgar o quão bom os mesmos foram. Além disso, o alinhamento manual foi utilizado também em outro trabalho de mestrado (Camargo, 2013), que visava investigar sumários multidocumento com o objetivo de identificar estratégias de sumarização humana multidocumento, ou seja, estratégias de produção manual de sumário multidocumento.

Considerações iniciais

A tarefa do alinhamento manual consistiu em julgar o relacionamento entre todas as sentenças dos sumários multidocumento do córpus CSTNews com sentenças dos documentos fonte. A ideia básica que foi utilizada para o julgamento dos pares de sentença é o conteúdo que elas transmitem. Sendo assim, é possível encontrar alinhamentos entre duas sentenças que não possuem nenhuma unidade lexical em comum, o que resultaria em um *Word overlap* com o valor 0, como exemplificado no Quadro 30. Cada sentença de cada sumário multidocumento pode ser alinhada com zero ou mais sentenças dos textos fonte, como é exemplificado nos Quadro 31 e 32, a seguir. No Quadro 31, uma sentença do sumário é alinhada a duas sentenças dos textos fonte e, no Quadro 32, uma sentença do sumário é alinhada a três sentenças dos textos fonte. É possível observar como os pares de sentença possuem informações em comum e que várias das palavras que possuem são diferentes.

Sentença do sumário multidocumento	**Sentença do documento**
Vários moradores e turistas nas regiões, inclusive brasileiros, foram retirados dos locais, enquanto outros estão se preparando para a passagem do furacão.	Na Jamaica, muitos estocaram alimentos, água, lanternas e velas.

Quadro 30: Exemplo de alinhamento que possui *Word overlap* 0

No exemplo do Quadro 30 em especial, a sentença do sumário contém a informação que moradores e turistas de uma região “estão se preparando para a passagem de um furacão”. Na sentença do documento, são citadas formas específicas de se preparar para um furacão (a estocagem de alimentos, água, lanternas e velas) e, por esse motivo, as duas sentenças foram consideradas alinhadas.

Sentença do sumário multidocumento	Sentenças dos documentos
O Brasil não fará parte do trajeto de 20 países do revezamento da tocha.	A tocha passará por vinte países, mas o Brasil não estará no percurso olímpico.
	O Brasil não faz parte do trajeto da tocha olímpica.

Quadro 31: Exemplo de alinhamento (1-2)

Sentença do sumário multidocumento	Sentenças dos documentos
A medalha de prata ficou com a americana April Steiner com 4m40 e a de bronze com a cubana Yarisley Silva com 4m30.	Murer - vice-campeã da Copa do Mundo de 2006 - conquistou o lugar mais alto do pódio com a marca de 4m60, contra 4m40 da norte-americana April Steiner, que ficou com a prata.
	Já o bronze pertence à cubana Yarisley Silva, com a marca de 4,30m.
	A medalha de prata ficou com a americana April Steiner, com a marca de 4m40 e o bronze foi para a cubana Yarisley Silva, com 4m30.

Quadro 32: Exemplo de alinhamento (1-3)

Regras

Para julgar cada par, partindo da premissa básica de que as sentenças devem conter alguma informação em comum, foram criadas 8 regras, 4 gerais (regras 1 a 4) e 4 específicas (regras 5 a 8). Dessa forma, a anotação poderá ser replicada novamente com consistência. Basicamente, as regras definem em que casos as sentenças devem, e não devem, ser alinhadas. Todas as regras criadas podem ser vistas a seguir, nos Quadro 33 a 40.

Regra 1	
Alinhar com base na sobreposição de conteúdo e não de forma	
Exemplo	
Sentença do sumário	**Sentença do documento**
A expectativa de lideranças da Câmara e do Conselho de Ética é que pouco mais de 10% dos 69 deputados denunciados no relatório parcial da CPI **abrirão mão de seus mandatos**.	As **renúncias** têm que ser publicadas até terça-feira, quando o presidente do Conselho de Ética, deputado Ricardo Izar (PTB-SP), vai instaurar os processos de perda de mandato contra os 69 deputados acusados pela CPI dos Sanguessugas de envolvimento com a máfia das ambulâncias.

Quadro 33: Exemplo de regra de alinhamento (1)

Como pode ser visto no exemplo do Quadro 33, as duas sentenças foram alinhadas mesmo que não possuam exatamente as mesmas palavras para expressar o conteúdo. No caso, "abrir mão de mandato" e "renúncia" tem o mesmo significado.

Regra 2	
Alinhar com base na sobreposição da informação principal	
Exemplo	
Sentença do sumário	**Sentença do documento**
Usando telescópios do Observatório Europeu Sul (ESO), Ray Jayawardhana, da Universidade de Toronto, e Valentin D. Ivanov, do ESO, **descobriram** um planemo com sete vezes a massa de Júpiter, o planeta mais pesado do Sistema Solar, e outro com o dobro desse peso, que giram um ao redor do outro,	Os pesquisadores Ray Jayawardhana e Valentin D. Ivanov **informam** a descoberta na edição de quinta-feira do serviço online Science Express, mantido pela revista Science.

denominado Oph 162225-240515, o primeiro planemo duplo.	

Quadro 34: Exemplo de regra de alinhamento (2)

No exemplo do Quadro 34, as duas sentenças não são alinhadas graças a regra 2, pois a sentença do sumário expressa a descoberta do planemo pelos pesquisadores e a sentença do documento expressa o ato de informar essa descoberta (“descobriram” *versus* “informam”).

Regra 3	
Alinhar com base na sobreposição de informação secundária	
Exemplo	
Sentença do sumário	**Sentença do documento**
Usando telescópios do Observatório Europeu Sul (ESO), Ray Jayawardhana, da Universidade de Toronto, e Valentin D. Ivanov, do ESO, descobriram um planemo com sete vezes a massa de Júpiter, o planeta mais pesado do Sistema Solar, e outro com o dobro desse peso, que **giram um ao redor do outro**, denominado Oph 162225-240515, o primeiro planemo duplo.	Ambos os mundos têm massa semelhante à de outros exoplanetas já catalogados, mas não giram em torno de uma estrela - na verdade, **giram em torno um do outro**.

Quadro 35: Exemplo de regra de alinhamento (3)

No exemplo do Quadro 35, as duas sentenças são alinhadas por conterem uma informação secundária em comum, sendo esta o fato de que os planemos giram um ao redor do outro.

Regra 4
Alinhar todas as sobreposições de um mesmo conteúdo

Exemplo
Sentença do sumário
Usando telescópios do Observatório Europeu Sul (ESO), Ray Jayawardhana, da Universidade de Toronto, e Valentin D. Ivanov, do ESO, descobriram um planemo com sete vezes a massa de Júpiter, o planeta mais pesado do Sistema Solar, e outro com o dobro desse peso, que giram um ao redor do outro, denominado Oph 162225-240515, o primeiro planemo duplo.
Sentenças do documento
Astrônomos do Observatório Europeu Austral, localizado no Chile, anunciaram a descoberta de uma dupla de planetas errantes (sem estrela-mãe) que giram ao redor deles mesmos e que vagam livremente pelo espaço.
O fato extraordinário é que ele não gira em volta de uma estrela,mas em torno de outro corpo frio com o dobro de sua massa.

Quadro 36: Exemplo de regra de alinhamento (4)

Como pode ser visto no exemplo do Quadro 36, todas as sentenças do documento foram alinhadas a do sumário por conterem informação em comum. No caso, as duas sentenças são provenientes de um mesmo documento, mas é interessante lembrar que elas podem ser provenientes de documentos diferentes.

Regra 5	
Alinhar com base na sobreposição da informação principal mesmo diante de dado numérico contraditório.	
Exemplo	
Sentença do sumário	**Sentença do documento**
Às **9h**, a cidade tinha oito pontos de alagamento, sendo dois intransitáveis.	O CGE (Centro de Gerenciamento de Emergências) da Prefeitura de São Paulo registrava oito pontos de alagamento na cidade, às **9h30** desta segunda-feira.

Quadro 37: Exemplo de regra de alinhamento (5)

No Quadro 37, as duas sentenças foram alinhadas por que expressam a mesma informação, apesar de conterem com uma contradição numérica. Contradições desse tipo podem ser causadas pela atualização de uma notícia, por exemplo, quando a contagem do número de feridos de um desastre aumenta.

Regra 6	
Alinhar com base na sobreposição da informação principal mesmo diante de diferentes graus de generalização	
Exemplo	
Sentença do sumário	**Sentença do documento**
A Companhia de Engenharia de Tráfego (CET) anunciou que o índice de congestionamento era de **54 quilômetros** às 8h, **113 km** às 9h e **110 km** meia hora depois, valores bem acima das médias para os horários, que eram de **36, 82 e 76 quilômetros** respectivamente, mas não havia registro de acidentes graves, apesar de haver feridos.	Com o asfalto molhado, o trânsito ficou mais lento e **o congestionamento ficou o dobro da média**.

Quadro 38: Exemplo de regra de alinhamento (6)

Como pode ser visto no exemplo do Quadro 38, as duas sentenças são alinhadas por conterem a mesma informação, ou seja, o "índice de congestionamento acima da média". Porém, na sentença do sumário, são especificados tanto os índices acima da média, quanto os da média, enquanto que na do documento a mesma informação é generalizada ("o congestionamento ficou o dobro da média").

Regra 7
Alinhar sentenças com sobreposição da informação principal e diferença no grau de assertividade
Exemplo

Sentença do sumário	Sentença do documento
As ações **são atribuídas** à facção criminosa Primeiro Comando da Capital (PCC), que já comandou outros ataques em duas ocasiões.	As ações criminosas **podem ter sido ordenadas** pelos líderes do Primeiro Comando da Capital (PCC), que haviam prometido retomar os ataques no Estado de São Paulo no Dia dos Pais, no próximo domingo.

Quadro 39: Exemplo de regra de alinhamento (7)

As duas sentenças no exemplo do Quadro 39 são alinhadas por conterem a mesma informação principal, ou seja, a autoria dos ataques em questão, apesar da sentença do sumário apresentar um grau maior de assertividade em relação à sentença do documento (que pode ser verificado pelo verbo "poder" na sentença do documento).

Regra 8	
Não alinhar sentenças com sobreposição da informação principal quando uma expressar um todo e a outra uma parte do todo.	
Exemplo	
Sentença do sumário	**Sentença do documento**
Somente neste ano, o senador se internou por três vezes no Incor.	Em abril, o senador foi internado no InCor com insuficiência cardíaca.

Quadro 40: Exemplo de regra de alinhamento (8)

Por fim, no Quadro 40, as duas sentenças não devem ser alinhadas apesar de seu conteúdo ser parecido, pois a sentença do sumário indica uma repetição do fato ocorrido (uma sequência de internações no Incor) enquanto que a sentença do documento descreve uma das ocorrências (uma internação no Incor em abril).

Anotação e resultados

A anotação do alinhamento do córpus CSTNews durou aproximadamente dois meses e foi realizada por duas linguístas computacionais com sessões de uma a duas horas por dia. Inicialmente, alguns textos foram anotados em conjunto como teste, para depois serem alinhados individualmente, sempre tirando possíveis dúvidas, e criando regras, em conjunto. Cinco grupos de texto foram separados e anotados individualmente para computar a concordância kappa, que atingiu um valor de 0,831. Sendo a kappa compreendida entre 0 e 1, um valor alto como o obtido sugere que, apesar da tarefa de alinhamento ser subjetiva, ela era bastante clara para as anotadoras.

Como informado anteriormente, os sumários possuem, em conjunto, 331 sentenças e os textos fonte 2067, sendo a média de sentenças de 6,62 e 41,34, respectivamente. Como resultado, 877 sentenças foram alinhadas, resultando em 1011 alinhamentos, sendo que a maioria das sentenças do sumário foi alinhada a duas sentenças dos textos fonte. Todos os tipos de alinhamento podem ser vistos na Tabela 10.

Tabela 10: Tipos de alinhamento

	Tipos de alinhamento												
	1-0	**1-1**	**1-2**	**1-3**	**1-4**	**1-5**	**1-6**	**1-7**	**1-8**	**1-9**	**1-10**	**1-11**	**1-12**
Quantidade	2	71	90	67	36	37	13	5	5	1	1	2	1

Como pode ser visto na Tabela 10, a maioria das sentenças alinhadas foi do tipo 1-2, tendo 90 ocorrências. Em seguida, várias sentenças foram do tipo 1-1 (71), do tipo 1-3 (67), e assim por diante.

Pode-se notar que houve casos extremos, como sentenças não alinhadas (tipo 1-0, com 2 ocorrências) e sentenças que alinharam com 12 sentenças (tipo 1-12, com 1 ocorrência). Um exemplo de sentença não alinhada (1-0) foi: "*Neste domingo, o esporte brasileiro alegrou a torcida verde-amarela*", por se tratar de uma informação inferida pelo criador do sumário. Um exemplo do caso 1-12 pode ser visto no Quadro 41, a seguir, em que cada sentença dos documentos foi alinhada à sentença do sumário em questão, pois todas elas se tratam, por inferência do anotador, de "belas atuações dos craques Ronaldinho e Kaká".

Sentença do sumário	Sentenças dos documentos
O jogo contou com belas atuações de craques como Ronaldinho e Kaká.	Aos 27min, Kaká arrancou e chutou de fora da área.
	Aos 32min, Kaká tentou de novo.
	De fora da área, ele chutou.
	Desta vez, a bola não desviou em ninguém e entrou no ângulo.
	Aos 26 minutos, a torcida xingava e pedia Obina na seleção, quando Kaká chutou forte de longe e Ronaldinho Gaúcho deu uma leve desviada na bola, enganando o goleiro equatoriano.
	Kaká acertou um belíssimo chute de longe no ângulo aos 31 e fez 3 a 0.
	Na volta da Seleção Brasileira ao Maracanã, os jogadores não decepcionaram e o Brasil goleou o Equador por 5 a 0, com direito a golaço, jogada bonita, show de dribles e frango do goleiro adversário.
	A Seleção voltou para a segunda etapa com vontade de abrir o placar e logo aos três minutos Kaká soltou uma bomba e o goleiro equatoriano se atrapalhou todo para defender.
	Kaká fez excelente jogada na direita e virou o jogo para Robinho na esquerda
	Aos 27, Kaká arriscou de muito longe e Ronaldinho colocou o desviou o chute.
	Cinco minutos depois, aos 31, Kaká fez o gol mais bonito da partida.
	Ele chutou de fora da área, colocado, e acertou o ângulo.

Quadro 41: Exemplo alinhamento (1-12)

Um exemplo de alinhamento de uma coleção inteira de textos do CSTNews pode ser visto na Página 8 (Figura 1), na Introdução desta dissertação.

Disponibilização

Todo o alinhamento foi disponibilizado em XML (*Extensible Markup Language*) juntamente com outras anotações contidas no córpus CSTNews. Um exemplo de como o alinhamento foi demarcado em XML pode ser visto no Quadro 42. Nele é possível observar que cada bloco representa uma sentença do sumário, e que cada linha do bloco representa uma sentença do documento que foi alinhada à do sumário. Para cada alinhamento, é anotado o nome do documento cuja sentença faz parte ("DOC"), o número daquela sentença no documento ("SENT"), o tipo do alinhamento ("TYPE"), e o nome do juiz ("JUDGE"). O campo "TYPE" será explicado com mais detalhes na Página 74, ainda nesta seção.


```
<align SENT="1">
    <DOC="D1_C31_Folha.txt.seg" SENT="1" TYPE="none" JUDGE="veronica"/>
    <DOC="D2_C31_Estadao.txt.seg" SENT="1" TYPE="none" JUDGE="veronica"/>
    <DOC="D2_C31_Estadao.txt.seg" SENT="2" TYPE="none" JUDGE="veronica"/>
</align>
<align SENT="2">
    <DOC="D1_C31_Folha.txt.seg" SENT="2" TYPE="none" JUDGE="veronica"/>
    <DOC="D2_C31_Estadao.txt.seg" SENT="3" TYPE="none" JUDGE="veronica"/>
    <DOC="D2_C31_Estadao.txt.seg" SENT="6" TYPE="none" JUDGE="veronica"/>
</align>
<align SENT="3">
    <DOC="D1_C31_Folha.txt.seg" SENT="3" TYPE="none" JUDGE="veronica"/>
    <DOC="D2_C31_Estadao.txt.seg" SENT="3" TYPE="none" JUDGE="veronica"/>
</align>
```

Quadro 42: Exemplo da representação XML do alinhamento

Considerações finais

Em alguns tipos de textos, eram mais difíceis os julgamentos do alinhamento. Destacam-se os textos que continham assuntos políticos e alguns textos esportivos, por possuírem informações de regras dos jogos que eram desconhecidas para as anotadoras, sendo necessário buscar informações na WEB e, até mesmo, em um dos casos, assistir vídeos sobre o assunto da notícia. Um exemplo de um caso mais difícil de ser julgado, por esse motivo, pode ser visto no Quadro 43.

Sentença do sumário	Sentença do documento

Fabiana conseguiu o ouro em três tentativas.	Fabiana Murer, no entanto, não parecia incomodada, mas errou sua primeira tentativa de alcançar 4,50 m; o erro não se repetiu e a brasileira chegou a tal marca na segunda tentativa, quebrando o recorde pan-americano, garantindo o ouro.

Quadro 43: Exemplo de dificuldade encontrada na tarefa do alinhamento

À primeira vista, as sentenças do Quadro 43 parecem conter uma contradição numérica, ou mesmo não retratarem o mesmo fato. Porém, com uma busca mais aprofundada sobre o assunto específico dos textos, chegou-se a conclusão que a "tentativa" contida na sentença do sumário se referia à "terceira tentativa" do atleta naquela prova, em um cômputo geral, e na sentença do documento o autor se referia à "segunda tentativa" que o atleta possui naquela marca específica de 4,50 metros.

Anotação do tipo do alinhamento

Além do alinhamento manual aqui apresentado, todos os alinhamentos foram posteriormente julgados em uma nova anotação para definir seus tipos, registrados no campo "TYPE" do XML da anotação do alinhamento. Essa nova tarefa foi chamada de **tipificação dos alinhamentos** (Camargo *et al*., 2013). Os tipos dos alinhamentos podem ser utilizados, entre outras funções, para se avaliar de que forma pessoas sumarizam textos, ou seja, quais transformações são realizadas por elas, e também podem indicar os alinhamentos mais fáceis ou difíceis de serem avaliados manualmente e automaticamente. Essa segunda anotação durou também aproximadamente 2 meses com sessões de aproximadamente 1 hora por dia. Dessa vez, os pares de sentença eram julgados em duas categorias: **forma** e **conteúdo**. A forma indica o quão parecidas as sentenças eram como um todo, levando em conta principalmente as palavras que as compunham, e o conteúdo demonstra quais tipos de alinhamento estão contidos naquele par, sendo a avaliação realizada por segmentos sentenciais. Além disso, os alinhamentos também eram caracterizados como contendo elementos onomásticos, no caso, nomes próprios de lugares ou pessoas. Os tipos possíveis de forma e conteúdo podem ser vistos no Quadro 44.

Forma	Conteúdo	Onomástica
Idêntico Parcial Diferente	Especificação Generalização Contradição Inferência Neutro Outro	Toponímia Antroponímia

Quadro 44: Tipos da tipificação

Os tipos possíveis que um alinhamento pode ser julgado, em relação à forma, são: "idêntico", quando as duas sentenças possuem todas as palavras em comum, ou seja, são idênticas; "parcial", quando as duas sentenças são similares, ou sejam, possuem algumas ou várias palavras em comum; e "diferente", quando as duas sentenças possuem poucas ou nenhuma palavra em comum. Os tipos possíveis, em relação ao conteúdo, são: "especificação", que indica a presença de uma informação mais específica na sentença do sumário em relação à sentença do documento; "generalização", que indica a presença de uma informação mais genérica na sentença do sumário em relação à sentença do documento; "contradição", quando há uma informação contraditória entre as duas sentenças; "inferência", quando há uma informação em uma das sentenças que foi inferida pelo autor do texto; "neutro", quando há alguma informação que não resulta de alguma transformação nas sentenças; e "outro", que indica uma não concordância pelo anotador em relação ao alinhamento em si (o que pode ocorrer, pois a tarefa de alinhamento é subjetiva). É interessante destacar que a generalização e a especificação são duas operações de fusões comumente utilizadas por humanos para produzir um sumário multidocumento (Mani, 2001).

Alguns exemplos de pares de sentenças tipificados podem ser vistos nos Quadro 45, 46 e 47.

Sentença do sumário	Alinhamento	Sentença do documento
Mais de 300 policiais federais de vários estados participaram das buscas e prisões durante a operação.	Forma: Parcial Conteúdo: Neutro, generalização e toponímia	A PF divulgou que mais de 300 policiais federais do Amazonas, Distrito Federal, Mato Grosso, Acre e Rondônia fazem parte das investigações da "Operação Dominó".

Quadro 45: Exemplo de alinhamento com tipificação (1)

No exemplo do Quadro 45, o alinhamento recebe o tipo “parcial” por possuir algumas palavras em comum; o tipo “neutro”, pois parte do par não possuía nenhuma transformação; o tipo “generalização”, por que, na sentença do sumário, o nome de cada estado citado foi generalizado para “vários estados”; e recebe o tipo “toponímia” por conter nomes de locais.

Sentença do sumário	Alinhamento	Sentença do documento
O Brasil, mesmo sem duas de suas estrelas, bateu a Argentina, que tinha a melhor campanha do campeonato e contava com seus principais jogadores.	Forma: Diferente Conteúdo: Inferência e antroponímia	Após acompanhar o belo futebol apresentado pelos “hermanos” durante toda a Copa América, que foi conduzida pelos habilidosos pés de Riquelme e Messi, o Brasil foi a campo sem o tradicional status de favorito que o acompanha há muito.

Quadro 46: Exemplo de alinhamento com tipificação (2)

No exemplo do Quadro 46, o alinhamento recebe o tipo “diferente” por não conter quase nenhuma palavra em comum; recebe o tipo “inferência”, pois o fato de “Riquelme” e “Messi” serem os jogadores principais foi uma inferência do autor que criou o sumário; e, por fim, recebe o tipo “antroponímia” por que foram identificados nomes dos jogares em uma das sentenças.

Sentença do sumário	Alinhamento	Sentença do documento
O Brasil conseguiu um gol logo nos primeiros 4 minutos do jogo, fazendo os argentinos apertarem o ataque no jogo, restando ao Brasil os contragolpes, chegando ao segundo gol, que foi um gol contra.	Forma: Parcial Conteúdo: Neutro e especificação	É verdade que o Brasil deu sorte de conseguir um gol logo no início da partida

Quadro 47: Exemplo de alinhamento com tipificação (3)

No exemplo do Quadro 47, o par de sentenças recebe o tipo "parcial" por conterem algumas palavras em comum; recebe o tipo "neutro" por que as duas sentenças transmitem a mesma informação (o fato do Brasil ter pontuado); e recebe o tipo "especificação" por que "4 minutos do jogo" é uma especificação de "no início da partida".

Como resultado, a maioria dos alinhamentos foi classificada como do tipo "parcial" e "neutro", sendo esses 714 dentro dos 867 alinhamentos classificados como "parcial". Todos os resultados podem ser vistos na Tabela 11.

Tabela 11: Quantidade dos tipos na tarefa de tipificação

Tipos na tipificação	Quantidade	Porcentagem
Parcial	867	86%
Idêntico	58	5,7%
Diferente	82	8,1%
Neutro	949	94,2%
Generalização	82	8,1%
Especificação	48	4,7%
Contradição	37	3,6%
Inferência	33	3,2%

Outro	6	0,5%
Antroponímia	20	1,9%
Toponímia	4	0,3%

Para essa tarefa, também foi calculada a concordância kappa, tanto para forma e conteúdo separadamente, como em conjunto, utilizando 5 conjuntos de textos do córpus. Essa anotação posterior foi um pouco mais complexa de ser definida, por isso se observou uma queda na concordância em relação à da tarefa de alinhamento. Os resultados de concordância podem ser vistos na Tabela 12. Considerando apenas o tipos relativos à forma, a kappa obtida foi de 0,717; considerando apenas os relativos ao conteúdo, 0,318; e considerando todos os tipos, o valor de kappa obtido foi 0,452.

Tabela 12: Concordância kappa da tipificação

Forma	**Conteúdo**	**Forma e conteúdo**
0,717	0,318	0,452

No capítulo seguinte, são apresentados os métodos que foram desenvolvidos para realizar o alinhamento automático.

Capítulo 4. Métodos

Nesse capítulo, são apresentados os métodos propostos que foram desenvolvidos para realizar o alinhamento automático. Foram propostas três abordagens, cada qual com seu nível de conhecimento linguístico, sendo elas: (i) os métodos superficiais, (ii) a que faz uso de uma teoria discursiva, e (iii) a que utiliza aprendizado de máquina. Cada abordagem será explicada com detalhes a seguir.

4.1. Métodos Superficiais

Os métodos superficiais, como o nome sugere, são baseados em suposições que não contém muito conhecimento linguístico, e são vistos como *baselines*, a linha de base que os resultados de um método devem atingir, sendo o mínimo esperado. Foram propostos três métodos, que podem ser utilizados em conjunto ou separadamente, para realizar o alinhamento. O primeiro método proposto é o método que usa a ***Word overlap***, ou **sobreposição de palavras**, muito utilizada em várias tarefas na área do processamento de língua natural, por exemplo, na filtragem de pares de sentença do CSTParser. A *Word overlap* mede a quantidade de palavras em comum entre unidades textuais, sendo seu valor compreendido entre 0 e 1, como pode ser visto exemplificado no Quadro 48, a seguir. Quanto mais próximo o valor é de 1, mais palavras em comum o par de unidades possui.

S1	O agressor morreu, mas **ainda** **não** **foi** confirmado **se** **ele** foi baleado pela polícia **ou** **se** **cometeu** **suicídio**. (18 palavras)
S2	**Ainda** **não** **se** sabe **se** **ele** **cometeu** **suicídio** **ou** **foi** morto por policiais. (13 palavras)

Quadro 48: Exemplo de Word overlap

No exemplo, S1 possui 18 palavras ao todo, S2, 13, e entre elas existem 9 palavras em comum. Seguindo os cálculos, que podem ser vistos a seguir, o *Word overlap* entre essas duas sentenças é de 0,58.

$$\frac{palavras\ em\ comum * 2}{quantidade\ de\ palavras\ menor + quantidade\ de\ palavras\ menor} = \frac{9 * 2}{18 + 13} = 0{,}58$$

É necessário dobrar o valor de palavras em comum para garantir que o valor final fique entre 0 e 1, e não entre 0 e 0,5.

Ainda, é possível excluir as *stopwords*[23] dos cálculos, garantindo assim que não sejam contabilizadas nos cálculos artigos e preposições, entre outras palavras que são bastante recorrentes em qualquer texto em português do Brasil. Alguns exemplos de *stopwords* são os artigos "o" e "uma", e as preposições "para" e "por".

Excluindo-se as *stopwords* do par de sentenças do Quadro 48 (ou seja, as palavras "se", "ele", "ou" e "se"), o novo valor de *Word overlap* é de 0,5. Os cálculos podem ser vistos a seguir.

$$\frac{palavras\ em\ comum * 2}{quantidade\ de\ palavras\ menor + quantidade\ de\ palavras\ menor} = \frac{5 * 2}{12 + 8} = 0{,}5$$

Se duas sentenças possuem informação em comum e, portanto, devem ser alinhadas, é normal supor que possuam também palavras em comum. Mas, como foi visto no Capítulo 3, no exemplo do Quadro 30, nem todos os pares de sentença alinhados possuem palavras em comum, e por isso a *Word Overlap* é considerado um *baseline*.

O segundo método superficial a ser apresentado é a diferença de tamanho entre sentenças, nomeado de **tamanho relativo**. Essa medida, também compreendida entre 0 e 1, denota a diferença de tamanho, em caracteres, entre um par de sentenças, como pode ser visto no exemplo a seguir (Quadro 49).

[23] *Stopwords* são palavras muito frequentes que não trazem muito sentido, como artigos ou preposições.

S1	Na sexta-feira, choveu 12 centímetros em algumas regiões, e há previsão de mais tempestades hoje. (98 caracteres)
S2	Na sexta-feira, choveu muito acima do esperado e há previsão de mais tempestades hoje. (86 caracteres)

Quadro 49: Exemplo de tamanho relativo

$$\frac{quantidade\ de\ caracteres\ maior - quantidade\ de\ caracteres\ menor}{quantidade\ de\ caracteres\ maior} = \frac{98 - 86}{98}$$
$$= 0{,}12$$

Diferente da medida anterior, quanto mais próximo de 0, mais próximas em tamanho as duas sentenças são. A diferença entre o tamanho maior e o tamanho menor deve ser dividida pelo tamanho maior para que o número fique compreendido entre 0 e 1.

No alinhamento na tradução automática, como foi comentado na Seção 2.3.1, é sabido que uma sentença fonte e sua versão traduzida possuem tamanhos aproximados, e há trabalhos que utilizam essa suposição para realizar o alinhamento (por exemplo, Gale e Church, 1993). Por isso, vê-se razoável supor que sentenças que estejam falando de um mesmo fato possam também ter tamanhos mais próximos do que sentenças que contenham outro assunto completamente diferente. Por esse motivo, o método foi desenvolvido e é utilizado como *baseline*.

A terceira medida superficial é a **distância relativa** entre sentenças. Essa medida, cujo valor é compreendido entre 0 e 1, mede o quão distante duas sentenças estão uma da outra em relação aos seus textos de origem. Por exemplo, a primeira sentença de um documento e a primeira sentença de um sumário deverão ter distância 0. A seguir, no Quadro 50, pode ser visto um exemplo dessa medida.

	Sentenças do sumário	Sentenças do documento	
S1	A equipe de revezamento 4x200 metros livre conquistou nesta terça-feira a segunda medalha de	A equipe formada por Thiago Pereira, Rodrigo Castro, Lucas Salatta e Nicolas Oliveira deu ao	S1

	ouro da natação brasileira nos Jogos Pan-Americanos do Rio.	Brasil nesta terça-feira o terceiro ouro nos Jogos Pan-Americanos do Rio-2007, na prova do revezamento 4 x 200 m livre da natação.	
S2	**Eles fizeram história ao cravar o tempo de 7min12s27 e superar os Estados Unidos - foi a primeira derrota norte-americana na competição.**	Jamais na história do Pan os Estados Unidos haviam perdido o ouro nesta modalidade.	S2
S3	Pouco antes, Thiago Pereira havia conquistado a segunda medalha de ouro brasileira no dia na final dos 400m medley, superando o norte-americano Robert Margalis e o canadense Keith Beavers.	**O quarteto brasileiro terminou a prova com o tempo de 7min12s27, recorde pan-americano, à frente dos EUA, segundo lugar, e do Canadá, que ficou com o bronze.**	S3
		O grande destaque da prova foi Nicolas Oliveira, quarto nadador brasileiro a cair na água.	S4
		Pouco antes Thiago Pereira já havia conquistado a segunda medalha de ouro brasileira no dia na final dos 400 m medley, superando o norte-americano Robert Margalis e o canadense Keith Beavers.	S5
		Também nesta terça, outras duas medalhas também foram conquistadas por brasileiros, ambas no remo.	S6
		Marcelus Marcili ficou com o	S7

		bronze no skiff simples, e a dupla formada por Anderson Nocetti e Alan Bitencourt também ficou em terceiro na final da categoria dois sem timoneiro.	

Quadro 50: Exemplo de posição relativa

Considerando as sentenças em negrito, cujas posições são 2, para a sentença do sumário, e 3, para a sentença do documento, os cálculos são feitos como se segue.

Primeiro é necessário descobrir uma faixa para se saber o quanto um texto é maior que o outro. No exemplo, cada sentença do sumário equivale a 2,33 sentenças do documento, valor que deve ser aproximado ao inteiro mais próximo, portanto 2. A faixa serve para descobrir se as posições das sentenças nos textos são equivalentes, da forma que se duas sentenças estiverem dentro da mesma faixa, elas terão posições equivalentes.

$$faixa = \frac{número\ de\ sentenças\ do\ texto\ maior}{número\ de\ sentenças\ do\ texto\ menor} = \frac{7}{3} = 2{,}33 = 2$$

Sendo assim, para as posições 2 e 3, no sumário e no documento, respectivamente, os cálculos são feitos como se segue.

$$\frac{posição\ menor - \lceil \frac{posição\ maior}{faixa} \rceil}{número\ de\ faixas - 1} = \frac{2 - \lceil \frac{3}{2} \rceil}{3 - 1} = \frac{2 - 2}{2} = 0$$

Por convenção, a divisão entre a posição maior e a faixa deve ser aproximada para cima, ou seja, deve-se utilizar o teto da divisão. Dessa forma, o par de sentenças do exemplo recebe o valor 0 de posição relativa, que indica que as duas sentenças estão na mesma faixa. Quanto mais próximo de 0, mais próximas as sentenças estarão em seus textos de origem (sumário e documento).

Essa medida foi pensada levando em conta a escrita de sumários em relação a seus textos de origem. Pressupõe-se que a ordem dos fatos contados em uma notícia será mantida em uma versão condensada da mesma.

Os métodos superficiais, como já foi dito, são considerados uma linha de base, utilizados como o valor mínimo que os resultados de uma tarefa devem alcançar. São suposições mais fáceis de serem desenvolvidas e, portanto, podem ser muito úteis. O conhecimento linguístico inerentes às suposições que deram origem aos métodos superficiais é menor do que o das outras abordagens. Além disso, um trabalho da literatura foi escolhido para ser desenvolvido, o de Jing e McKeown (1999), que foi anteriormente apresentado na Seção 2.3.2. Uma das abordagens que contém mais conhecimento linguístico será apresentada a seguir.

4.2. Método da Teoria Discursiva

Como citado anteriormente, a CST é uma teoria que denota relações discursivas entre textos. As relações podem ser de redundância, sobreposição de parte do conteúdo, entre outras. A ideia dessa segunda abordagem é explorar o uso da CST para realizar o alinhamento entre um par de sentenças.

Como apresentado na Seção 3.1, o córpus CSTNews possui uma anotação utilizando essa teoria, porém essa anotação foi realizada apenas entre os textos fonte do córpus. Por esse motivo, vê-se necessário o uso do CSTParser para encontrar as relações CST entre os sumários multidocumento e seus textos de origem.

A ideia da segunda abordagem de alinhamento pode ser vista exemplificada a seguir na Figura 15 e no Quadro 51. A Figura 15 compreende um cenário hipotético em que é possível observar que há uma relação CST entre a primeira sentença do sumário e a segunda sentença do documento 1, duas relações entre a terceira sentença do sumário e a quinta sentença do documento 1 e a primeira sentença do documento 2, e uma relação entre a quinta sentença do sumário e a quarta sentença do documento 2. Sendo assim, de acordo com a segunda abordagem, esses também seriam os alinhamentos entre este sumário multidocumento e seus dois textos de origem.

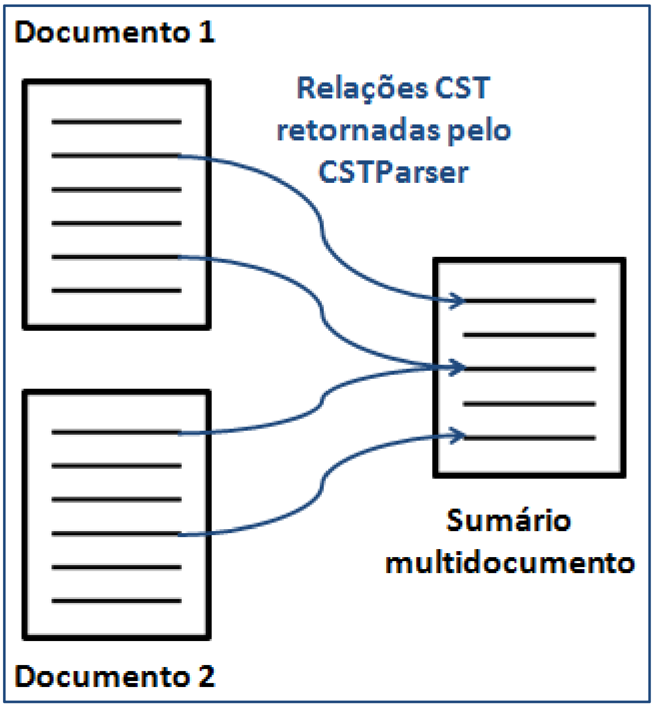


Figura 15: Abordagem CST - esquema

De fato, se for observado o exemplo do Quadro 51, é possível notar que a ideia da abordagem é justificada (especialmente para os casos de relações CST do tipo redundância, total ou parcial, vistos na tipologia na Figura 4, na Página 21). No quadro, as duas sentenças possuem uma relação do tipo *Overlap*, ou seja, as duas sentenças possuem alguma informação em comum (e também existe alguma informação que não é comum às duas sentenças), que é justamente o que o alinhamento denota.

Sentença do Sumário	Relação CST	Sentença do Documento
A outra brasileira, Joana Costa, ficou na quinta posição, com 4m20, mostrando que o nervosismo pode atrapalhar as competições em casa.	*Overlap* ⟺	Já a outra brasileira que participou da prova, Joana Costa, não subiu ao pódio, uma vez que não alcançou a marca da cubana.

Quadro 51: Exemplo da abordagem com CST – relação (sentenças retiradas do Cluster 24 do CSTNews)

Portanto, se o CSTParser julgar que um par de sentenças contém alguma relação (no caso do exemplo, uma do tipo *Overlap*) as duas sentenças podem ser consideradas alinhadas. Como já foi dito, é claro perceber que, se duas sentenças contiverem uma relação do tipo redundância (*Overlap*, entre outras), as duas sentenças provavelmente seriam alinhadas. É interessante avaliar também quais outros tipos de relações podem contribuir para o alinhamento.

A desvantagem dessa abordagem é o fato dos possíveis erros do parser serem levados até o alinhamento, porém contém mais conhecimento linguístico do que os métodos superficiais, o que, teoricamente, implica em um resultado melhor.

4.3. Método do Aprendizado de Máquina

O **aprendizado de máquina** é uma subárea da Inteligência Artificial, e uma técnica bastante poderosa capaz de aprender muito a partir de um conjunto de dados, o que apenas um olhar humano não seria capaz, muitas vezes. No caso do alinhamento, a partir de exemplos de pares de sentenças alinhadas e não alinhadas, tendo alguns atributos que os definam, é possível aprender quais características são importantes para realizar o alinhamento, criando assim um modelo para decidir sobre o alinhamento ou não de novos pares de sentenças. A tabela de exemplos utilizada no aprendizado de máquina, chamada de **tabela atributo valor**, é criada a partir de todos os pares de alinhamento possíveis do CSTNews. Cada par possível, chamados de **instância**, é caracterizado com alguns **atributos**, e, para cada atributo é atribuído um valor. A classe dos pares, ou seja, "sim", para um alinhamento, e "não", nos casos em que o par não contém um alinhamento, é dada utilizando o alinhamento manual que foi apresentado na Seção 3.3. Um resumo do que é uma tabela atributo valor pode ser visto no Quadro 52.

Instâncias	Atributo 1	Atributo 2	Atributo 3	Classe
D1S1_D2S1	Valor 1	Valor 4	Valor 7	Classe 1
D1S1_D2S2	Valor 2	Valor 5	Valor 8	Classe 1
D1S1_D2S3	Valor 3	Valor 6	Valor 9	Classe 2

Quadro 52: Esquema de uma tabela atributo valor

No caso da tarefa do alinhamento, a classe pode receber apenas dois valores, “sim” ou “não”.

Para os atributos dos pares, foram utilizados os valores das medidas dos métodos superficiais e também informações relativas a CST, sendo esta terceira abordagem, portanto, uma abordagem híbrida. Os atributos escolhidos podem ser vistos na Tabela 13.

Tabela 13: Atributos do Aprendizado de Máquina

Atributo	Tipo	Valor
Word overlap	Numérico	Entre 0 e 1
Tamanho relativo	Numérico	Entre 0 e 1
Posição relativa	Numérico	Entre 0 e 1
Quantidade de relações CST	Numérico	0 ou mais
Tipo de relação CST	Simbólico	“redundância”, “complemento”, etc.

Por fim, um exemplo da tabela atributo valor utilizada para a classificação do aprendizado de máquina pode ser visto na Figura 16. A tabela é armazenada no formato CSV (*Comma-Separated Values*), em que cada valor dos atributos é separado por vírgula.

```
0.0, 0, NOT_RELATION, 0.4, 0.30327868852459017, nao
0.23076923076923078, 0, NOT_RELATION, 0.2, 0.4325581395348837, nao
0.29411764705882354, 0, REDUNDANCIA, 0.2, 0.05699481865284974, nao
0.11764705882352941, 0, NOT_RELATION, 0.4, 0.17616580310880828, nao
0.058823529411764705, 0, NOT_RELATION, 0.4, 0.1645021645021645, nao
0.35294117647058826, 0, REDUNDANCIA, 0.6, 0.20207253886010362, nao
0.9411764705882353, 0, REDUNDANCIA, 0.0, 0.19246861924686193, sim
0.058823529411764705, 0, NOT_RELATION, 0.0, 0.5181347150259067, nao
0.0, 0, NOT_RELATION, 0.16666666666666666, 0.5407725321888412, nao
0.058823529411764705, 0, NOT_RELATION, 0.6, 0.025906735751295335, nao
0.058823529411764705, 0, NOT_RELATION, 0.8, 0.45077720207253885, nao
```

Figura 16: Tabela atributo valor

Por exemplo, a primeira linha da Figura 16, representa uma instância, de acordo com o esquema do Quadro 52. Essa instância possui os valores “0”, “0”, “not_relation”, “0,4”, aproximadamente “0,3” para os atributos “*Word overlap*”, “quantidade de relações CST”, “tipo de relação CST”, “posição relativa” e “tamanho relativo”, respectivamente, e a classe “não”. A tabela atributo valor com todos os pares de sentença possíveis do córpus CSTNews é utilizada para o aprendizado.

Os resultados das três abordagens são comparados com os do alinhamento manual, tido com um *gold standard*, para assim julgar a qualidade dos métodos. A comparação dos resultados dos métodos com o alinhamento manual será vista a seguir, no Capítulo 5, juntamente com os resultados e algumas avaliações dos mesmos.

Capítulo 5. Resultados e Avaliações

Nesse capítulo, são apresentados os resultados obtidos, juntamente com algumas avaliações e os erros comuns encontrados.

5.1. Métodos Superficiais

Como foi visto no Capítulo 4, os primeiros métodos criados para realizar o alinhamento automático entre duas sentenças são considerado os *baselines* do trabalho, pois vem de algumas suposições bastante simples, sendo elas: (i) duas sentenças podem ser alinhadas se possuírem certa quantidade de palavras em comum, (ii) duas sentenças podem ser alinhadas se ocuparem uma posição próxima em seus textos de origem, e (iii) duas sentenças podem ser alinhadas se possuírem um tamanho aproximado. Os parâmetros desses três métodos podem variar, sendo eles encontrados empiricamente, ou seja, vários parâmetros são testados até que seja obtido o maior valor possível de medida-f. Tais suposições foram exploradas em conjunto e separadamente.

Separadamente, os melhores resultados de medida-F, juntamente com suas respectivas precisões e coberturas, obtidos pelos três métodos podem ser vistos na Tabela 14, a seguir, obtidos com uma comparação com o alinhamento manual (considerado o *gold standard*). Como é possível notar, o valor que mede a quantidade de palavras em comum entre as sentenças (*Word overlap*) obteve significantemente os melhores resultados.

Tabela 14: Resultados dos métodos superficiais isolados

Medidas	Precisão	Cobertura	Medida-F
Word overlap	71,88%	61,38%	66,22%
Tamanho relativo	10,15%	63,33%	17,57%
Distância relativa	12,72%	68,07%	21,43%

Para atingir esses valores, as medidas *Word overlap*, tamanho relativo e posição relativa receberam, respectivamente, os parâmetros 0,295; 0,420 e 0,260 (obtidos empiricamente). Ou seja, pares de sentenças que obtinham 0,295 ou mais de *Word overlap*, eram consideradas um alinhamento, e assim por diante.

Em conjunto, o melhor resultado de medida-F obtido, junto com suas respectivas precisões e coberturas, pode ser visto na Tabela 15.

Tabela 15: Resultados dos métodos superficiais em conjunto

Medidas	Precisão	Cobertura	Medida-F
***Word overlap*, tamanho relativo e posição relativa**	71,88%	61,38%	66,22%

A ideia de utilizar as três medidas superficiais em conjunto era melhorar os resultados, porém, para obter esses resultados, as medidas *Word overlap*, tamanho relativo e posição relativa receberam, respectivamente, os valores: 0,295; 1 e 1 (obtidos empiricamente). Ou seja, apenas a medida *Word overlap* sozinha já obtinha os mesmos bons resultados. Porém, isso não significa que as medidas tamanho relativo e posição relativa não sejam úteis, por exemplo, como atributos do aprendizado de máquina.

5.2. Trabalho da Literatura

Como foi dito anteriormente, um trabalho da literatura de alinhamento na sumarização, o de Jing e Mckeown (1999), foi desenvolvido a fim de comparar os resultados utilizando os mesmos textos provenientes do CSTNews e na língua portuguesa do Brasil. Esse trabalho foi escolhido por ser bastante conhecido e também por apresentar o alinhamento final na saída de seu programa, o que poderia ser útil para comparações.

Sobre o desenvolvimento do método, não foram necessárias adaptações. Basicamente, o método é independente de língua, pois ele apenas utiliza uma lista de *stopwords*, que varia de uma língua para outra. Os parâmetros utilizados (P1 a P6), receberam, primeiramente, os mesmos valores sugeridos pelos autores, ou seja, o valor máximo atribuído é 1 e os outros são atribuídos decrescendo os valores em 0,9; 0,8, e assim por diante. Os valores foram substituídos empiricamente, da forma que os que fizeram o método atingir os melhores resultados foram: 1; 0,5; 0,6; 0,3; 0,2 e 0,1, respectivamente para os parâmetros P1 a P6.

Os resultados obtidos com essa abordagem podem ser vistos na Tabela 16.

Tabela 16: Resultados do trabalho da literatura

Precisão média	Cobertura média	Medida-F média
35,66%	80,59%	49,45%

Sem realizar um teste estatístico, é impossível afirmar que a abordagem de Jing e Mckeown (1999), quando utilizada para alinhar os textos do córpus CSTNews, obteve resultados inferiores aos dos já apresentados métodos superficiais (no caso, o método que utiliza a medida *Word overlap* para guiar o alinhamento). Especialmente por que os valores não ficaram muito distantes.

5.3. Método da Teoria Discursiva

Como explicado no Capítulo 4, a ideia da segunda abordagem proposta é utilizar as relações CST como guia para o alinhamento.

De acordo com o julgamento do *parser*, apenas foram encontradas relações dos tipos “redundância” e “complemento”, não resultando em nenhum dos tipos “contradição”, “fonte/autoria” e “estilo”. Por isso, foi assumido que todas as relações CST retornadas pelo CSTParser indicam um alinhamento entre sentenças dos sumários multidocumento e textos fonte do córpus CSTNews. Para cada conjunto de textos, o que resultou foi comparado com o alinhamento manual, tido como *gold standard*, sendo obtidos os valores da Tabela 17.

Tabela 17: Resultados da abordagem CST

Precisão	Cobertura	Medida-F
55,00%	67,83%	60,74%

Os resultados são bons, próximos aos obtidos na abordagem dos métodos superficiais. Porém, não é possível afirmar com exatidão que o método é melhor ou pior sem um teste estatístico.

Graças a esses bons resultados, é possível afirmar que o alinhamento realmente reflete os fenômenos multidocumento, pelo menos parte deles.

5.4. Método do Aprendizado de Máquina

Para o aprendizado de máquina, foi utilizada uma tabela com 15689 exemplos de alinhamento, sendo esses todos os possíveis alinhamentos de todos os conjuntos de texto do CSTNews. Como há de se imaginar, há muito mais exemplos de pares de sentença que não foram alinhados do que dos que foram alinhados, sendo eles 14678, ou seja, 93,55% do total, contra 1011, apenas 6,44% do total, respectivamente. Apesar disso, os cálculos foram realizados, primeiramente, sobre o conjunto de dados desbalanceado e, mesmo assim, bons resultados foram obtidos.

Utilizando a ferramenta WEKA[24] (Hall *et al.*, 2009), os dados foram classificados utilizando-se alguns classificadores, sendo eles: (i) J48, um exemplo de classificador por árvore de decisão, (ii) Naive Bayes, um classificador bayesiano, (iii) OneR, que utiliza regras para classificar, e (iv) SVM, um exemplo de classificador que utiliza uma função matemática para realizar a classificação. Os classificadores foram escolhidos por serem bastante utilizados e para contemplar quatro tipos diferentes de classificador. Para a classificação, foi utilizado o método de *Ten-fold cross-validation*, que consiste em realizar a classificação em 10 etapas, de forma que todo o conjunto de dados seja considerado em algum momento como treinamento e também como teste, da seguinte maneira: na primeira etapa, um décimo do conjunto de dados é separado para teste, enquanto que o restante (nove décimos) é utilizado como treinamento; na

[24] Uma ferramenta poderosa que contém vários classificadores para aplicações de aprendizado de máquina. Disponível em: http://www.cs.waikato.ac.nz/ml/weka

segunda etapa, outra parte do conjunto de dados, equivalente a um décimo do mesmo, é separado para teste, enquanto que o restante é utilizado para treinamento, e assim por diante. Os resultados mais importantes podem ser vistos na Tabela 18, a seguir (todos os resultados estão detalhadamente resumidos na Tabela 20).

Tabela 18: Principais resultados do Aprendizado de Máquina (desbalanceado)

Classificador	Instâncias classificadas corretamente	Classe "sim"		
		Precisão	Cobertura	Medida-F
OneR	96,13%	86,2%	47,6%	61,3%
J48	95,94%	78,7%	50,7%	61,7%
NaiveBayes	93,78%	51,3%	69,3%	59,0%
SVM	95,87%	88,2%	41,4%	56,4%

Nessa tabela (Tabela 18), os resultados para a classe "não" foram ignorados, mas como pode ser visto na Tabela 20, todos eles são acima de 96%. Todos os valores de medida-F, para a classe "sim", são aproximadamente 60%, que é um resultado bom.

Foram classificadas corretamente sempre 93,78% ou mais das instâncias, porém esse grande número pode mascarar a classe minoritária (ou seja, a classe "sim"), já que apenas 6,44% dos pares de sentença eram desse tipo. Um classificador que chutasse a classe majoritária iria acertar, nesse caso, em 93,55% dos casos.

O classificador do tipo regra OneR cria uma regra utilizando um dos atributos para realizar a classificação. Nesse experimento, o classificador utilizou o atributo *Word overlap*. A regra que o classificador criou pode ser vista a seguir, no Quadro 53. De acordo com a regra criada, se o *Word Overlap* possuir um valor abaixo de aproximadamente 0,202, o par de sentenças não deve ser alinhado (ou seja, recebe a classe "não"); se o atributo possuir um valor aproximadamente menor que 0,211 (e maior que 0,202), o par de sentenças deve ser alinhado (ou seja, recebe a classe "sim"); se o atributo possuir um valor menor que aproximadamente 0,274 (e maior que 0,211) o par de sentenças não deve ser alinhado; e assim por diante, em que um par que obtiver o valor de aproximadamente 0,433, ou mais, deve ser alinhado.

```
WO:
< 0.20294117647058824   -> nao
< 0.2113237639553429    -> sim
< 0.274294670846395     -> nao
< 0.2788888888888889    -> sim
< 0.36038961038961037   -> nao
< 0.3651515151515151    -> sim
< 0.37797619047619047   -> nao
< 0.396969696969697     -> sim
< 0.40454545454545454   -> nao
< 0.41801075268817206   -> sim
< 0.4330357142857143    -> nao
>= 0.4330357142857143   -> sim
```

Quadro 53: Regra criada - OneR (balanceado)

Já o classificador J48, utiliza uma árvore, chamada de Árvore de Decisão, para decidir a classe de uma instância. A árvore gerada pode ser vista a seguir, na Figura 17.

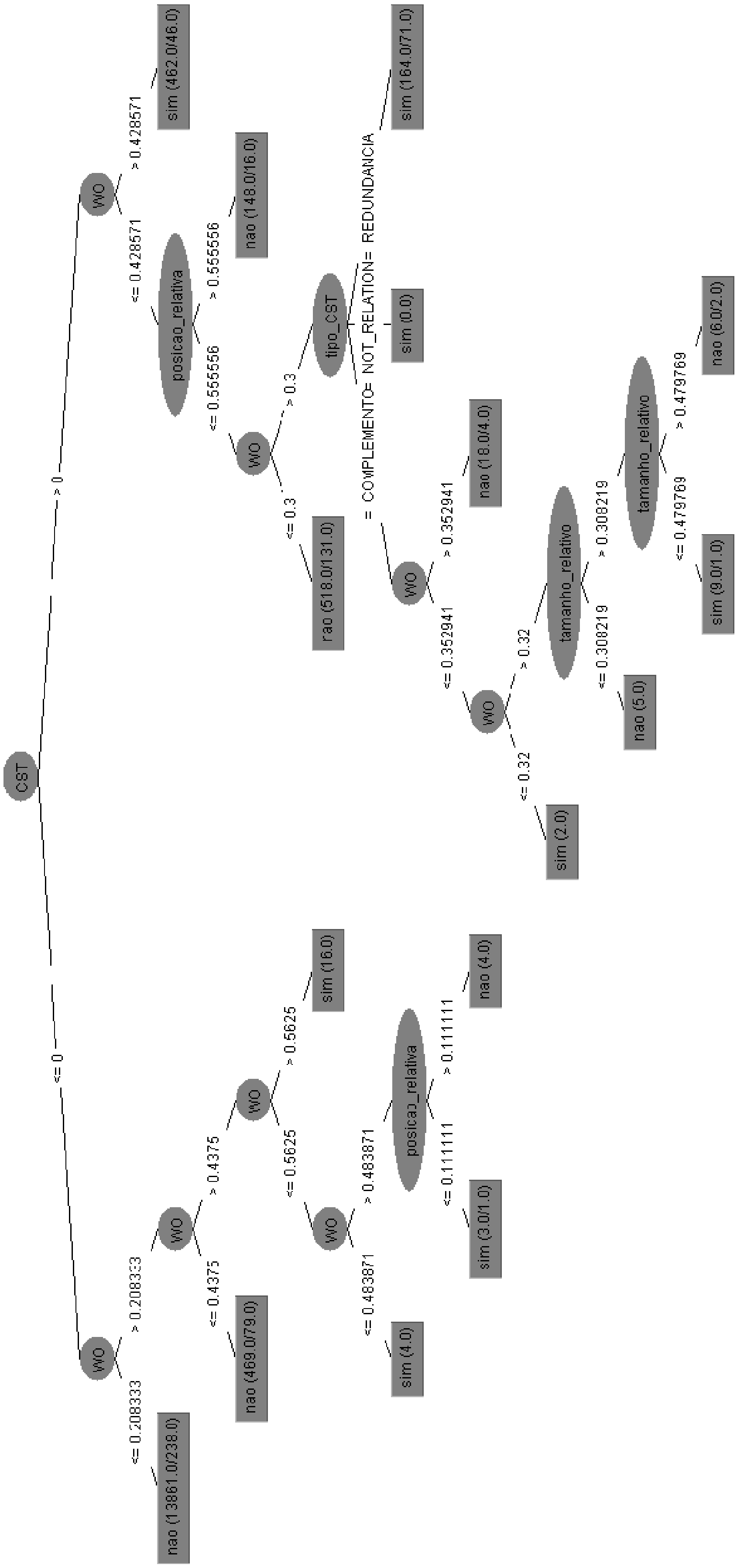


Figura 17: Árvore de decisão - J48 (desbalanceado)

Como é possível notar, o atributo que foi escolhido para ocupar a raiz da árvore é o atributo relativo a CST. De acordo com a árvore, primeiro deve ser avaliado se um par possui, ou não, uma relação CST. Se o mesmo possuir, deve-se caminhar para o lado direito da árvore, e se o mesmo não possuir, para o lado esquerdo da árvore. No caso que foi caminhado para o lado esquerdo, será avaliado o atributo *Word overlap*, em que se uma instância obtiver, para esse atributo, um valor igual ou menor a aproximadamente 0,208, ela será julgada como da classe "não" (ou seja, não ocorre um alinhamento para o par em questão). Já se o valor para o atributo *Word overlap* for maior que o valor de aproximadamente 0,208, deverá ser observado se o mesmo é maior ou não que o valor aproximado de 0,437, e assim por diante, até que todos os casos sejam cobertos pela árvore.

Como exemplo, de acordo com a árvore gerada, um par que possuir os valores 1, para o atributo CST, 0,5, de posição relativa e 0,25, de *Word overlap*, será considerado da classe "não", sem que seja preciso levar em consideração os valores dos outros atributos da instância.

5.4.1. *Balanceamento*

Ainda, foi realizada uma segunda avaliação utilizando-se uma tabela atributo valor balanceada, manualmente, resultando nos mesmos 14678 exemplos da classe "não" e em 14154 exemplos da classe "sim" (um número 14 vezes maior que o anterior). Com isso, a taxa de acerto diminuiu em dois dos três classificadores, mas a medida-F da classe "sim" aumentou consideravelmente. Os novos resultados podem ser vistos na Tabela 19, a seguir.

Tabela 19: Principais resultados do Aprendizado de Máquina (balanceado)

Classificador	Instâncias classificadas corretamente	Classe "sim"		
		Precisão	Cobertura	Medida-F
OneR	85,65%	81,0%	92,5%	86,4%
J48	97,22%	94,6%	100%	97,2%

NaiveBayes	83,15%	92,8%	71,2%	80,6%
SVM	83,71%	92,3%	72,9%	81,5%

Na tabela, destaca-se o classificador J48, que obteve um desempenho muito bom, resultando, inclusive, em uma cobertura de 100% para a classe “sim” e uma precisão de 100% para a classe “não” (como pode ser visto na Tabela 20). Apesar do ótimo desempenho obtido, os dados balanceados não correspondem a um cenário real, pois o problema do alinhamento sempre irá apresentar mais pares de sentença que não foram alinhados (classe “não”) do que pares de alinhamento que foram alinhados (classe “sim”).

5.4.2. *Seleção de Atributos*

Também foi realizado ranqueamento dos atributos, tarefa que o WEKA também é capaz de realizar. Para isso, foi utilizado o método InfoGain, que ranqueia os atributos. Como resultado, para os conjuntos de dados tanto desbalanceado quanto balanceado, foram ranqueados todos os atributos, resultando na ordenação que pode ser vista no Quadro 54.

Ordenação	**Tabela atributo valor desbalanceada**	**Tabela atributo valor balanceada**
1º	*Word overlap*	*Word overlap*
2º	Tipo da relação CST	Tipo da relação CST
3º	Quantidade de relações CST	Quantidade de relações CST
4º	Posição relativa	Tamanho relativo
5º	Tamanho relativo	Posição relativa

Quadro 54: Ranqueamento dos atributos do Aprendizado de Máquina

Como é possível notar pelo Quadro 54, o *Word overlap* foi tido como o melhor atributo nas duas seleções. Em seguida, os melhores atributos foram o relativo ao tipo da relação CST e o relativo a quantidade de relações CST, respectivamente. As posições finais se alternaram entre os atributos posição relativa e tamanho relativo.

Em um último experimento, os atributos posição relativa e tamanho relativo foram ignorados na avaliação, resultando em um decréscimo na utilização dos três classificadores. Por isso, apesar de não melhorarem os resultados nos métodos superficiais, as medidas ainda se vêem úteis, pois melhoram o resultado no aprendizado de máquina. Os resultados desse experimento podem ser visto na **Erro! Fonte de referência não encontrada.**.

Tabela 20: Resultados do Aprendizado de Máquina

Classificador	Tipo do Experimento	Instâncias corretamente classificadas	Classe "sim"			Classe "não"			Média
			Precisão	Cobertura	Medida-f	Precisão	Cobertura	Medida-f	Medida-f
OneR	Dados não balanceados	96,13%	86,2%	47,6%	61,3%	96,5%	99,5%	98,0%	95,6%
J48		95,94%	78,7%	50,7%	61,7%	96,7%	99,1%	97,9%	95,5%
Naive Bayes		93,78%	51,3%	69,3%	59,0%	97,8%	95,5%	96,6%	94,2%
SVM		95,87%	88,2%	41,4%	56,4%	96,1%	99,6%	97,8%	95,2%
OneR	Dados balanceados manualmente	85,65%	81,0%	92,5%	86,4%	91,6%	79,0%	84,9%	85,6%
J48		97,21%	94,6%	100%	97,2%	100%	94,5%	97,2%	97,2%
Naive Bayes		83,15%	92,8%	71,2%	80,6%	77,3%	94,7%	85,1%	82,9%
SVM		83,71%	92,3%	72,9%	81,5%	78,3%	94,2%	85,5%	83,5%
OneR	Seleção de atributos – não balanceado	96,13%	86,2%	47,6%	61,3%	96,5%	99,5%	98,0%	95,6%
J48		95,99%	88,0%	43,7%	58,4%	96,3%	99,6%	97,9%	95,4%
Naive Bayes		93,62%	50,4%	69,2%	58,3%	97,8%	95,3%	96,5%	94,1%
SVM		95,86%	88,2%	41,3%	56,3%	96,1%	99,6%	97,8%	95,2%
OneR	Seleção de atributos – balanceado	84,63%	89,6%	77,7%	83,2%	80,9%	91,3%	85,8%	84,6%
J48		85,57%	91,2%	78,2%	84,2%	81,5%	92,7%	86,7%	85,5%
Naive Bayes		83,16%	92,7%	71,3%	80,6%	77,4%	94,6%	85,1%	82,9%
SVM		82,89%	93,2%	70,3%	80,1%	76,9%	95,0%	85,0%	82,6%

5.5. Avaliação das Suposições dos Métodos

Utilizando o WEKA, é possível observar os erros comuns dos métodos e também avaliar as suposições feitas para a criação dos mesmos. Em relação ao método, e ao atributo, *Word overlap*, é possível verificar, na Figura 18, que a suposição de "quanto maior for o *Word overlap*, maior será a chance das sentenças serem alinhadas" é parcialmente verdadeira. De fato, há uma maior aglomeração de exemplos da classe "não" com *Word overlap* baixo (canto inferior direito da Figura 18), porém essa relação não parece se manter no caso dos exemplos da classe "sim", onde há uma grande aglomeração de exemplos com para essa classe e *Word overlap* bem baixos (canto inferior esquerdo da Figura 18).

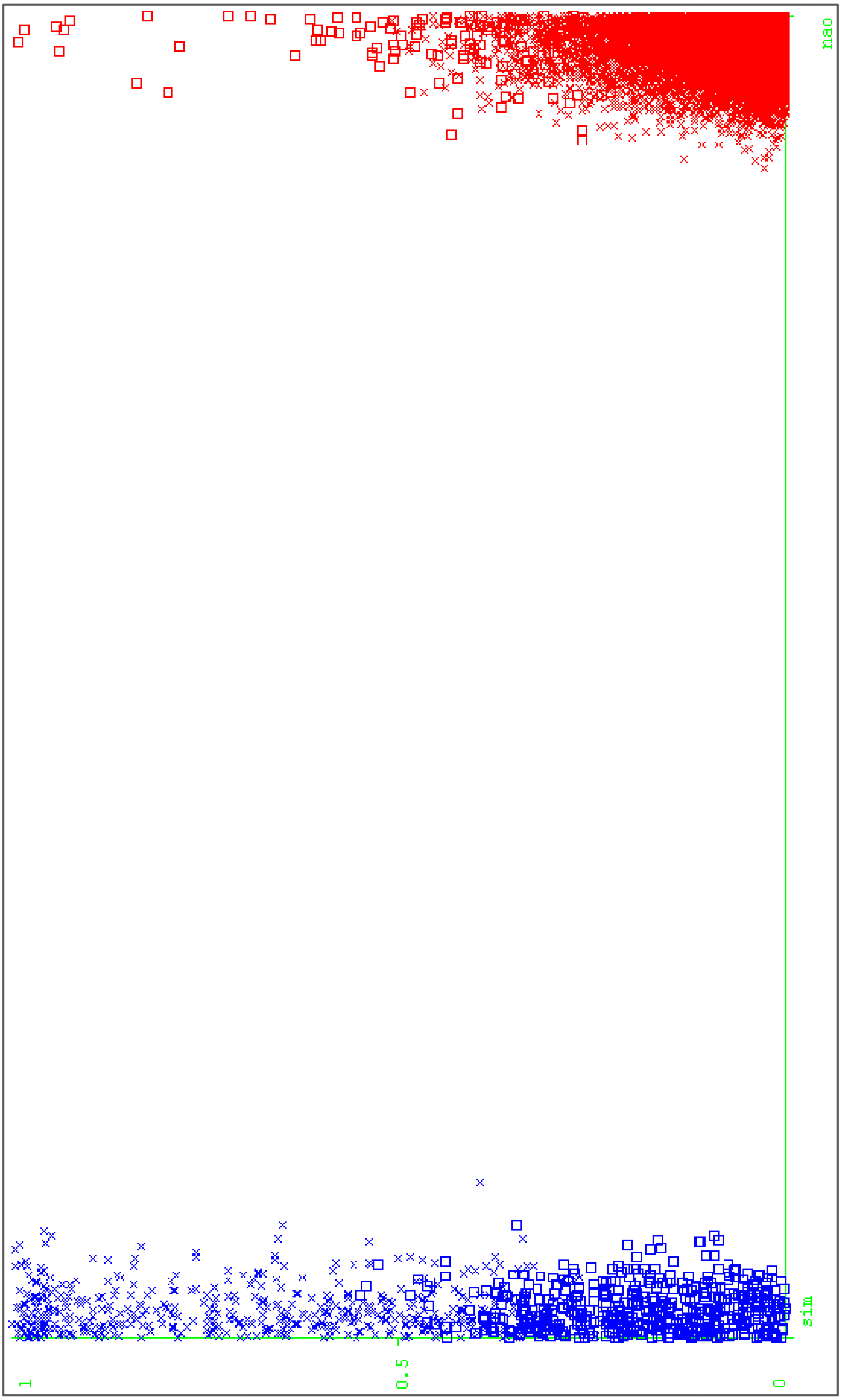


Figura 18: ***Word overlap*** **em relação à classe (J48)**

Ainda em relação a essa figura, e as próximas que se seguem (Figuras 19, 20, 21 e 22), instâncias representadas por quadrados foram erroneamente classificadas (sendo “x” a representação de instâncias classificadas corretamente). No canto superior direito da Figura 18, é possível verificar alguns exemplos com *Word overlap* igual a 1, ou seja, pares de sentenças com todas as unidades lexicais em comum, que são da classe “não” erroneamente classificados como da classe “sim” (quadrados). Nesse caso, os pares de sentença apenas eram da classe “não” graças a um erro que foi trazido do alinhamento manual. Na época, decidiu-se deixar a anotação congelada, para que novas possíveis suposições não interferissem na anotação original. É interessante lembrar que é esperado que erros desse tipo existam, devido a se tratar de uma anotação humana.

Quanto ao segundo método superficial, e atributo, considerado *baseline*, a posição relativa, pode ser visto na Figura 19 que de fato há uma relação entre seu valor e a classe “sim”, pois existe uma maior aglomeração de exemplos para essa classe com valores baixos para a medida, sendo que os exemplos diminuem conforme a posição relativa se aproxima de 1 (lado esquerdo da Figura 19). Porém, em relação à classe “não”, fica impossível afirmar que essa relação é mantida.

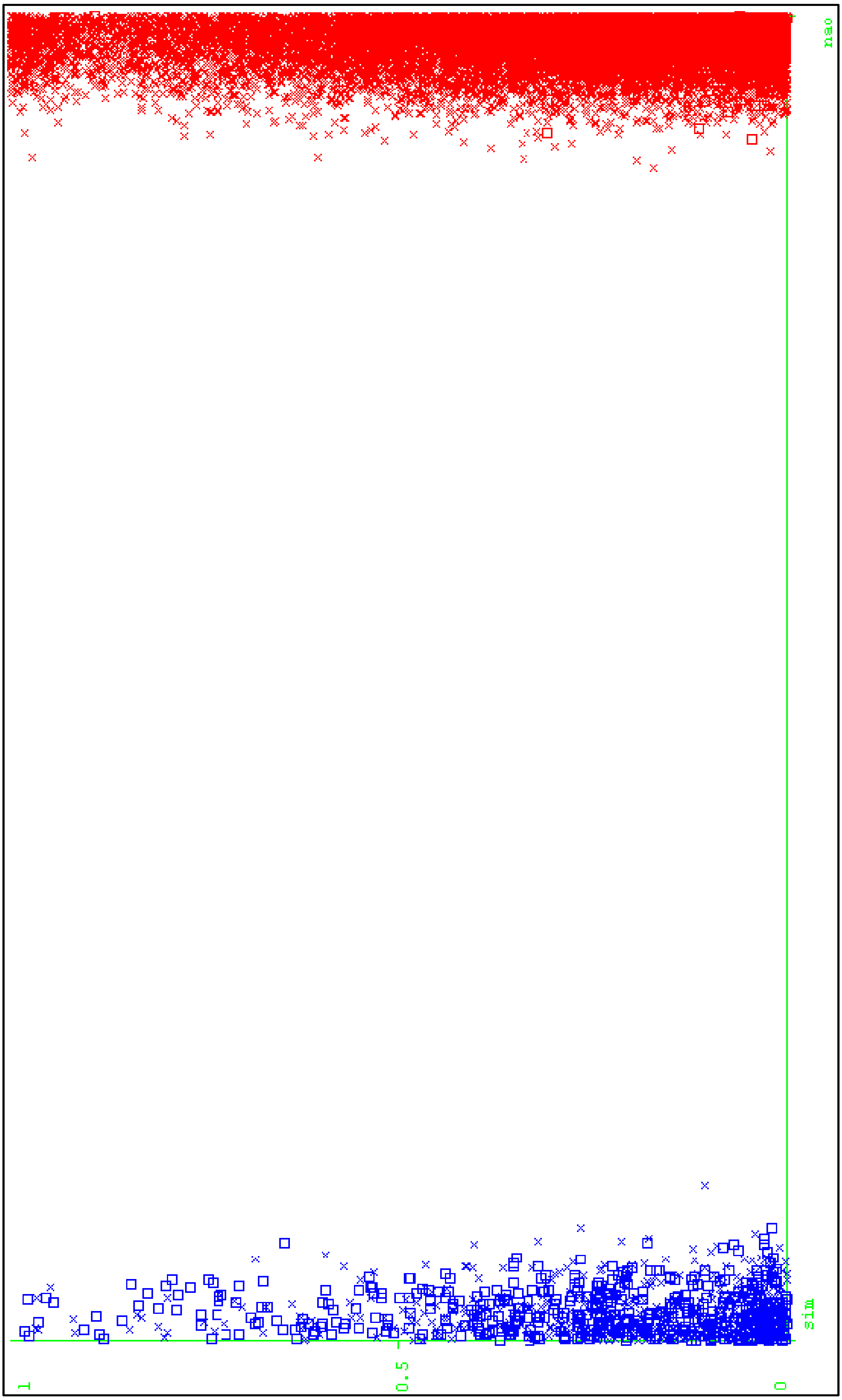


Figura 19: Posição relativa em relação à classe (J48)

Em relação ao método, e atributo, tamanho relativo, pode ser visto na Figura 20, que não há exemplos de classe “sim” quando uma sentença é o dobro da outra (ou seja, tamanho relativo igual a 1), e que, a partir de aproximadamente 0,75, os exemplos parecem diminuir. Porém, há de fato poucos casos em que o tamanho das sentenças são tão diferentes.

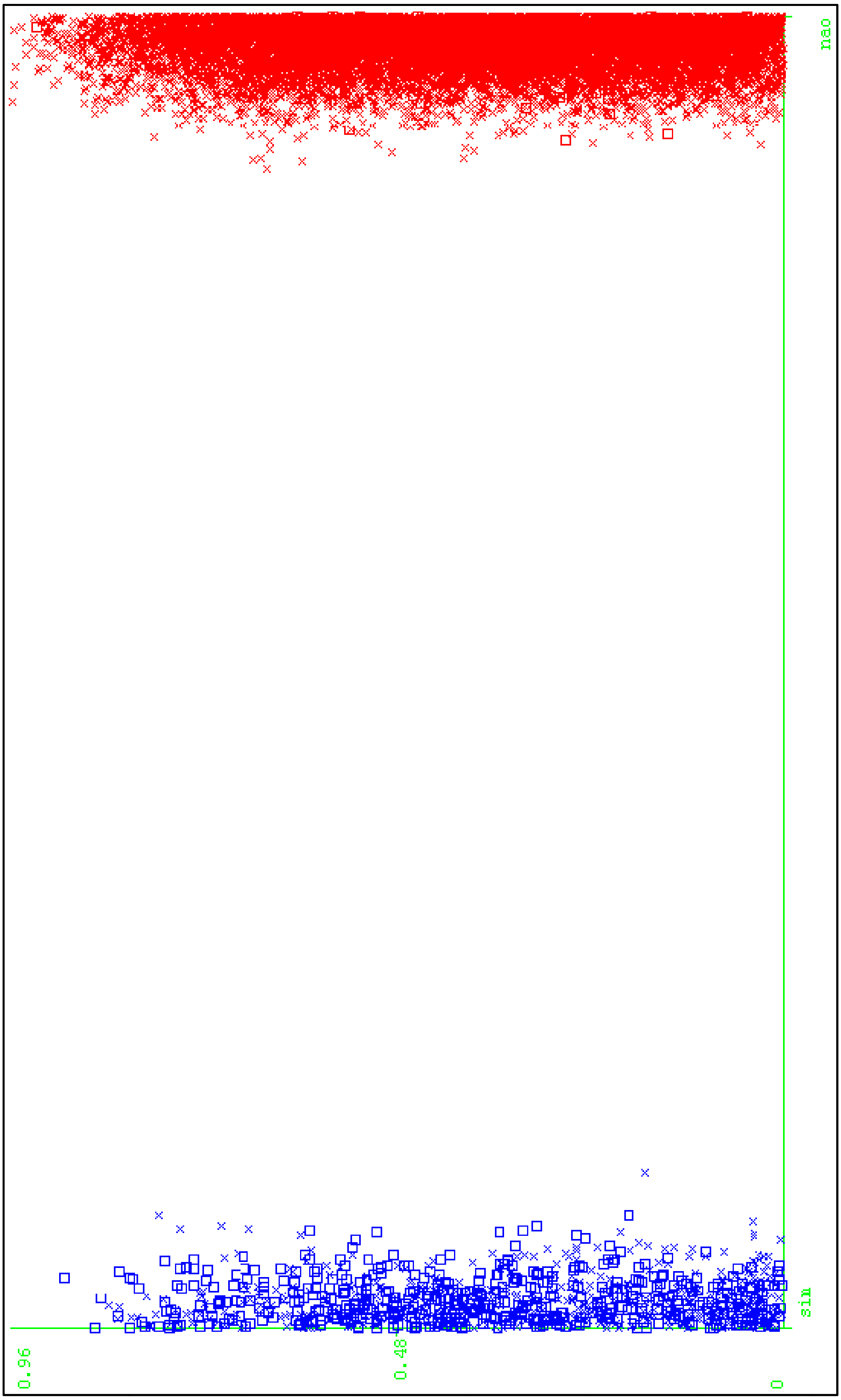


Figura 20: Tamanho relativo em relação à classe (J48)

Avaliando a correspondência entre a quantidade, juntamente com o tipo, de relações CST e a classe, pode-se observar a Figura 21. Nela, é possível notar que há muitos mais casos da classe “não” que não possuem nenhuma relação CST do que casos em que possuem alguma (lado direito da figura). Em relação à classe “sim”, é possível notar que há mais exemplos com o tipo “redundância” do que com o tipo “complemento”, como era de se esperar. No entanto, também há vários exemplos de sentenças alinhadas que não foram classificadas como contendo alguma relação CST. Lembrando que sempre pode haver erros que foram trazidos de anotações anteriores, no caso as do CSTParser.

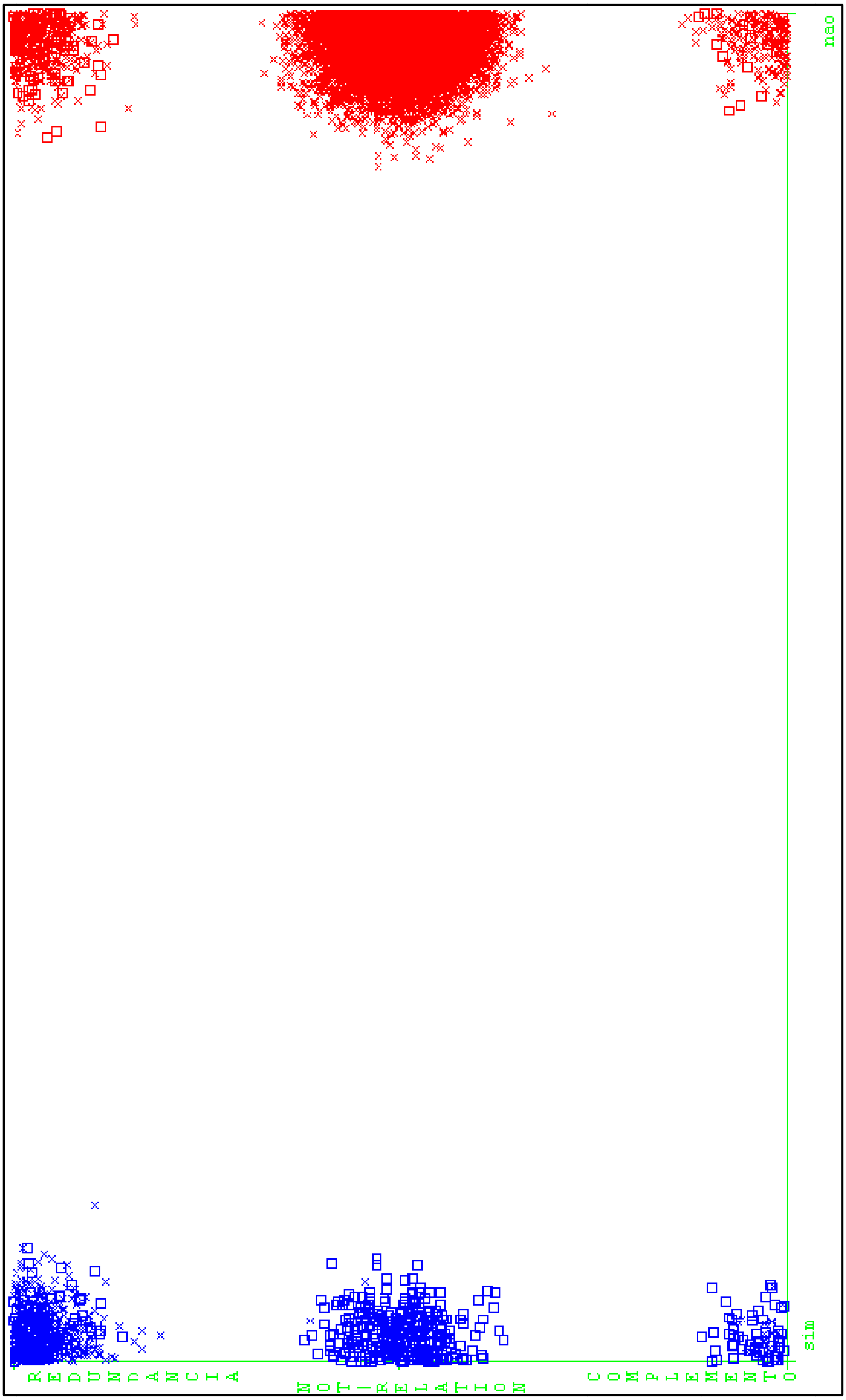


Figura 21: CST em relação à classe (J48)

Por fim, observando visualmente os erros do classificador, na Figura 22, é possível mais uma vez perceber como é muito mais fácil classificar corretamente alinhamentos do tipo “não”, dado o seu grande número de exemplos.

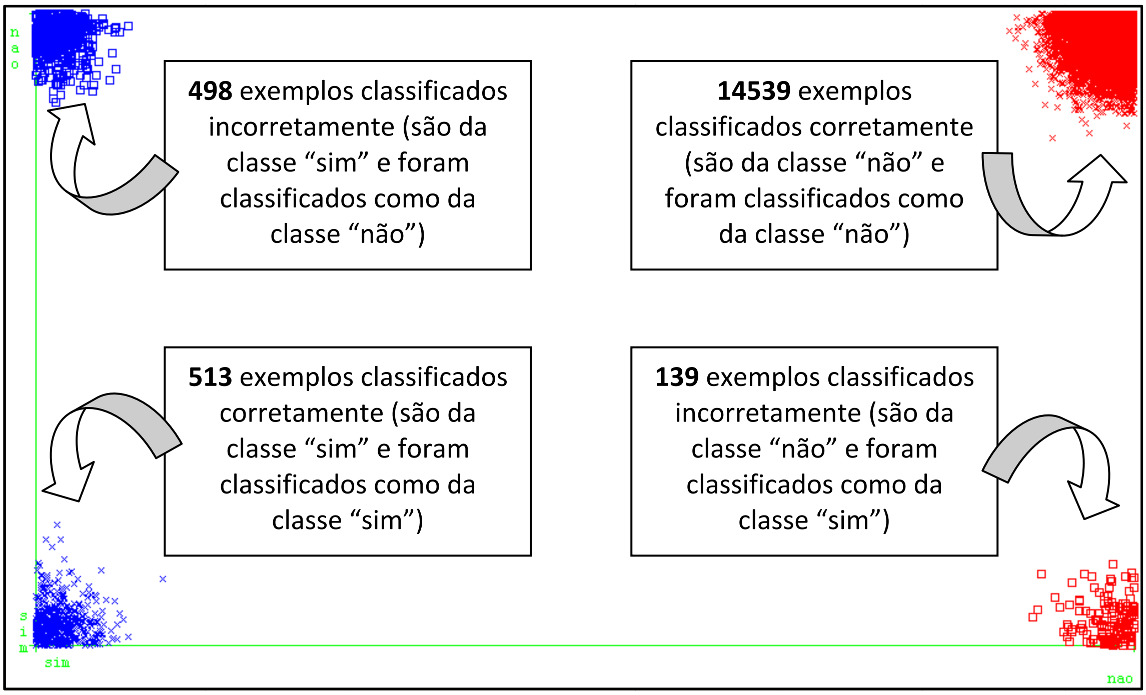


Figura 22: Erros do classificador (J48)

Quanto aos erros comuns, é esperado que vários alinhamentos do tipo de forma “diferente” sejam classificados de forma errada. Além disso, é esperado que alinhamentos do tipo de conteúdo “inferência”, e possivelmente “outros”, também não sejam capturados de forma correta pelos métodos das abordagens propostas.

Sendo assim, no capítulo seguinte, são apresentadas algumas considerações finais.

Capítulo 6. Considerações Finais

Este trabalho de mestrado é o primeiro exemplo de trabalho a realizar o alinhamento de um sumário multidocumento e seus textos de origem, de forma automática, para a língua portuguesa do Brasil, sendo, portanto, uma grande contribuição para a área do alinhamento e da sumarização automática multidocumento. Ainda não há exemplos, em português, de trabalhos que fazem uso do alinhamento manual, ou automático, dos sumários e seus textos de origem para auxiliar na sumarização automática, porém isso se mostra bastante interessante, como foi dito anteriormente. Além disso, os tipos de alinhamentos descobertos na tarefa de tipificação poderiam ser utilizados para julgar que tipo de transformações realizadas por sumarizadores profissionais são mais importantes em um sumário, mais uma vez sendo útil para a sumarização automática. A anotação manual do alinhamento e de seus tipos é uma das contribuições desse trabalho.

Em especial, a terceira abordagem, a que faz uso de aprendizado de máquina, obteve ótimos resultados quando se utilizando uma tabela atributo valor balanceada, atingindo, inclusive, os valores de 100% de cobertura (para a classe “sim”) e 100% de precisão (para a classe “não”) e uma medida-F média de 97,2%. Porém, sabe-se que o cenário real em que o alinhamento se encontra nunca será balanceado. Para o aprendizado de máquina fazendo uso de um conjunto de dados desbalanceado, que representa a realidade, os melhores resultados também foram muito bons, sendo eles 61,7% de medida-F para a classe “sim” e 97,9% de medida-F para a classe “não”, com uma média de 95,5% de medida-F. Além disso, um dos métodos superficiais, o *Word overlap*, também obteve bons resultados, em que a medida-F obteve o valor de 66,22%.

Com isso, os objetivos deste trabalho de mestrado foram completamente atingidos. Técnicas de alinhamento sentencial foram exploradas, e desenvolvidas, resultando em um alinhador automático (futuramente disponível na página do NILC[25]). Porém, não é possível concluir como verdadeira a hipótese de que abordagens que

[25] http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources

contenham mais conhecimento linguístico possuem melhores resultados do que abordagens que possuem menos conhecimento linguístico, sendo que a medida superficial que computa a quantidade de palavras em comum entre um par de segmentos (o *Word overlap*) obteve um resultado bom, superior a medida-F da classe "sim" do método do aprendizado de máquina. Foi comprovado que, em certa medida, os fenômenos multidocumento refletem o alinhamento. Esse fato foi comprovado pelos resultados obtidos com a segunda abordagem, o método que usa uma teoria discursiva, a CST, sendo a medida-F de 60,74%.

Como trabalhos futuros, destaca-se a possibilidade de utilizar os alinhamentos, e o alinhador automático desenvolvido com a junção das três abordagens aqui descritas, para auxiliar diretamente na sumarização automática, ou seja, na criação de um sumarizador baseado no alinhamento de sumários multidocumento e seus textos de origem. Também, os alinhamentos, e seus tipos, podem ser reavaliados em nível de n-gramas e esta nova avaliação ser utilizada em um sumarizador capaz de criar extratos a partir de n-gramas, e não de sentenças. Ainda, seria possível um terceiro nível, o do alinhamento entre palavras, que, muito provavelmente, seria bem mais complexo de ser desenvolvido.

Referências

Agostini, V., Camargo, R. T., Di Felippo, A., & Pardo, T. A. S. (2012). *Alinhamento manual dos sumários humanos e dos textos-fonte do corpus Multidocumento CSTNews*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, no. 380. NILC-TR-12-01. São Carlos-SP, Junho, 20p.

Agostini, V., Camargo, R. T., Di Felippo, A. & Pardo, T. A. S. (Forthcoming 2014). *Manual Alignment of News Texts and their Multi-document Human Summaries*. Cambridge: Cambridge Scholars Publishing.

Aleixo, P., & Pardo, T. A. S. (2008). *CSTTool: Uma Ferramenta Semi-automática para Anotação de Córpus pela Teoria Discursiva Multidocumento CST*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, n. 321. São Carlos/SP, 14p.

Banko, M., Mittal, V., Kantrowitz, M., & Goldstein, J. (1999). Generating Extraction-Based Summaries from Hand-Written Summaries by Aligning Text Spans. In the *Proceedings of the 4th Conference of the Pacific Association for Computational Linguistics*, 5p.

Barzilay, R. (2003). *Information Fusion for Multidocument Summarization: Paraphrasing and Generation*, Ph. D. Thesis, Columbia University, New York, 221p.

Barzilay, R., & Elhadad, N. (2003). Sentence Alignment for Monolingual Comparable Corpora. In the *Proceedings of the Empirical Methods for Natural Language*, pp. 25-32.

Barzilay, R., & McKeown, K. (2005). Sentence Fusion for Multi-document News Summarization. *Computational Linguistics*, v. 31, n. 3, pp. 297-327.

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process. *Inequalities*, v. 3, pp. 1-8.

Black, P. E. (2005). "Greedy algorithm" in *Dictionary of Algorithms and Data Structures [online]*, U.S. National Institute of Standards and Technology, February, webpage: http://xlinux.nist.gov/dads//HTML/greedyalgo.html.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., & Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, v. 16, n. 2, pp. 79-85.

Brown, P. F., Della Pietra, S. A., Dela Pietra, V. J., & Mercer, R. L. (1993). The Mathematics of Machine Translation: Parameter Estimation. *Computational Linguistics*, v. 19, n. 2, pp. 263-311.

Camargo, R. (2013). *Investigação de estratégias de sumarização humana multidocumento*. (Masters dissertation). Universidade Federal de São Carlos (UFSCar).

Camargo, R. T., Agostini, V., Di Felippo, A., & Pardo, T. A. S. (2013). Manual Typification of Source Texts and Multi-document Summaries Alignments. *Procedia – Social and Behavioral Sciences*, Vol. 95, pp. 498-506.

Cancedda, N., Gaussier, E., Goutte, C., & Renders, J-M. (2003). Word-Sequence Kernels. *Journal of Machine Learning Research*, pp. 1059-1082.

Cardoso, P. C. F., Pardo, T. A. S., & Nunes, M. G. V. (2011a). Métodos para Sumarização Automática Multidocumento Usando Modelos Semântico-Discursivos. In the *Proceedings of the 3rd RST Brazilian Meeting*, pp. 59-74. October 26, Cuiabá/MT, Brazil.

Cardoso, P. C. F., Maziero, E. G., Castro Jorge, M. L. C., Seno, E. M. R., Di Felippo, A., Rino, L. H. M., Nunes, M. G. V., & Pardo, T. A. S. (2011b). CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In the *Proceedings of the 3rd RST Brazilian Meeting*, pp. 88-105. October 26, Cuiabá/MT, Brazil.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, v. 22, n. 2, pp. 249-254.

Caseli, H. M. (2003). *Alinhamento sentencial de textos paralelos português-inglês*. (Masters dissertation). Instituto de Ciências Matemáticas e de Computação (ICMC), Universidade de São Paulo (USP), Fevereiro, 101p.

Castro Jorge, M. L. R., & Pardo, T. A. S. (2010). Experiments with CST-based Multidocument Summarization. In the *Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing*, pp. 74-82.

Castro Jorge, M. L. R., Agostini, V., & Pardo, T. A. S. (2011). Multi-document Summarization Using Complex and Rich Features. In *Anais do VIII Encontro Nacional de Inteligência Artificial*, pp. 1-12. July 19-22, Natal/RN, Brazil.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, v. 20, n. 1, pp. 37-46.

Cohen, W. (1996). Learning Trees and Rules with Set-Valued Features. In the *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-96)*. American Association for Artificial Intelligence, pp. 709-716.

Cremmins, E. T. (1996). *The Art of Abstracting*. Arlington, Virgina: Information Resources Press, 230p.

Curran, J. R. (2003). *From Distributional to Semantic Similarity*. PhD thesis, University of Edinburgh, 177p.

Dagan, I. (2000). Contextual word similarity. In Dale , R., Moisl, H., & Somers, H.L. (Eds.), *Handbook of Natural Language Processing*, 964p.

Daumé III, H., & Marcu, D. (2004). A Phrase-Based HMM Approach to Document/Abstract Alignment. In the *Empirical Methods in Natural Language Processing (EMNLP)*, 8p.

Daumé III, H., & Marcu, D. (2005). Induction of Word and Phrase Alignments for Automatic Document Summarization. *Computational Linguistics*, v. 31, n. 4, pp. 505-530.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, pp. 1-37

Fellbaum, C. (1998) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Gale, W. A., & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, v. 19, n. 1, pp. 75-102.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Hatzivassiloglou, V., Klavans, J. L., & Eskin, E. (1999). Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning**.** In the *Proceedings of the Empirical Methods for Natural Language Processing*, pp. 203-212.

Hatzivassiloglou, V., Klavans, J. L., Holcombe, M. L., Barzilay, R., Kan, M., & McKeown, K. R. (2001). SIMFINDER: A Flexible Clustering Tool for Summarization**.** In the *Proceedings of the NAACL Workshop for Summarization*, pp. 41-49.

Hirao, T., Suzuki, J., Isozaki, H., & Maeda, E. (2004). Dependency-based Sentence Alignment for Multiple Document Summarization. In the *COLING '04 Proceedings of the 20th international conference on Computational Linguistics*, pp. 446-452.

Hutchins, J. (1987). Summarization: Some problems and Methods**.** In the *Meaning: the frontier of informatics. Informatics 9*. In the *Proceedings of a conference jointly sponsored by Aslib, the Aslib Informatics Group, and the Information Retrieval Specialist Group of the British Computer Society*, King's College Cambridge, 26-27 March; edited by Kevin P. Jones., pp. 151-173.

Jing, H., & McKeown, K. (1999). The Decomposition of Human-Written Summary Sentences.
In the *Proceedings of the 22nd Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, pp. 129-136.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. 2nd edition. Prentice-Hall, 988p.

Kantrowitz, M. (2000). *Term-length term-frequency method for measuring document similarity and classifying text*. Justsystem Pittsburgh Research Center.

Kupiec, J., Pedersen, J., & Chen, F. (1995) A trainable document summarizer In the *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 68-73

Lee, L. (1999). Measures of distributional similarity. In *ACL-99*, pp. 25-32.

Mani, I. (2001). *Automatic Summarization. Natural Language Processing Vol. 3*. Amsterdam/Philadelphia: John Benjamins Publishing Company. [Centro de Linguística da Universidade do Porto, Cota: N/35], 285p.

Mann, W. C., & Thompson, S. A. (1987). *Rhetorical structure theory: A theory of text organization*. Tech. rep. ISI/RS-87-190, University of Southern California, 83p.

Marcu, D. (1999). The automatic construction of large-scale corpora for summarization research. In the *Proceedings of the 22nd Conference on Research and Development in Information Retrieval*, pp. 137-144.

Maziero, E. G., & Pardo, T. A. S. (2011). Multi-Document Discourse Parsing Using Traditional and Hierarchical Machine Learning. In the *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*, pp. 1-10. October 24-26, Cuiabá/MT, Brazil.

Maziero, E. G., Castro Jorge, M. L. C., & Pardo, T. A. S. (2010). Identifying Multidocument Relations. In the *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science - NLPCS*, pp. 60-69. June 8-12, Funchal/Madeira, Portugal.

Melamed, I. D. (2000). Pattern recognition for mapping bitext correspondence. In *VÉRONIS, J. (ed.). Parallel text processing: Alignment and use of translation corpora*. s.l: Kluwer Academic Publishers, pp. 25-47.

Miller, G. A. (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

Och, F. J., Christoph, T., & Hermann, N. (1999). Improved alignment models for statistical machine translation. In the *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 20-28. University of Maryland, College Park, June.

Okumura, M., Fukusima, T., & Nanba, H. (2003). Text Summarization Challenge 2 - Text Summarization Evaluation at NTCIR Workshop 3. *HLT-NAACL 2003 Workshop: Text Summarization (DUC03)*, pp. 49-56

Papageorgiou, H., Cranias, L., & Piperidis, S. (1994). Automatic alignment in parallel corpora. In the *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL 94)*, Las Cruses, New Mexico, pp. 334-336.

Pardo, T. A. S., & Rino, L. H. M. (2002). DMSumm: Um Gerador Automático de Sumários. In *Anais do I Workshop de Teses e Dissertações em Inteligência Artificial – WTDIA*, pp. 1-10. Recife-PE, Brasil. 11 a 14 de Novembro.

Pinto Molina, M. (1995). Document Abstracting: Toward a Methodological Model. *Journal of the American Society for Information Science* 46(3), pp. 225-234.

Piperidis, S., Papageorgiou, H., & Boutsis, S. (2000). From sentences to words and clauses. *In VÉRONIS, J. (ed.). Parallel text processing: Alignment and use of translation corpora*. s.l.: Kluwer Academic Publishers, pp. 117-138.

Radev, D.R. (2000). A common theory of information fusion from multiple text sources, step one: Cross ‑ document structure. In the *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*, pp. 74-83.

Radev, D. R., & McKeown, K. (1998). Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, v. 24, n. 3, pp. 469-500.

Radev, D.R., Otterbacher, J., & Zhang, Z. (2004). CST Bank: A Corpus for the Study of Crossdocument Structural Relationships. In the *Proceedings of Fourth International Conference on Language Resources and Evaluation*, 4p.

Radev, D. R., Zhang, Z., & Otterbacher, J. (2008). Cross-document relationship classification for text summarization. Available at http://clair.si.umich.edu/~radev/papers/progress/p1.pdf.

Seno, E. R. M., & Nunes, M. G. V. (2008). Some Experiments on Clustering Similar Sentences of Texts in Brazilian Portuguese. In the *Proceedings of the International Conference on Computational Processing of Portuguese Language – PROPOR (Lecture Notes in Artificial Intelligence, 5190)*, pp. 133-144.

Seno, E. R. M., & Nunes, M. G. V. (2009). Reconhecimento de Informações Comuns para a Fusão de Sentenças Comparáveis do Português. *Linguamática*, v. 1, pp. 71-87.

Seno, E. R. M. & Rino, L. H. M. (2005). Co-referential chaining for coherent summaries through rhetorical and linguistic modeling. In the *Proceedings of the Workshop on Crossing Barriers in Text Summarization Research/RANLP*. Borovets-Bulgaria, 6p.

Soricut, R. & Brill, E. (2004). *Automatic Question Answering: Beyond the Factoid.* In the *Proceedings of HLT-NAACL*, pp. 57-64.

Sparck Jones, K. (1999). Automatic Summarizing: Factors and Directions. In the *Advances in Automatic Text Summarization*, Mani, I. and Maybury, M. T. (eds.), pp. 1-12. Cambridge, Massachusetts: MIT Press.

Specia, L. (2010). Translating from Complex to Simplified Sentences. In the *Proceedings of PROPOR*, pp. 30-39.

Trigg, R. (1983). *A Network-Based Approach to Text Handling for the Online Scientific Community.* PhD thesis. Department of Computer Science, University of Maryland, 149p.

Trigg, R., & Weiser, M. (1986). TEXTNET: A network-based approach to text handling. *ACM Transactions on Office Information Systems*, v. 4, n. 1, pp. 1-23.

Uzêda, V. R., Pardo, T. A. S., & Nunes, M. G. V. (2010). A Comprehensive Comparative Evaluation of RST-Based Summarization Methods. *ACM Transactions on Speech and Language Processing*, v. 6, N. 4, pp. 1-20.

Vogel, S., Hermann, N., & Christoph, T. (1996). HMM-based word alignment in statistical translation. In *COLING '96: The 16th International Conference on Computational Linguistics*, pp. 836-841. Copenhagen, Denmark, August.

Viterbi, A. J. (1967). Error bounds for convolution codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, v. 13, pp. 260-269.

Yamada, K., & Knight, K. (2001). A syntax-based statistical translation model. In the *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 523-530. Toulouse, France, July.

Zhang, Z., Otterbacher, J., & Radev, D.R. (2003). Learning Cross-document Structural Relationships using Boosting. In the *Proceedings of the twelfth international conference on Information and knowledge management*, pp. 124-130.